Is Inverse Reinforcement Learning Harder than Standard Reinforcement Learning? A Theoretical Perspective

Lei Zhao[1] **Mengdi Wang**[2] **Yu Bai**[3]

Abstract

Inverse Reinforcement Learning (IRL)—the problem of learning reward functions from demonstrations of an *expert policy*—plays a critical role in developing intelligent systems. While widely used in applications, theoretical understandings of IRL present unique challenges and remain less developed compared with standard RL. For example, it remains open how to do IRL efficiently in standard *offline* settings with pre-collected data, where states are obtained from a *behavior policy* (which could be the expert policy itself), and actions are sampled from the expert policy.

This paper provides the first line of results for efficient IRL in vanilla offline and online settings using polynomial samples and runtime. Our algorithms and analyses seamlessly adapt the pessimism principle commonly used in offline RL, and achieve IRL guarantees in stronger metrics than considered in existing work. We provide lower bounds showing that our sample complexities are nearly optimal. As an application, we also show that the learned rewards can *transfer* to another target MDP with suitable guarantees when the target MDP satisfies certain similarity assumptions with the original (source) MDP.

1. Introduction

Inverse Reinforcement Learning (IRL) aims to recover reward functions from demonstrations of an *expert policy* (Ng & Russell, 2000; Abbeel & Ng, 2004), in contrast to standard reinforcement learning which aims to learn optimal policies for a given reward function. IRL has applications in numerous domains such as robotics (Argall et al., 2009; Finn et al., 2016), target-driven navigation tasks (Ziebart et al., 2008; Sadigh et al., 2017; Kuderer et al., 2015; Pan et al., 2020; Barnes et al., 2023), game AI (Ibarz et al., 2018; Vinyals et al., 2019), and medical decision-making (Woodworth et al., 2018; Hantous et al., 2022). The learned reward functions in these applications are typically used for replicating the expert behaviors in similar or varying downstream environments. Broadly, the problem of learning reward functions from data is of rising importance beyond the scope of IRL, and is used in procedures such as Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017) for aligning large language models (Ouyang et al., 2022; Bai et al., 2022; OpenAI, 2023; Touvron et al., 2023).

Despite the success of IRL in practical applications (Agarwal et al., 2020; Finn et al., 2016; Sadigh et al., 2017; Kuderer et al., 2015; Woodworth et al., 2018; Wu et al., 2020; Ravichandar et al., 2020; Vasquez et al., 2014), theoretical understanding is still in an early stage and presents several unique challenges, especially when compared with standard RL (finding optimal policy under a given reward) where the theory is more established. First, the solution is **inherently non-unique** for *any* IRL problem—For example, for any given expert policy, zero reward is always a feasible solution (making the expert policy optimal under this reward). A sensible definition of IRL would require not just recovering a single reward function but instead a *set* of feasible rewards (Metelli et al., 2021; Lindner et al., 2023). Second, theoretical results for IRL is **lacking even for some standard learning settings**, such as learning from an offline dataset of trajectories from the expert policy (akin to an imitation setting). Finally, as a more nuanced challenge (but related to both challenges above), so far there is **no commonly agreed performance metric** for measuring the distance between the estimated reward set and the ground truth reward set. Existing performance metrics in the literature either require strong feedback such as a simulator (Metelli et al., 2021; 2023), or do not require the returned solution to be aware of the transition dynamics Lindner et al. (2023) (see Section 3.3 for a discussion). These challenges motivate the following open question:

Is IRL more difficult than standard RL?

[1] School of Mathematical Sciences, University of Science and Technology of China, Hefei, China [2] Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ, USA [3] Salesforce AI Research, Palo Alto, CA, USA. Correspondence to: Lei Zhao <zl20021031@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

In this paper, we theoretically study IRL in standard episodic tabular Markov Decision Processes without Rewards (MDP\R's) under vanilla offline and online learning settings. Our contributions can be summarized as follows.

- The goal of IRL is to output a set of rewards that approximate the ground truth set of *feasible* rewards, i.e. rewards under which the expert policy is optimal. We define new metrics for both reward functions and for IRL using the concept of *reward mapping*, which can be viewed as a "generating function" of the (ground truth) set of feasible rewards (Section 2.1 & 3.1). We show that our metrics are stronger / more appropriate than existing metrics in certain aspects (Section 3.3).
- We show that any estimated reward that is similar in our metric and satisfies monotonicity with respect to the true reward admits an approximate planning/learning guarantee (Section 3.2).
- We design an algorithm, REWARD LEARNING WITH PESSIMISM (RLP) that performs IRL from any given offline demonstration dataset (Section 4). Our algorithm returns an estimated reward mapping that is ϵ-close in our metric and satisfies monotonicity, and requires a number of episodes that is polynomial in the size of the MDP as well as the single-policy concentrability coefficient between the *evaluation policy* and the behavior policy that generated the states of the offline dataset. To our best knowledge, this is the first provably sample-efficient algorithm for IRL in the standard offline setting.

  Technically, the algorithm seamlessly adapts the pessimism principle from the offline RL literature to achieve the desired monotonicity and closeness conditions, demonstrating that IRL is "not much harder than standard RL" in a certain sense.
- We next design an algorithm REWARD LEARNING WITH EXPLORATION (RLE), which operates in a natural online setting where the learner can both actively explore the environment and query the expert policy, and achieves IRL guarantee in a stronger metric from polynomial samples (Section 5). Algorithm RLE builds on a simple reduction to reward-free exploration (Jin et al., 2020; Li et al., 2023) and the RLP algorithm.
- We establish sample complexity lower bounds for both the offline and online settings, showing that our upper bounds are nearly optimal up to a small factor (Section 4.4 & 5.3).
- We extend our results to a *transfer learning* setting, where the learned reward mapping is transferred to and evaluated in a target MDP\R different from the source MDP\R. We provide guarantees for RLP and RLE under certain similarity assumptions between the source and target MDP\Rs (Section 6 & Appendix I).

1.1. Related work

Inverse reinforcement learning Inverse reinforcement learning (IRL) was first proposed by (Ng & Russell, 2000) and since then significantly developed in various follow-up approaches such as feature matching (Abbeel & Ng, 2004), maximum margin (Ratliff et al., 2006), maximum entropy (Ziebart et al., 2008), relative entropy (Boularias et al., 2011), and generative adversarial imitation learning (Ho & Ermon, 2016). Other notable approaches include Bayesian IRL (Ramachandran & Amir, 2007) which subsume IRL, and the reduction method (Brantley et al., 2019).

IRL has been successfully applied in many domains including target-driven navigation tasks (Ziebart et al., 2008; Sadigh et al., 2017; Kuderer et al., 2015; Pan et al., 2020), robotics (Argall et al., 2009; Finn et al., 2016; Hadfield-Menell et al., 2016; Kretzschmar et al., 2016; Okal & Arras, 2016; Kumar et al., 2023; Jara-Ettinger, 2019), medical decision-making (Woodworth et al., 2018; Hantous et al., 2022; Gong et al., 2023; Yu et al., 2019; Chadi & Mousannif, 2022), and game AI (Finn et al., 2016; Fu et al., 2017; Qureshi et al., 2018; Brown et al., 2019).

Theoretical understandings of IRL Despite their successful applications, theoretical understandings of IRL are still in an early stage. Prior theoretical work (Ziebart et al., 2008) considers how to efficiently pick a single reward that best differentiates the expert policy from other policies in the learner's policy class. The IRL with this as the learning goal has achieved some significant theoretical results (Swamy et al., 2021; 2022; 2023). Recently, Metelli et al. (2021) pioneered the investigation of the sample complexity of reward-set estimation for IRL (a different learning goal from prior work) under the simulator (generative model) setting where the learner can directly query feedback from any (state, action) pair. This work was later extended by Metelli et al. (2023), who introduced a framework based on Hausdorff-based metrics for measuring distances between reward sets, examined relationships between different metrics, and provided corresponding lower bounds. However, their results critically rely on the simulator setting and do not generalize to more realistic offline/online learning settings. Dexter et al. (2021) also performed a theoretical analysis for IRL in the simulator setting with continuous states and discrete actions.

The recent work of Lindner et al. (2023) considers IRL in the online setting where the learner can interact with the MDP\R in an online fashion, which is closely related to our results for the online setting. Compared with our metric, their metric is defined for an estimated IRL problem (instead of an estimated reward set). Further, their metric does not effectively take into account the estimated transitions, which can lead to a family of counter-exmaples where the estimated IRL problem achieves perfect recovery under their metric, but the induced reward sets are actually far from the true feasible reward set in our metric (cf. Section 3.3 for a detailed discussion). Our work improves upon the above works by introducing new performance metrics for IRL, and providing new algorithms for standard learning settings such as offline learning.

Relationship with standard RL theory Our work builds upon various existing techniques from the sample-efficient RL literature to design our algorithms and establish our theoretical results. For the offline setting, our algorithm and analysis build upon the pessimism principle and the single-policy concentrability condition commonly used in offline RL (Kidambi et al., 2020; Jin et al., 2021; Yu et al., 2020; Kumar et al., 2020; Rashidinejad et al., 2021; Xie et al., 2021; 2022). For the online setting, we adapt the reward-free learning algorithm of Li et al. (2023) to find a policy that achieves a certain concentrability-like condition with respect to all policies.

We note theoretical results on imitation learning (Abbeel & Ng, 2004; Ratliff et al., 2006; Ziebart et al., 2008; Levine et al., 2011; Fu et al., 2017; Chang et al., 2021) and RLHF (Zhu et al., 2023a;b; Wang et al., 2023; Zhan et al., 2023), which are related to but different from (and do not imply) our results. Additional related work is discussed in Appendix A due to the space limit.

2. Preliminaries

Markov Decision Processes without Reward We consider episodic Markov Decision Processes without Reward (MDP\R), specified by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P})$, where $\mathcal{S}$ is the state space with $|\mathcal{S}| = S$, $\mathcal{A}$ is the action space with $|\mathcal{A}| = A$, H is the horizon length, $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$ where $\mathbb{P}_h(\cdot|s,a) \in \Delta(\mathcal{S})$ is the transition probability at step h. Without loss of generality, we assume that the initial state is deterministically some $s_1 \in \mathcal{S}$.

Reward functions A reward function $r : [H] \times \mathcal{S} \times \mathcal{A} \to [-1, 1]$ maps a state-action-time step triplet (h, s, a) to a reward $r_h(s, a)$. Given an MDP\R $\mathcal{M}$ and a reward function r, we denote the MDP induced by $\mathcal{M}$ and r as $\mathcal{M} \cup r$. A policy $\pi = \{\pi_h(\cdot \mid s)\}_{h\in[H],s\in\mathcal{S}}$, where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ maps a state to an action distribution.

Values and visitation distributions A policy $\pi = (\pi_h)_{h\in[H]}$, where each $\pi_h(\cdot|s) \in \Delta(\mathcal{A})$ for each $s \in \mathcal{S}$. Let $\mathrm{supp}(\pi_h(\cdot|s)) := \{a : \pi_h(a|s) > 0\}$ denote the support set of $\pi_h(\cdot|s)$. For any policy π and any reward function r, we define the value function $V_h^\pi(\cdot; r) : \mathcal{S} \to \mathbb{R}$ at each time step $h \in [H]$ by the expected cumulative reward: $V_h^\pi(s; r) = \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \mid s_h = s\right]$, where $\mathbb{E}_\pi$ denotes the expectation with respect to the random trajectory induced by π in the MDP\R, that is, $(s_1, a_1, s_2, a_2, ..., s_H, a_H)$, where $a_h \sim \pi_h(s_h), r_h = r_h(s_h, a_h), s_{h+1} \sim \mathbb{P}_h(\cdot \mid s_h, a_h)$. Similarly, we denote the Q-function at time step h as : $Q_h^\pi(s, a; r) = \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a\right]$. For any reward r, the corresponding advantage function $A_h^\pi(\cdot; r) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as $A_h^\pi(s, a; r) := Q_h^\pi(s, a; r) - V_h^\pi(s; r)$ and we say a policy is an optimal policy of $\mathcal{M} \cup r$ if $A_h^\pi(s, a; r) \le 0$ holds for all $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$[1]. Additionally, we represent the set of all optimal policies for $\mathcal{M} \cup r$ as $\Pi^\star_{\mathcal{M}\cup r}$ and denote the set of all deterministic policies for $\mathcal{M} \cup r$ as $\Pi^{\mathrm{det}}_{\mathcal{M}\cup r}$.

We introduce d_h^π to denote the state(-action) visitation distributions associated with policy at time step $h \in [H]$: $d_h^\pi(s) := \mathbb{P}(s_h = s|\pi)$ and $d_h^\pi(s, a) := \mathbb{P}(s_h = s, a_h = a|\pi)$. Lastly, we define the operators $\mathbb{P}_h$ and $\mathbb{V}_h$ by $[\mathbb{P}_h V_{h+1}](s, a) := \mathbb{E}[V_{h+1}(s_{h+1})|s_h = s, a_h = a]$ and $[\mathbb{V}_h V_{h+1}](s, a) := \mathsf{Var}[V_{h+1}(s_{h+1})|s_h = s, a_h = a]$ applying to any value function V_{h+1} at time step $h+1$. In this paper, we will frequently employ $\widehat{\mathbb{P}}_h$ and $\widehat{\mathbb{V}}_h$ to represent empirical counterparts of these operators constructed based on estimated models. For any function $f : \mathcal{S} \to \mathbb{R}$, define its infinity norm as $||f||_\infty := \sup_{s\in\mathcal{S}} |f(s)|$ (and we define similarly for any $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$).

2.1. Inverse Reinforcement Learning

An Inverse Reinforcement Learning (IRL) problem is denoted as a pair $(\mathcal{M}, \pi^{\mathsf{E}})$, where $\mathcal{M}$ is an MDP\R and π^{E} is a policy called the *expert policy*. We say a reward r is a feasible reward for $(\mathcal{M}, \pi^{\mathsf{E}})$ if π^{E} is an optimal policy of $\mathcal{M} \cup r$, i.e., $A_h^{\pi^{\mathsf{E}}}(s, a; r) \le 0$ for all s, a. The goal of IRL is to interact with $(\mathcal{M}, \pi^{\mathsf{E}})$, and recover reward function r's that are *feasible* for $(\mathcal{M}, \pi^{\mathsf{E}})$.

Reward mapping Noting that learning *one* feasible reward function is trivial (the zero reward $r \equiv 0$ is feasible for any π^{E}), we consider the stronger goal of recovering the *set* of all feasible rewards, which can be characterized by an explicit formula by the classical result of Ng & Russell (2000). Here we restate this result through the concept of a *reward mapping*.

Let $\mathcal{R}^{\mathsf{all}} = \{r : \mathcal{S} \times \mathcal{A} \times [H] \to \mathbb{R}\}$ denote the set of all possible reward functions, and $\mathcal{R}^{\mathsf{feas}}_{[-B,B]} := \{r \in \mathcal{R}^{\mathsf{all}} : r \text{ is feasible and } |r| \le B\}$ denote the set of all feasible rewards bounded by B for any $B > 0$. Let $\overline{\mathcal{V}} := \overline{\mathcal{V}}_1 \times \cdots \times \overline{\mathcal{V}}_H$ and $\overline{\mathcal{A}} := \overline{\mathcal{A}}_1 \times \cdots \times \overline{\mathcal{A}}_H$, where $\overline{\mathcal{V}}_h := \left\{V_h \in \mathbb{R}^{\mathcal{S}} \mid \|V_h\|_\infty \le H - h + 1\right\}$ and $\overline{\mathcal{A}}_h := \left\{A_h \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\ge 0} \mid \|A_h\|_\infty \le H - h + 1\right\}$ denote the set of all possible "value functions" and "advantage functions"

[1] This definition of optimal policy requires π to be optimal starting from any time step h and state $s \in \mathcal{S}$ (not necessarily visitable ones), which is stronger than the standard definition but is commonly adopted in the IRL literature (Ng & Russell, 2000).

respectively.

Definition 1 (Reward mapping)**.** *The (ground truth)* reward mapping $\mathscr{R}^\star : \overline{\mathcal{V}} \times \overline{\mathcal{A}} \mapsto \mathcal{R}^{\mathsf{all}}$ *of an IRL problem* $(\mathcal{M}, \pi^{\mathsf{E}})$ *is the mapping that maps any* $(V, A) \in \overline{\mathcal{V}} \times \overline{\mathcal{A}}$ *to the following reward function* r*:*

$$r_h(s,a) = [\mathscr{R}^\star(V,A)]_h(s,a) := -A_h(s,a) \times \mathbf{1}\left\{a \notin \operatorname{supp}\left(\pi^{\mathsf{E}}_h(\cdot\,|\,s)\right)\right\} + V_h(s) - [\mathbb{P}_h V_{h+1}](s,a), \tag{1}$$

where we recall that $\mathbb{P}_h$ *is the transition probability of* $\mathcal{M}$ *at step* $h \in [H]$.

With the definition of reward mapping ready, we now restate the classical result of Ng & Russell (2000), which shows that the reward mapping $\mathscr{R}^\star$ generates a set of rewards that is a superset of $\mathcal{R}^{\mathsf{feas}}_{[-1,1]}$—the set of all $[-1,1]$-bounded feasible rewards—by ranging over $(V, A) \in \overline{\mathcal{V}} \times \overline{\mathcal{A}}$.

Lemma 2 (Reward mapping produces all bounded feasible rewards)**.** *The set of rewards* $\mathscr{R}^\star(\overline{\mathcal{V}} \times \overline{\mathcal{A}}) = \{\mathscr{R}^\star(V,A) : (V,A) \in \overline{\mathcal{V}} \times \overline{\mathcal{A}}\}$ *induced by* $\mathscr{R}^\star$ *satisfies*

$$\mathcal{R}^{\mathsf{feas}}_{[-1,1]} \subseteq \mathscr{R}^\star(\overline{\mathcal{V}} \times \overline{\mathcal{A}}) \subseteq \mathcal{R}^{\mathsf{feas}}_{[-3H,3H]}. \tag{2}$$

In words, $\mathscr{R}^\star$ *always produces feasible rewards bounded in* $[-3H, 3H]$*, and the set* $\mathscr{R}^\star(\overline{\mathcal{V}} \times \overline{\mathcal{A}})$ *contains (is a superset of) all* $[-1,1]$*-bounded feasible rewards.*

As IRL is concerned precisely with the recovery of the set $\mathcal{R}^{\mathsf{feas}}_{[-1,1]}$, we consider the recovery of the reward mapping $\mathscr{R}^\star$ itself as a natural learning goal—An accurate estimator $\widehat{\mathscr{R}} \approx \mathscr{R}^\star$ guarantees $\widehat{\mathscr{R}}(V,A) \approx \mathscr{R}^\star(V,A)$ for any $(V,A) \in \overline{\mathcal{V}} \times \overline{\mathcal{A}}$, and thus imply accurate estimation of $\mathscr{R}^\star(\overline{\mathcal{V}} \times \overline{\mathcal{A}})$ in precise ways which we specify in the sequel.

We will also consider recovering the reward mapping on a *subset* $\Theta \subset \overline{\mathcal{V}} \times \overline{\mathcal{A}}$. We use the following standard definition of covering numbers to measure the capacity of such Θ's:

Definition 3 (Covering number)**.** *The* ϵ-covering number *of* $\Theta \subset \overline{\mathcal{V}} \times \overline{\mathcal{A}}$ *is defined as*

$$\mathcal{N}(\Theta; \epsilon) := \max\nolimits_{h \in [H]} \mathcal{N}(\overline{\mathcal{V}}^{\Theta}_h; \epsilon),$$

where $\overline{\mathcal{V}}^{\Theta}_h := \{V_h : (V,A) \in \Theta\}$ *denotes the restriction of* Θ *onto* $\overline{\mathcal{V}}_h$*, and* $\mathcal{N}(\overline{\mathcal{V}}^{\Theta}_h; \epsilon)$ *is the* ϵ*-covering number of* $\overline{\mathcal{V}}^{\Theta}_h$ *in* $\|\cdot\|_\infty$ *norm.*

Note that $\log \mathcal{N}(\Theta; \epsilon) \le \min\{\log|\Theta|, \mathcal{O}(S \log(H/\epsilon))\}$ by combining the (trivial) bound for the finite case and the standard covering number bound for $\Theta = \overline{\mathcal{V}} \times \overline{\mathcal{A}}$ (Vershynin, 2018). In addition, the left-hand side may be much smaller than the right-hand side if Θ admits additional structure (for example, if $\overline{\mathcal{V}}^{\Theta}_h$ lies in a low-dimensional subspace of $\mathbb{R}^{\mathcal{S}}$).

3. Performance metrics for IRL

3.1. Metric for IRL

We now define our performance metric for IRL based on the recovery of reward mapping $\mathscr{R}^\star$. Fixing any MDP\R $\mathcal{M}$, we begin by defining our *base metric* d^π (indexed by a policy π) and d^{all} between two rewards.

Definition 4 (Base metric for rewards)**.** *We define the metric*[2] d^π *(indexed by any policy* π*) between any pair of rewards* $r, r' \in \mathcal{R}^{\mathsf{all}}$ *as*

$$d^\pi(r, r') := \sup_{h \in [H]} \mathbb{E}_{s_h \sim \pi} |V^\pi_h(s_h; r) - V^\pi_h(s_h; r')|. \tag{3}$$

We further define $d^{\mathsf{all}}(r, r') := \sup_\pi d^\pi(r, r')$.

In words, metric d^π compares the rewards r and r' when executing π. Concretely, (3) compares the difference in the value functions $V^\pi_h(\cdot; r)$ and $V^\pi_h(\cdot; r')$ averaged over the visitation distribution $s_h \sim \pi$, which is sensible for our learning settings as it takes into account the transition structure of $\mathcal{M}$ (compared with other existing metrics based the sup-distance over all states; cf. Section 3.3). The stronger metric d^{all} takes the supremum of d^π over all policy π's. We also note that our metric is related to EPIC distance in Gleave et al. (2020).

We now define our main metric D^π_Θ for the recovery of reward mappings, which simply takes the supremum of d^π between all pairs of rewards induced by the two reward mappings using the *same* parameter $(V, A) \in \Theta$. Here, a reward mapping represents a mapping from $\overline{\mathcal{V}} \times \overline{\mathcal{A}}$ to $\mathcal{R}^{\mathsf{all}}$.

Definition 5 (Metric for reward mappings)**.** *Given any policy* π *and any parameter set* Θ*, we define the metric*[2] D^π_Θ *between any pair of reward mappings* $\mathscr{R}, \mathscr{R}'$ *as*

$$D^\pi_\Theta(\mathscr{R}, \mathscr{R}') := \sup_{(V,A) \in \Theta} d^\pi(\mathscr{R}(V,A), \mathscr{R}'(V,A)). \tag{4}$$

We further define $D^{\mathsf{all}}_\Theta(\mathscr{R}, \mathscr{R}') := \sup_\pi D^\pi_\Theta(\mathscr{R}, \mathscr{R}')$.

(4) compares two reward mappings $\mathscr{R}$ and $\mathscr{R}'$ by measuring the distance between $\mathscr{R}(V,A)$ and $\mathscr{R}'(V,A)$ using our base metric and taking the sup over all $(V,A) \in \Theta$. Another common choice in the IRL literature is the Hausdorff distance (based on some base metric) between the two sets $\mathscr{R}(\overline{\mathcal{V}} \times \overline{\mathcal{A}})$ and $\mathscr{R}'(\overline{\mathcal{V}} \times \overline{\mathcal{A}})$ (Metelli et al., 2021; 2023; Lindner et al., 2023). We show that (4) is always stronger than the Hausdorff distance in the sense that a metric of the form (4) is greater or equal to the Hausdorff distance regardless of the base metric (Lemma D.3), and the inequality can be strict for some base metric (Lemma D.4).

[2] Technically a semi-metric.

3.2. Implications for learning with estimated reward

For IRL, a natural desire for a base metric between rewards is that, a small metric between r and $\widehat{r}$ should imply that learning (planning) using reward $\widehat{r}$ in $\mathcal{M}$ should at most incur a small error when the true reward is r. The following result shows that our metric d^{π} satisfies such a desiderata. The proof can be found in Appendix D.5.

Proposition 6 (Planning with estimated reward)**.** *Given an MDP\R $\mathcal{M}$, let $r, \widehat{r}$ be a pair of rewards such that*

- *(a) (Small d^{π} on near-optimal policy) $d^{\pi}(r, \widehat{r}) \le \epsilon$ for some $\bar{\epsilon}$ near-optimal policy π for MDP $\mathcal{M} \cup r$;*
- *(b) (Monotonicity) $\widehat{r}_h(s,a) \le r_h(s,a)$ for any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$.*

Then, letting $\widehat{\pi}$ be any ϵ' near-optimal policy for MDP $\mathcal{M} \cup \widehat{r}$, i.e, $V_1^{\star}(s_1; \widehat{r}) - V_1^{\widehat{\pi}}(s_1; \widehat{r}) \le \epsilon'$, we have

$$V_1^{\star}(s_1; r) - V_1^{\widehat{\pi}}(s_1; r) \le \epsilon + \epsilon' + 2\bar{\epsilon}, \qquad (5)$$

i.e. $\widehat{\pi}$ is also $(\epsilon + \epsilon' + 2\bar{\epsilon})$ near-optimal for $\mathcal{M} \cup r$.

Proposition 6 ensures that any estimated reward $\widehat{r}$ that satisfies (a) small d^{π} and (b) monotonicity with respect to the true reward will incur a small error when used in planning. We emphasize that monotonicity is necessary in order for (5) to hold, similar to how pessimism is necessary for near-optimal learning in offline bandits/RL (Jin et al., 2021). Throughout the rest of the paper, we focus on designing IRL algorithms that satisfy (a) & (b). These guarantees can then directly yield planning/learning guarantees as corollaries by Proposition 6, and we will omit such statements.

3.3. Relationship with existing metrics

Our metrics d^{π} and d^{all} differ from several metrics for IRL used in existing theoretical work, which we discuss here.

Lindner et al. (2023) measures the difference between two reward mappings implicitly by a metric D^L (see (23)) between the two inducing IRL problems (the ground truth problem $(\mathcal{M}, \pi^{\mathsf{E}})$ and the estimated problem $(\widehat{\mathcal{M}}, \widehat{\pi}^{\mathsf{E}})$ returned by an algorithm). The following result shows that D^L is weaker than our metric $D_{\Theta}^{\mathsf{all}}$ in a strong sense.

Theorem 7 (Relationship with D^L; informal)**.** *The metric D^L defined in (23) satisfies the following:*

- *(a) (Informal version of Prop. D.1) Under the same setting as Theorem 11 (in which our algorithm* RLE *achieves ϵ error in $D_{\Theta}^{\mathsf{all}}$),* RLE *also achieves ϵ error in D^L with the same sample complexity therein.*
- *(b) (Informal version of Prop.D.2) Conversely, there exists a family of pairs of IRL problems which has distance 0 in the D^L metric but distance 1 in the $D_{\Theta}^{\mathsf{all}}$ metric between the induced reward mappings.*

In a separate thread, the works of Metelli et al. (2021; 2023) consider IRL under access to a simulator. Their metric between two reward functions requires the induced value/Q functions to be close *uniformly over all* $(s,a) \in \mathcal{S} \times \mathcal{A}$ (cf. Appendix D.2), regardless of whether the state is visitable by a policy in this particular MDP\R), which is tailored to the simulator setting and does not applicable to the standard offline/online settings considered in this work. By contrast, our metrics d^{π} and d^{all} measure the distance between the induced value functions *averaged over visitation distributions*, which are more tractable for the offline/online settings.

4. IRL in the offline setting

4.1. Setting

In the offline setting, the learner does not know $(\mathcal{M}, \pi^{\mathsf{E}})$, and only has access to a dataset $\mathcal{D} = \{(s_h^k, a_h^k, e_h^k)\}_{k=1,h=1}^{K,H}$ consisting of K iid trajectories without reward from $\mathcal{M}$, where actions are obtained by executing some *behavior policy* π^{b} in $\mathcal{M}$: $a_h^k \sim \pi_h^{\mathsf{b}}(\cdot|s_h^k)$ for all (k,h), and the *expert feedback* e_h^k's are obtained from the expert policy π^{E} using one of the following two options:

$$e_h^k = \begin{cases} a_h^{\mathsf{E},k} \sim \pi_h^{\mathsf{E}}(\cdot|s_h^k) & \text{in option 1,} \\ \mathbf{1}\left\{a_h^k \in \operatorname{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s_h^k)\right)\right\} & \text{in option 2.} \end{cases} \qquad (10)$$

Option 1, where the learner directly observes an *expert action* $a_h^{\mathsf{E},k}$, is the commonly employed setting in the IRL literature (Metelli et al., 2021; Lindner et al., 2023; Metelli et al., 2023). In the special case where $\pi^{\mathsf{b}} = \pi^{\mathsf{E}}$, we can take $a_h^{\mathsf{E},k} := a_h^k$, i.e. no need for additional expert feedback when the behavior policy coincides with the expert policy. We also allow option 2, in which e_h^k indicates whether a_h^k "is an expert action" (belongs to the support of $\pi_h^{\mathsf{E}}(\cdot|s)$). As we will see, both options suffice for performing IRL.

Additionally, for option 1, we require the following well-posedness assumption on the expert policy π^{E}.

Assumption A (Well-posedness)**.** *For any $\Delta \in (0,1]$, we say policy π^{E} is Δ-well-posed if*

$$\min_{(h,s,a):\pi_h^{\mathsf{E}}(a|s) \neq 0} \pi_h^{\mathsf{E}}(a|s) \ge \Delta. \qquad (11)$$

This assumption is also made by Metelli et al. (2023, Assumption D.1), and is necessary for ruling out the edge case where $\pi_h^{\mathsf{E}}(a|s)$ is positive but extremely small for some action $a \in \mathcal{A}$, in which case a large number of samples is required to determine $\mathbf{1}\left\{a \in \operatorname{supp}(\pi_h^{\mathsf{E}}(\cdot|s))\right\}$.

4.2. Algorithm

We now present our algorithm REWARD LEARNING WITH PESSIMISM (RLP; full description in Algorithm 1) for IRL

Algorithm 1 Reward Learning with Pessimism

```
Input: Dataset $\mathcal{D} = \{(s_h^k, a_h^k, e_h^k)\}_{k=1,h=1}^{K,H}$, parameter set $\Theta \subset \overline{\mathcal{V}} \times \overline{\mathcal{A}}$, confidence level $\delta > 0$, error tolerance $\epsilon > 0$.
for $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$ do
    Compute the empirical transition kernel $\widehat{\mathbb{P}}_h$, the empirical expert policy $\widehat{\pi}^{\mathsf{E}}$ and the penalty term $b_h^\theta$ for all $\theta \in \Theta$ as follows:
    $$\widehat{\mathbb{P}}_h(s' \mid s, a) = \frac{1}{N_h^b(s,a) \vee 1} \sum_{(s_h, a_h, s_{h+1}) \in \mathcal{D}} \mathbf{1}\left\{(s_h, a_h, s_{h+1}) = (s, a, s')\right\}, \qquad (6)$$
    $$\widehat{\pi}_h^{\mathsf{E}}(a \mid s) = \begin{cases} \frac{1}{N_h^b(s) \vee 1} \cdot \sum_{(s_h, a_h, e_h) \in \mathcal{D}} \mathbf{1}\left\{(s_h, e_h) = (s, a)\right\} & \text{in option 1,} \\ \frac{1}{N_{h,1}^b(s) \vee 1} \cdot \sum_{(s_h, a_h, e_h) \in \mathcal{D}} \mathbf{1}\left\{(s_h, a_h, e_h) = (s, a, 1)\right\} & \text{in option 2,} \end{cases} \qquad (7)$$
    $$b_h^\theta(s,a) = C \cdot \min\left\{ \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H)\iota}{N_h^b(s,a) \vee 1} \left[\widehat{\mathbb{V}}_h V_{h+1}\right](s,a)} + \frac{H \log \mathcal{N}(\Theta; \epsilon/H)\iota}{N_h^b(s,a) \vee 1} + \frac{\epsilon}{H}\left(1 + \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H)\iota}{N_h^b(s,a) \vee 1}}\right), H \right\}, \qquad (8)$$
    where the visitation counts $N_h^b(s,a) := \sum_{(s_h, a_h) \in \mathcal{D}} \mathbf{1}\left\{(s_h, a_h) = (s,a)\right\}$, $N_h^b(s) := \sum_{a \in \mathcal{A}} N_h^b(s,a)$, $N_{h,1}^b(s) := \sum_{(s_h, a_h, e_h)} \mathbf{1}\left\{(s_h, e_h) = (s, 1)\right\}$, $\iota := \log\left(HSA/\delta\right)$ and $C > 0$ is an absolute constant.
end for
Output: Estimated reward mapping $\widehat{\mathscr{R}}$ defined as follows: For all $(V, A) \in \Theta$,
    $$[\widehat{\mathscr{R}}(V,A)]_h(s,a) := -A_h(s,a) \cdot \mathbf{1}\left\{a \notin \operatorname{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\} + V_h(s) - [\widehat{\mathbb{P}}_h V_{h+1}](s,a) - b_h^\theta(s,a). \qquad (9)$$
```

in the offline setting. RLP returns an estimated reward mapping $\widehat{\mathscr{R}}$ given any offline dataset $\mathcal{D}$. At a high level, RLP consists of two main steps:

- (Empirical MDP) We estimate the transition probabilities $\mathbb{P}_h$ and expert policy π^{E} by standard empirical estimates $\widehat{\mathbb{P}}_h$ and $\widehat{\pi}^{\mathsf{E}}$, as in (6) and (7).
- (Pessimism) We compute a bonus function $b_h^\theta(s,a)$ for any $\theta = (V, A) \in \Theta$, $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$ as in (8). The final estimated reward (and thus the reward mapping) (9) is defined by the empirical version of the ground truth reward (1) combined with the negative bonus $-b_h^\theta(s,a)$, for every parameter $(V, A) \in \Theta$.

  The specific design of $b_h^\theta(s,a)$ is based on Bernstein's inequality, and ensures that with high probability, for all (h, s, a, θ) simultaneously,

  $$b_h^\theta(s,a) \geq A_h(s,a) \times \left|\mathbf{1}\left\{a \notin \operatorname{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\} - \mathbf{1}\left\{a \notin \operatorname{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\}\right| + \left|\left[\left(\widehat{\mathbb{P}}_h - \mathbb{P}\right)V_{h+1}\right](s,a)\right|.$$

  Combined with the form of the ground truth reward $\mathscr{R}(V, A)$ in (1), a standard pessimism argument ensures the monotonicity condition $[\widehat{\mathscr{R}}(V,A)]_h(s,a) \leq [\mathscr{R}(V,A)]_h(s,a)$ for all (h, s, a) and all (V, A).

Here, we provide the rationale for introducing the pessimism principle. Proposition 6 states that any estimated reward satisfying (a) small distance to true reward under any near-optimal policy + (b) monotonicity (less or equal to true reward) ensures that learning with the estimated reward also gives a near-optimal policy. By Proposition 6, it suffices for an IRL algorithm to satisfy both (a) and (b) in order for the learned reward to be useful. In Algorithm 1, the empirical estimates ensure (a) and the negative bonus is used in Eq.(9) to ensure the monotonicity condition while choosing the right bonus to not harm (a). We also note that Algorithm 1 is computationally efficient.

4.3. Theoretical guarantee

We now state our theoretical guarantee for Algorithm 1. To measure the quality of the recovered reward mappings, we will be considering the d^π and D_Θ^π metric with $\pi = \pi^{\mathsf{eval}}$ being any given *evaluation policy*. We assume that π^{eval} satisfies the standard single-policy concentrability condition with respect to the behavior policy π^{b}.

Assumption B (Average form single-policy concentrability)**.** *We say π^{eval} satisfies $C^\star$-single-policy concentrability with respect to π^{b} if (with the convention $0/0 = 0$)*

$$\frac{1}{HS} \sum_{h \in [H]} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{d_h^{\pi^{\mathsf{eval}}}(s,a)}{d_h^{\pi^{\mathsf{b}}}(s,a)} \leq C^\star. \qquad (12)$$

Assumption B is standard in the offline RL literature (Jin et al., 2021; Rashidinejad et al., 2021; Xie et al., 2021), though we remark that our (12) only requires the *average* form, instead of the worst-case form made in (Rashidinejad et al., 2021; Xie et al., 2021) which requires the distribution ratio to be bounded for all (h, s, a).

We are now ready to present the guarantee for RLP (Algorithm 1). The proof can be found in Appendix E.2.

Theorem 8 (Sample complexity of RLP)**.** *Let π^{eval} be any policy that satisfies $C^\star$ single-policy concentrability (Assumption B) with respect to π^{b}. Assume that π^{E} is Δ-well-posed (Assumption A) if we choose option 1 in (10).*

Then for both options, with probability at least $1-\delta$, RLP *(Algorithm 1) outputs a reward mapping $\widehat{\mathscr{R}}$ such that*

$$D_\Theta^{\pi^{\mathsf{eval}}}\left(\mathscr{R}^\star, \widehat{\mathscr{R}}\right) \le \epsilon, \left[\widehat{\mathscr{R}}(V,A)\right]_h(s,a) \le [\mathscr{R}^\star(V,A)]_h(s,a)$$

for all $(V,A) \in \Theta \subset \overline{\mathcal{V}} \times \overline{\mathcal{A}}$ and $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, as long as the number of episodes

$$K \ge \widetilde{\mathcal{O}}\left(\frac{H^4 SC^\star \log \mathcal{N}}{\epsilon^2} + \frac{H^2 SC^\star \eta}{\epsilon}\right).$$

Above, $\log \mathcal{N} := \log \mathcal{N}(\Theta; \epsilon/H)$, $\eta := \Delta^{-1}\mathbf{1}\{\text{option 1}\}$, and $\widetilde{\mathcal{O}}(\cdot)$ hides $\mathrm{polylog}(H,S,A,1/\delta)$ factors.

To our best knowledge, Theorem 8 provides the first theoretical guarantee for reward-set estimation under the standard offline setting, showing that RLP achieves the desired monotonicity condition and small D_Θ^π distance for any evaluation policy π^{eval} that satisfies single-policy concentrability with respect to π^{b}. For small enough ϵ, the sample complexity (number of episodes required) scales as $\widetilde{\mathcal{O}}(H^4 SC^\star \log \mathcal{N}/\epsilon^2)$, which depends on the number of states S, the concentrability coefficient $C^\star$, as well as the log-covering number $\log \mathcal{N}$ which always admits the bound $\log \mathcal{N} \le \widetilde{\mathcal{O}}(S)$ in the worst case and may be smaller.

Apart from the $\log \mathcal{N}$ factor, this rate resembles that of standard offline RL under single-policy concentrability (Rashidinejad et al., 2021; Xie et al., 2021). This is no coincidence, as our algorithm and proof (for both the $D_\Theta^{\pi^{\mathsf{eval}}}$ bound and the monotonicity condition) can be viewed as an adaptation of the pessimism technique for all rewards $(\mathscr{R}(V,A))_{(V,A)\in\Theta}$ simultaneously, demonstrating that IRL is "no harder than standard RL" in this setting. We remark that the Δ^{-1} factor brought by Assumption A appears only in the $\widetilde{\mathcal{O}}(\epsilon^{-1})$ burn-in term in the rate when the feedback $\{e_h^k\}_{k,h}$ in (10) comes from option 1.

Result for $\pi^{\mathsf{eval}} = \pi^{\mathsf{E}}$ In the special case where $\pi^{\mathsf{eval}} = \pi^{\mathsf{E}}$, we establish a slightly stronger result where we can improve over Theorem 8 by one H factor ($H^4 \to H^3$) in the main term. The proof uses the specific form of our Bernstein-like bonus (8) combined with a total variance argument (Azar et al., 2017; Zhang et al., 2020; Xie et al., 2021), and can be found in Appendix E.3.

Theorem 9 (Improved sample complexity for $\pi^{\mathsf{eval}} = \pi^{\mathsf{E}}$)**.** *Suppose $\pi^{\mathsf{eval}} = \pi^{\mathsf{E}}$ which achieves $C^\star$ single-policy concentrability with respect to π^{b} (Assumption B), and in addition $\sup_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} |[\mathscr{R}^\star(V,A)]_h(s,a)| \le 1$ for all $(V,A) \in \Theta \subset \overline{\mathcal{V}} \times \overline{\mathcal{A}}$. Then under both options in (10), with probability at least $1-\delta$,* RLP *(Algorithm 1) achieves the same guarantee as in Theorem 8 ($D_\Theta^{\pi^{\mathsf{eval}}}(\mathscr{R}^\star, \widehat{\mathscr{R}}) \le \epsilon$ and monotonicity), as long as the number of episodes*

$$K \ge \widetilde{\mathcal{O}}\left(\frac{H^3 SC^\star \log \mathcal{N}}{\epsilon^2} + \frac{H^2 SC^\star (A + H\log \mathcal{N})}{\epsilon}\right).$$

Theorem 9 no longer requires well-posedness of π^{E} (Assumption A) in option 1. This happens due to the assumed concentrability between $\pi^{\mathsf{E}} (= \pi^{\mathsf{eval}})$ and π^{b}, which can aid the learning of $\mathrm{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)$ even without well-posedness.

IRL from full expert trajectories An important special case of Theorem 9 is when π^{b} further coincides with π^{E}. This represents a natural and clean setting where dataset $\mathcal{D}$ consists of full trajectories drawn from the expert policy π^{E}, and our goal is to recover a reward mapping with a small $D_\Theta^{\pi^{\mathsf{E}}}$. This case is covered by Theorem 9 by taking $C^\star = 1$ and admits a sample complexity $\widetilde{\mathcal{O}}(H^3 S \log \mathcal{N}/\epsilon^2)$.

4.4. Lower bound

We present an information-theoretic lower bound showing that the upper bound in Theorem 8 is nearly tight.

Theorem 10 (Informal version of Theorem H.2)**.** *For any (H,S,A,ϵ) and any $C^\star \ge 1$, there exists a family of offline IRL problems where $\mathcal{D}$ consists of K episodes, π^{eval} satisfies $C^\star$-concentrability at most $C^\star$, $\Theta = \overline{\mathcal{V}} \times \overline{\mathcal{A}}$, and π^{E} is Δ well-posed with $\Delta = 1$, such that the following holds.*

Suppose any IRL algorithm achieves $D_\Theta^{\pi^{\mathsf{eval}}}(\mathscr{R}^\star, \widehat{\mathscr{R}}) \le \epsilon$ for every problem in this family with probability at least $2/3$, then we must have $K \ge \Omega\big(H^2 SC^\star \min\{S,A\}/\epsilon^2\big)$.

For $\Theta = \overline{\mathcal{V}} \times \overline{\mathcal{A}}$, the upper bound in Theorem 8 scales as $\widetilde{\mathcal{O}}(H^4 S^2 C^\star/\epsilon^2)$. Ignoring H and polylogarithmic factors, Theorem 10 assert that this rate is tight for $S \le A$ (so that $\min\{S,A\} = S$). The form of this $\min\{S,A\}$ factor in Theorem 10 is due to certain technicalities in the hard instance construction; whether this can be improved to an S factor would be an interesting question for future work.

5. IRL in the online setting

5.1. Setting

We now consider IRL in a natural online learning setting (also known as "active exploration IRL" (Lindner et al., 2023)). In each episode, the learner interacts with the IRL problem $(\mathcal{M}, \pi^{\mathsf{E}})$ as follows: At each $h \in [H]$, the learner receives the state $s_h \in \mathcal{S}$ and chooses their action $a_h \in \mathcal{A}$ from an arbitrary policy. The environment then provides the expert feedback e_h as in (10) (from one of the two options) and transits to the next state $s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h)$. This

Algorithm 2 REWARD LEARNING WITH EXPLORATION

```
1: Input: Parameter set $\Theta \subseteq \overline{\mathcal{V}} \times \overline{\mathcal{A}}$, confidence level $\delta > 0$, error tolerance $\epsilon > 0$, $N, K \in \mathbb{Z}_{\geq 0}$, threshold $\xi = c_\xi H^3 S^3 A^3 \log \frac{10HSA}{\delta}$.
2: Call Algorithm 3 to play in the environment for $NH$ episodes and obtain an explorative behavior policy $\pi^{\mathsf{b}}$.
3: Collect a dataset $\mathcal{D} = \{(s_h^k, a_h^k, e_h^k)\}_{k=1,h=1}^{K,H}$ by executing $\pi^{\mathsf{b}}$ in $\mathcal{M}$.
4: Subsampling: subsample $\mathcal{D}$ to obtain $\mathcal{D}^{\mathrm{trim}}$, such that for each $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, $\mathcal{D}^{\mathrm{trim}}$ contains $\min\left\{\widehat{N}_h^{\mathsf{b}}(s,a), N_h(s,a)\right\}$
   sample transitions randomly drawn from $\mathcal{D}$, where $\widehat{N}_h^{\mathsf{b}}(s,a)$ and $N_h(s,a)$ are defined by
   $N_h(s,a) := \sum_{k=1}^{K} \mathbf{1}\left\{(s_h^k, a_h^k) = (s,a)\right\} \quad \widehat{N}_h^{\mathsf{b}}(s,a) := \min\left[\frac{K}{4}, \mathbb{E}_{\pi\sim\mu^{\mathsf{b}}}[\widehat{d}_h^{\pi}(s,a)] - \frac{K\xi}{8N} - 3\log\frac{10HSA}{\delta}\right]_+, \quad (13)$
   where $\widehat{d}_h^{\pi}(s,a)$ is specified in Algorithm 3.
5: Call RLP (Algorithm 1) on dataset $\mathcal{D}^{\mathrm{trim}}$ with parameters $(\Theta, \delta/10, \epsilon/10)$ to compute the recovered reward mapping $\widehat{\mathscr{R}}$.
6: Output: Estimated reward mapping $\widehat{\mathscr{R}}$.
```

setting shares the same expert feedback model (e_h) with the offline setting, and differs in that the learner can interact with the environment, instead of learning from a fixed dataset pre-collected by some fixed behavior policy. We note that classic IRL work does not require assuming the ability to query the expert online; however, prior work (Lindner et al., 2023) on reward-set estimation for IRL has considered this setting.

5.2. Algorithm and guarantee

Our algorithm REWARD LEARNING WITH EXPLORATION (RLE; Algorithm 2) performs IRL in the online setting by a simple reduction to reward-free learning and the RLP algorithm. RLE consists of two main steps: (1) Call a reward-free exploration subroutine (Algorithm 3, building on the algorithm of Li et al. (2023)) to explore the environment $\mathcal{M}$ and obtain an explorative behavior policy π^{b} (Line 2); (2) Collect K episodes of data $\mathcal{D}$ using π^{b}, subsample the data, and call the RLP algorithm on the subsampled data $\mathcal{D}^{\mathrm{trim}}$ to obtain the estimated reward mapping $\widehat{\mathscr{R}}$.

We now present the theoretical guarantee of RLE. The proof can be found in Appendix F.2.

Theorem 11 (Sample complexity of RLE)**.** *Suppose π^{E} is Δ-well-posed (Assumption A) when we receive feedback in option 1 of (10). Then for the online setting, for sufficiently small $\epsilon \leq H^{-9}(SA)^{-6}$, with probability at least $1-\delta$,* RLE *(Algorithm 2) with $N = \widetilde{\mathcal{O}}(\sqrt{H^9 S^7 A^7 K})$ outputs a reward mapping $\widehat{\mathscr{R}}$ such that*

$$D_\Theta^{\mathsf{all}}\left(\mathscr{R}^\star, \widehat{\mathscr{R}}\right) \leq \epsilon, \left[\widehat{\mathscr{R}}(V,A)\right]_h(s,a) \leq [\mathscr{R}^\star(V,A)]_h(s,a)$$

for all $(V,A) \in \Theta$ and $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, as long as the total the number of episodes

$$K + NH \geq \widetilde{\mathcal{O}}\left(\frac{H^4 SA \log\mathcal{N}}{\epsilon^2} + \frac{H^2 SA\eta}{\epsilon}\right).$$

Above, $\log\mathcal{N} := \log\mathcal{N}(\Theta; \epsilon/H)$*, $\eta := \Delta^{-1}\mathbf{1}\{\text{option } 1\}$, and $\widetilde{\mathcal{O}}(\cdot)$ hides* $\mathrm{polylog}(H,S,A,1/\delta)$ *factors.*

For small enough ϵ, RLE requires $\widetilde{\mathcal{O}}(H^4 SA \log\mathcal{N}/\epsilon^2)$ episdoes for finding $\mathscr{R}$ with $D_\Theta^{\mathsf{all}}(\mathscr{R}^\star, \widehat{\mathscr{R}}) \leq \epsilon$. Compared with the offline setting (Theorem 8), the main differences here are that the metric is stronger (D_Θ^{all} versus $D_\Theta^{\pi^{\mathsf{eval}}}$ therein), and that the concentrability coefficient $C^\star$ in the sample complexity is replaced with the number of actions A. This is because using online interaction, our reward-free exploration subroutine (Algorithm 3) can find a policy π^{b} that achieves a form of "single-policy concentrability" A with respect to any policy π; see (16).

To our best knowledge, the only existing work that studies IRL in the same online setting is Lindner et al. (2023), who also achieve a sample complexity[3] of $\widetilde{\mathcal{O}}(H^4 S^2 A/\epsilon^2 + H^2 SA\eta/\epsilon)$ (for $\Theta = \overline{\mathcal{V}} \times \overline{\mathcal{A}}$) in their performance metric D^L (cf. (23)). However, our metric D_Θ^{all} is stronger than their D^L and avoids certain indistinguishability issues of theirs, as we have shown in Theorem 7.

5.3. Lower bound

We also provide a lower bound for IRL in the online setting in the D_Θ^{all} metric. The rate of the lower bound is similar to Theorem 10, and ensures that the rate in Theorem 11 is tight up to H and polylogarithmic factors when $S \leq A$.

Theorem 12 (Informal version of Theorem G.2)**.** *For any (H,S,A,ϵ), there exists a family of online IRL problems where $\Theta = \overline{\mathcal{V}} \times \overline{\mathcal{A}}$, and π^{E} is Δ well-posed with $\Delta = 1$, such that the following holds. Suppose any IRL algorithm achieves $D_\Theta^{\mathsf{all}}(\mathscr{R}^\star, \widehat{\mathscr{R}}) \leq \epsilon$ for every problem in this family with probability at least $2/3$, then we must have $K \geq \Omega\big(H^3 SA \min\{S,A\}/\epsilon^2\big)$.*

[3] Extracted from the proof of Lindner et al. (2023, Theorem 8) and taking into account the uniform convergence over $\overline{\mathcal{V}} \times \overline{\mathcal{A}}$ and dependence on $\eta = \Delta^{-1}\mathbf{1}\{\text{option } 1\}$; cf. Appendix D.1.

6. Transfer learning

As a further application, we consider a transfer learning setting, where rewards learned in a source MDP\R are transferred to a target MDP\R (possibly different from the source MDP\R). Inspired by the single-policy concentrability assumption, we define two concepts called *weak-transferability* and *transferability* (Definition I.1 & I.2) that measure the similarity between two MDP\R's.

We show that when the target MDP\R exhibits a small week-transferability (transferability) with respect to the source MDP\R, our algorithms RLP and RLE can perform IRL with sample complexity polynomial in these transferability coefficients and other problem parameters (Theorem I.3 & I.4), and provide guarantees for performing RL algorithms with the learned rewards in the target environments (Corollary I.5 & I.6). We defer the detailed setups and results to Appendix I due to the space limit.

7. Conclusion

This paper designs the first provably sample-efficient algorithm for inverse reinforcement learning (IRL) in the offline setting. Our algorithms and analyses seamlessly adapt the pessimism principle in standard offline RL, and we also extend it to an online setting by a simple reduction aided by reward-free exploration. We believe our work opens up many important questions, such as generalization to function approximation settings and empirical verifications.

Acknowledgment

The authors thank David Lindner for the many valuable discussions. Mengdi Wang acknowledges the support by NSF IIS-2107304, NSF CPS-2312093, ONR 1006977 and Genmab.

Impact statement

As the contributions of this paper are primarily theoretical, we don't foresee any tangible societal impact directly. Our broader line of inquiry could impact a line of thinking about designing more sample-efficient algorithms for inverse reinforcement learning, which could be useful towards making such practice more resource and energy efficient.

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.

Agarwal, A., Jiang, N., and Kakade, S. M. *Reinforcement learning: Theory and algorithms*. MIT, 2020.

Argall, B. D., Chernova, S., Veloso, M., and Browning, B. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

Arora, S., Du, S., Kakade, S., Luo, Y., and Saunshi, N. Provable representation learning for imitation learning via bi-level optimization. In *International Conference on Machine Learning*, pp. 367–376. PMLR, 2020.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. *arXiv preprint arXiv:1703.05449*, 2017.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Bain, M. and Sammut, C. A framework for behavioural cloning. In *Machine Intelligence 15*, pp. 103–129, 1995.

Barnes, M., Abueg, M., Lange, O. F., Deeds, M., Trader, J., Molitor, D., Wulfmeier, M., and O'Banion, S. Massively scalable inverse reinforcement learning in google maps. *arXiv preprint arXiv:2305.11290*, 2023.

Boularias, A., Kober, J., and Peters, J. Relative entropy inverse reinforcement learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 182–189. JMLR Workshop and Conference Proceedings, 2011.

Brantley, K., Sun, W., and Henaff, M. Disagreement-regularized imitation learning. In *International Conference on Learning Representations*, 2019.

Brown, D., Goo, W., Nagarajan, P., and Niekum, S. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, pp. 783–792. PMLR, 2019.

Chadi, M.-A. and Mousannif, H. Inverse reinforcement learning for healthcare applications: A survey. 2022.

Chang, J. D., Uehara, M., Sreenivas, D., Kidambi, R., and Sun, W. Mitigating covariate shift in imitation learning via offline data without great coverage. *arXiv preprint arXiv:2106.03207*, 2021.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Dexter, G., Bello, K., and Honorio, J. Inverse reinforcement learning in a continuous state space with formal guarantees. *Advances in Neural Information Processing Systems*, 34:6972–6982, 2021.

Ding, Z., Chen, Y., Ren, A. Z., Gu, S. S., Dong, H., and Jin, C. Learning a universal human prior for dexterous manipulation from human preference. *arXiv preprint arXiv:2304.04602*, 2023.

Finn, C., Levine, S., and Abbeel, P. Guided cost learning: Deep inverse optimal control via policy optimization. In *International conference on machine learning*, pp. 49–58. PMLR, 2016.

Fu, J., Luo, K., and Levine, S. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

Gleave, A., Dennis, M., Legg, S., Russell, S., and Leike, J. Quantifying differences in reward functions. *arXiv preprint arXiv:2006.13900*, 2020.

Gong, W., Cao, L., Zhu, Y., Zuo, F., He, X., and Zhou, H. Federated inverse reinforcement learning for smart icus with differential privacy. *IEEE Internet of Things Journal*, 2023.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Hantous, K., Rejeb, L., and Hellali, R. Detecting physiological needs using deep inverse reinforcement learning. *Applied Artificial Intelligence*, 36(1):2022340, 2022.

Ho, J. and Ermon, S. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., and Amodei, D. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31, 2018.

Jain, A., Wojcik, B., Joachims, T., and Saxena, A. Learning trajectory preferences for manipulators via iterative improvement. *Advances in neural information processing systems*, 26, 2013.

Jara-Ettinger, J. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29: 105–110, 2019.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pp. 4870–4879. PMLR, 2020.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. MOReL: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*, 2020.

Kretzschmar, H., Spies, M., Sprunk, C., and Burgard, W. Socially compliant mobile robot navigation via inverse reinforcement learning. *The International Journal of Robotics Research*, 35(11):1289–1307, 2016.

Kuderer, M., Gulati, S., and Burgard, W. Learning driving styles for autonomous vehicles from demonstration. In *2015 IEEE international conference on robotics and automation (ICRA)*, pp. 2641–2646. IEEE, 2015.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative Q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

Kumar, S., Zamora, J., Hansen, N., Jangir, R., and Wang, X. Graph inverse reinforcement learning from diverse videos. In *Conference on Robot Learning*, pp. 55–66. PMLR, 2023.

Levine, S., Popovic, Z., and Koltun, V. Nonlinear inverse reinforcement learning with gaussian processes. *Advances in neural information processing systems*, 24, 2011.

Li, G., Yan, Y., Chen, Y., and Fan, J. Minimax-optimal reward-agnostic exploration in reinforcement learning, 2023.

Lindner, D., Krause, A., and Ramponi, G. Active exploration for inverse reinforcement learning, 2023.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Metelli, A. M., Ramponi, G., Concetti, A., and Restelli, M. Provably efficient learning of transferable rewards. In *International Conference on Machine Learning*, pp. 7665–7676. PMLR, 2021.

Metelli, A. M., Lazzati, F., and Restelli, M. Towards theoretical understanding of inverse reinforcement learning, 2023.

Nachum, O. and Yang, M. Provable representation learning for imitation with contrastive fourier features. *Advances in Neural Information Processing Systems*, 34:30100–30112, 2021.

Ng, A. Y. and Russell, S. J. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 663–670, 2000.

Novoseller, E., Wei, Y., Sui, Y., Yue, Y., and Burdick, J. Dueling posterior sampling for preference-based reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1029–1038. PMLR, 2020.

Okal, B. and Arras, K. O. Learning socially normative robot navigation behaviors with bayesian inverse reinforcement learning. In *2016 IEEE international conference on robotics and automation (ICRA)*, pp. 2889–2895. IEEE, 2016.

OpenAI. Gpt-4 technical report, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Pacchiano, A., Saha, A., and Lee, J. Dueling rl: reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*, 2021.

Pan, Y., Cheng, C.-A., Saigol, K., Lee, K., Yan, X., Theodorou, E., and Boots, B. Agile autonomous driving using end-to-end deep imitation learning. *arXiv preprint arXiv:1709.07174*, 2017.

Pan, Y., Cheng, C.-A., Saigol, K., Lee, K., Yan, X., Theodorou, E. A., and Boots, B. Imitation learning for agile autonomous driving. *The International Journal of Robotics Research*, 39(2-3):286–302, 2020.

Qureshi, A. H., Boots, B., and Yip, M. C. Adversarial imitation via variational inverse reinforcement learning. *arXiv preprint arXiv:1809.06404*, 2018.

Rajaraman, N., Yang, L., Jiao, J., and Ramchandran, K. Toward the fundamental limits of imitation learning. *Advances in Neural Information Processing Systems*, 33: 2914–2924, 2020.

Ramachandran, D. and Amir, E. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pp. 2586–2591, 2007.

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

Ratliff, N. D., Bagnell, J. A., and Zinkevich, M. A. Maximum margin planning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 729–736, 2006.

Ravichandar, H., Polydoros, A. S., Chernova, S., and Billard, A. Recent advances in robot learning from demonstration. *Annual review of control, robotics, and autonomous systems*, 3:297–330, 2020.

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

Ross, S., Gordon, G., and Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Sadigh, D., Dragan, A. D., Sastry, S., and Seshia, S. A. *Active preference-based learning of reward functions*. 2017.

Sun, W., Venkatraman, A., Gordon, G. J., Boots, B., and Bagnell, J. A. Deeply aggrevated: Differentiable imitation learning for sequential prediction. In *International conference on machine learning*, pp. 3309–3318. PMLR, 2017.

Swamy, G., Choudhury, S., Bagnell, J. A., and Wu, S. Of moments and matching: A game-theoretic framework for closing the imitation gap. In *International Conference on Machine Learning*, pp. 10022–10032. PMLR, 2021.

Swamy, G., Rajaraman, N., Peng, M., Choudhury, S., Bagnell, J., Wu, S. Z., Jiao, J., and Ramchandran, K. Minimax optimal online imitation learning via replay estimation. *Advances in Neural Information Processing Systems*, 35: 7077–7088, 2022.

Swamy, G., Wu, D., Choudhury, S., Bagnell, D., and Wu, S. Inverse reinforcement learning without reinforcement learning. In *International Conference on Machine Learning*, pp. 33299–33318. PMLR, 2023.

Syed, U. and Schapire, R. E. A game-theoretic approach to apprenticeship learning. *Advances in neural information processing systems*, 20, 2007.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Vasquez, D., Okal, B., and Arras, K. O. Inverse reinforcement learning algorithms and features for robot navigation in crowds: an experimental comparison. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1341–1346. IEEE, 2014.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Wang, Y., Liu, Q., and Jin, C. Is rlhf more difficult than standard rl? *arXiv preprint arXiv:2306.14111*, 2023.

Woodworth, B., Ferrari, F., Zosa, T. E., and Riek, L. D. Preference learning in assistive robotics: Observational repeated inverse reinforcement learning. In *Machine learning for healthcare conference*, pp. 420–439. PMLR, 2018.

Wu, Z., Sun, L., Zhan, W., Yang, C., and Tomizuka, M. Efficient sampling-based maximum entropy inverse reinforcement learning with application to autonomous driving. *IEEE Robotics and Automation Letters*, 5(4):5355–5362, 2020.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021.

Xie, T., Foster, D. J., Bai, Y., Jiang, N., and Kakade, S. M. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

Xu, T., Li, Z., and Yu, Y. Error bounds of imitating policies and environments. *Advances in Neural Information Processing Systems*, 33:15737–15749, 2020a.

Xu, Y., Wang, R., Yang, L., Singh, A., and Dubrawski, A. Preference-based reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 33:18784–18794, 2020b.

Yu, C., Ren, G., and Liu, J. Deep inverse reinforcement learning for sepsis treatment. In *2019 IEEE international conference on healthcare informatics (ICHI)*, pp. 1–3. IEEE, 2019.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., and Ma, T. MOPO: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

Zhan, W., Uehara, M., Kallus, N., Lee, J. D., and Sun, W. Provable offline reinforcement learning with human feedback. *arXiv preprint arXiv:2305.14816*, 2023.

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33:15198–15207, 2020.

Zhu, B., Jiao, J., and Jordan, M. I. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. *arXiv preprint arXiv:2301.11270*, 2023a.

Zhu, B., Sharma, H., Frujeri, F. V., Dong, S., Zhu, C., Jordan, M. I., and Jiao, J. Fine-tuning language models with advantage-induced policy alignment. *arXiv preprint arXiv:2306.02231*, 2023b.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., Dey, A. K., et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

A. Additional related work

Imitation learning A closely related field to IRL is Imitation Learning, which focuses on learning policies from demonstrations, in contrast to IRL's emphasis on learning rewards from expert demonstrations (Bain & Sammut, 1995; Abbeel & Ng, 2004; Ratliff et al., 2006; Ziebart et al., 2008; Pan et al., 2017; Finn et al., 2016). Imitation learning has been extensively studied in the active setting (Ross et al., 2011; Ross & Bagnell, 2014; Sun et al., 2017), and theoretical analyses for Imitation Learning have been provided by Rajaraman et al. (2020); Xu et al. (2020a); Chang et al. (2021). More recently, the concept of Representation Learning for Imitation Learning has gained considerable attention (Arora et al., 2020; Nachum & Yang, 2021). While Imitation learning can be implemented by IRL (Abbeel & Ng, 2004; Ratliff et al., 2006; Ziebart et al., 2008), it is important to note that IRL has wider capabilities than Imitation Learning since the rewards learned through IRL can be transferred across different environments (Levine et al., 2011; Fu et al., 2017).

Reinforcement learning from human feedback Reinforcement Learning from Human Feedback (RLHF) bears a close relation to IRL, particularly because the process of learning rewards is a crucial aspect of both approaches (Zhu et al., 2023a;b; Wang et al., 2023; Zhan et al., 2023). RLHF has been successfully applied in various domains, including robotics (Jain et al., 2013; Sadigh et al., 2017; Ding et al., 2023) and game playing (Ibarz et al., 2018). Recently, RLHF has attracted considerable attention due to its remarkable capability to integrate human knowledge with large language models (Ouyang et al., 2022; OpenAI, 2023). Furthermore, the theoretical foundations of RLHF have been extensively developed in both tabular and function approximation settings (Zhan et al., 2023; Xu et al., 2020b; Pacchiano et al., 2021; Novoseller et al., 2020; Zhu et al., 2023a; Wang et al., 2023).

B. Technical tools

Lemma B.1 (Xie et al. (2021)). *Suppose $N \sim \mathrm{Bin}(n, p)$ where $n \geq 1$ and $p \in [0, 1]$. Then with probability at least $1 - \delta$, we have*

$$\frac{p}{N \vee 1} \leq \frac{8 \log(1/\delta)}{n}.$$

Theorem B.2 (Metelli et al. (2023)). *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$\mathbb{P}(A) + \mathbb{Q}(A^c) \geq \frac{1}{2} \exp\left(-D_{KL}(\mathbb{P}, \mathbb{Q})\right),$$

where $A^c = \Omega \setminus \mathcal{A}$ is the complement of A.

Theorem B.3 (Metelli et al. (2023)). *Let $\mathbb{P}_0, \mathbb{P}_1, \ldots, \mathbb{P}_M$ be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A_1, \ldots, A_M \in \mathcal{F}$ be a partition of Ω. Then,*

$$\frac{1}{M} \sum_{i=1}^{M} \mathbb{P}_i(A_i^c) \geq 1 - \frac{\frac{1}{M} \sum_{i=1}^{M} D_{KL}(\mathbb{P}_i, \mathbb{P}_0) - \log 2}{\log M},$$

where $A^c = \Omega \setminus A$ is the complement of A.

C. Useful algorithmic subroutines from prior works

In this section, we give the algorithm procedures of finding behavior policy π^{b} in Algorithm 2. The algorithm procedures are directly quoted from Li et al. (2023), with slight modification.

C.1. Algorithm: finding behavior policy π^{b}

Algorithm 3, a component of Li et al. (2023, Algorithm 1), aims to identify a suitable behavior policy. This is achieved by estimating the occupancy distribution d^{π}, which is induced by any deterministic policy π, through a meticulously designed exploration strategy. At each stage h, Algorithm 3 invokes Algorithm procedure 4 to compute an appropriate exploration policy, denoted as $\pi^{\mathsf{explore},h}$, and subsequently collects N sample trajectories by executing $\pi^{\mathsf{explore},h}$. These steps facilitate the estimation of the occupancy distribution d^{π}_{h+1} for the next stage $h+1$. Finally, the behavior policy $\pi^{\mathsf{b}} \sim \mu_{\mathsf{b}}$ is computed by invoking Algorithm 5.

Algorithm 3 Subroutine for computing behavior policy (Li et al., 2023)

```
Input: state space $\mathcal{S}$, action space $\mathcal{A}$, horizon length $H$, initial state distribution $\rho$, target success probability $1-\delta$, threshold $\xi = c_\xi H^3 S^3 A^3 \log(HSA/\delta)$.
Draw $N$ i.i.d. initial states $s_1^{n,0} \overset{\text{i.i.d.}}{\sim} \rho$ $(1 \le n \le N)$, and define the following functions
    $\widehat{d}_1^{\pi}(s) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\{s_1^{n,0} = s\}, \qquad \widehat{d}_1^{\pi}(s,a) = \widehat{d}_1^{\pi}(s) \cdot \pi_1(a|s)$    (14)
for any deterministic policy $\pi : [H] \times \mathcal{S} \to \Delta(\mathcal{A})$ and any $(s,a) \in \mathcal{S} \times \mathcal{A}$.
for $h = 1, ..., H-1$ do
    Call Algorithm 4 to compute an exploration policy $\pi^{\mathsf{explore},h}$.
    Draw $N$ independent trajectories $\{s_1^{n,h}, a_1^{n,h}, \dots, s_{h+1}^{n,h}\}_{1\le n\le N}$ using policy $\pi^{\mathsf{explore},h}$ and compute
        $\widehat{\mathbb{P}}_h(s^{'}|s,a) = \frac{\mathbf{1}\{N_h(s,a) > \xi\}}{\max\{N_h(s,a), 1\}} \sum_{n=1}^{N} \mathbf{1}\left\{s_h^{n,h} = s, a_h^{n,h} = a, s_{h+1}^{n,h} = s^{'}\right\}, \qquad \forall (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S},$
    where $N_h(s,a) = \sum_{n=1}^{N} \mathbf{1}\left\{s_h^{n,h} = s, a_h^{n,h} = a\right\}$.
    For any deterministic policy $\pi : \mathcal{S} \times [H] \to \Delta(\mathcal{A})$ and any $(s,a) \in \mathcal{S} \times \mathcal{A}$, define
        $\widehat{d}_{h+1}^{\pi}(s) = \langle \widehat{\mathbb{P}}_h(s|\cdot,\cdot), \widehat{d}_h^{\pi}(\cdot,\cdot) \rangle, \qquad \widehat{d}_{h+1}^{\pi}(s,a) = \widehat{d}_{h+1}^{\pi}(s) \cdot \pi_{h+1}(a|s).$    (15)
end for
Call Algorithm 5 to compute a behavior policy $\pi^{\mathsf{b}}$.
Output: the behavior policy $\pi^{\mathsf{b}}$.
```

We highlight that the behavior policy distribution μ_{b} output by Algorithm 3 has following property Li et al. (2023)

$$\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\widehat{d}_h^{\pi}(s,a)}{\mathbb{E}_{\pi'\sim\mu_{\mathsf{b}}}\left[\widehat{d}_h^{\pi'}(s,a)\right]} \lesssim HSA, \tag{16}$$

for any deterministic policy $\pi \in \Pi^{\mathsf{det}}$.

C.2. Subroutine: computing exploration policy $\pi^{\mathsf{explore},h}$

We proceed to describe Algorithm 4, originally proposed in Li et al. (2023, Algorithm 3), which is designed to compute the desired exploration policy $\pi^{\mathsf{explore},h}$. At a high level, this algorithm calculates the exploration policy by approximately solving the subsequent optimization sub-problem, utilizing the Frank-Wolfe algorithm:

$$\widehat{\mu}^h \approx \arg\max_{\mu\in\Delta(\Pi)} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \log\left[\frac{1}{KH} + \mathop{\mathbb{E}}_{\pi\sim\mu}\left[\widehat{d}_h^{\pi}(s,a)\right]\right], \tag{17}$$

Here $\mathcal{M}_{\mathsf{b}}^h = (\mathcal{S} \cup \{s^{\mathsf{aug}}\}, \mathcal{A}, H, \widehat{\mathbb{P}}^{\mathsf{aug},h}, r_{\mathsf{b}}^h)$, where s_{aug} is an augmented state as before, and the reward function is chosen to be

$$r_{\mathsf{b},j}^h(s,a) = \begin{cases} \frac{1}{\frac{1}{KH} + \mathbb{E}_{\pi\sim\mu^{(t)}}\left[\widehat{d}_h^{\pi}(s,a)\right]} \in [0, KH], & \text{if } (s,a,j) \in \mathcal{S} \times \mathcal{A} \times \{h\}; \\ 0, & \text{if } s = s_{\mathsf{aug}} \text{ or } j \neq h. \end{cases} \tag{18}$$

Algorithm 4 Subroutine for solving Eq.(17) (Li et al., 2023).

```
Initialize: $\mu^{(0)} = \delta_{\pi_{\mathsf{init}}}$ for an arbitrary policy $\pi_{\mathsf{init}} \in \Pi$, $T_{\max} = \lfloor 50SA\log(KH)\rfloor$.
for $t = 0, 1..., T_{\max}$ do
    Compute the optimal deterministic policy $\pi^{(t),\mathsf{b}}$ of the MDP $\mathcal{M}_{\mathsf{b}}^{h} = (\mathcal{S} \cup \{s_{\mathsf{aug}}\}, \mathcal{A}, H, \widehat{\mathbb{P}}^{\mathsf{aug},h}, r_{\mathsf{b}}^{h})$, where $r_{\mathsf{b}}^{h}$ is defined in Eq.(18), and $\widehat{\mathbb{P}}^{\mathsf{aug},h}$ is defined in Eq.(19); let $\pi^{(t)}$ be the corresponding optimal deterministic policy of $\pi^{(t),\mathsf{b}}$ in the original state space.
    Compute
        $\alpha_t = \frac{\frac{1}{SA}g(\pi^{(t)}, \widehat{d}, \mu^{(t)}) - 1}{g(\pi^{(t)}, \widehat{d}, \mu^{(t)}) - 1}, \quad \text{where} \quad g(\pi, \widehat{d}, \mu) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\frac{1}{KH} + \widehat{d}_h^{\pi}(s,a)}{\frac{1}{KH} + \mathbb{E}_{\pi\sim\mu}[\widehat{d}_h^{\pi}(s,a)]}.$
    Here, $\widehat{d}_h^{\pi}(s,a)$ is computed via Eq.(14) for $h = 1$, and Eq.(15) for $h \geq 2$.
    If $g(\pi^{(t)}, \widehat{d}, \mu^{(t)}) \leq 2SA$ then exit for-loop.
    Update
        $\mu^{(t+1)} = (1 - \alpha_t)\, \mu^{(t)} + \alpha_t\, \mathbb{1}_{\pi^{(t)}}\,.$
end for
Output: the exploration policy $\pi^{\mathsf{explore},h} = \mathbb{E}_{\pi\sim\mu^{(t)}}[\pi]$ and the weight $\widehat{\mu}^{h} = \mu^{(t)}$.
```

In addition, the augmented probability transition kernel $\widehat{\mathbb{P}}^{\mathsf{aug},h}$ is constructed based on $\widehat{\mathbb{P}}$ as follows:

$$\widehat{\mathbb{P}}_j^{\mathsf{aug},h}(s^{'} \mid s,a) = \begin{cases} \widehat{\mathbb{P}}_j(s^{'} \mid s,a), & \text{if } s^{'} \in \mathcal{S} \\ 1 - \sum_{s^{'}\in\mathcal{S}} \widehat{\mathbb{P}}_j(s^{'} \mid s,a), & \text{if } s^{'} = s_{\mathsf{aug}} \end{cases} \qquad \text{for all } (s,a,j) \in \mathcal{S}\times\mathcal{A}\times[h]; \tag{19a}$$

$$\widehat{\mathbb{P}}_j^{\mathsf{aug},h}(s^{'} \mid s,a) = \mathbb{1}(s^{'} = s_{\mathsf{aug}}) \qquad \text{if } s = s_{\mathsf{aug}} \text{ or } j > h. \tag{19b}$$

C.3. Subroutine: computing final behavior policy π^{b}

We proceed to describe Algorithm 5, originally proposed in Li et al. (2023, Algorithm 2), which is designed to compute the final behavior policy π^{b} $\pi^{\mathsf{explore},h}$, based on the estimated occupancy distributions specified in Algorithm 3. Algorithm 5 follows a similar fashion of Algorithm 4. Algorithm 5 computes the behavior policy by approximately solving the subsequent optimization sub-problem, utilizing the Frank-Wolfe algorithm:

$$\widehat{\mu}^{\mathsf{b}} \approx \arg\max_{\mu\in\Delta(\Pi)} \left\{ \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \log\left[\frac{1}{KH} + \mathbb{E}_{\pi\sim\mu}\left[\widehat{d}_h^{\pi}(s,a)\right]\right] \right\}. \tag{20}$$

Algorithm 5 Subroutine for solving Eq.(20) (Li et al., 2023).

```
Initialize: $\mu_{\mathsf{b}}^{(0)} = \delta_{\pi_{\mathsf{init}}}$ for an arbitrary policy $\pi_{\mathsf{init}} \in \Pi$, $T_{\max} = \lfloor 50SAH\log(KH)\rfloor$.
for $t = 0, 1..., T_{\max}$ do
    Compute the optimal deterministic policy $\pi^{(t),\mathsf{b}}$ of the MDP $\mathcal{M}_{\mathsf{b}} = (\mathcal{S} \cup \{s_{\mathsf{aug}}\}, \mathcal{A}, H, \widehat{\mathbb{P}}^{\mathsf{aug}}, r_{\mathsf{b}})$, where $r_{\mathsf{b}}$ is defined in Eq.(21), and $\widehat{\mathbb{P}}^{\mathsf{aug}}$ is defined in Eq.(22); let $\pi^{(t)}$ be the corresponding optimal deterministic policy of $\pi^{(t),\mathsf{b}}$ in the original state space.
    Compute
        $\alpha_t = \frac{\frac{1}{SAH}g(\pi^{(t)}, \widehat{d}, \mu_{\mathsf{b}}^{(t)}) - 1}{g(\pi^{(t)}, \widehat{d}, \mu_{\mathsf{b}}^{(t)}) - 1}, \quad \text{where} \quad g(\pi, \widehat{d}, \mu) = \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\frac{1}{KH} + \widehat{d}_h^{\pi}(s,a)}{\frac{1}{KH} + \mathbb{E}_{\pi\sim\mu}\left[\widehat{d}_h^{\pi}(s,a)\right]}.$
    Here, $\widehat{d}_h^{\pi}(s,a)$ is computed via Eq.(14) for $h = 1$, and Eq.(15) for $h \geq 2$.
    If $g(\pi^{(t)}, \widehat{d}, \mu_{\mathsf{b}}^{(t)}) \leq 2HSA$ then exit for-loop. Update
        $\mu_{\mathsf{b}}^{(t+1)} = (1 - \alpha_t)\, \mu_{\mathsf{b}}^{(t)} + \alpha_t \mathbf{1}_{\pi^{(t)}}.$
end for
Output: the behavior policy $\pi^{\mathsf{b}} = \mathbb{E}_{\pi\sim\mu_{\mathsf{b}}^{(t)}}[\pi]$ and the associated weight $\widehat{\mu}_b = \mu_{\mathsf{b}}^{(t)}$.
```

Here, $\mathcal{M}_{\mathsf{b}} = (\mathcal{S} \cup \{s_{\mathsf{aug}}\}, \mathcal{A}, H, \widehat{\mathbb{P}}^{\mathsf{aug}}, r_{\mathsf{b}})$, where s_{aug} is an augmented state and the reward function is chosen to be

$$r_{\mathsf{b},h}(s,a) = \begin{cases} \frac{1}{\frac{1}{KH} + \mathbb{E}_{\pi\sim\mu_{\mathsf{b}}^{(t)}}\left[\widehat{d}_h^{\pi}(s,a)\right]} \in [0, KH], & \text{if } (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]; \\ 0, & \text{if } (s,a,h) \in \{s_{\mathsf{aug}}\}\times\mathcal{A}\times[H]. \end{cases} \tag{21}$$

In addition, the augmented probability transition kernel $\widehat{\mathbb{P}}^{\mathsf{aug}}$ is constructed based on $\widehat{\mathbb{P}}$ as follows:

$$\widehat{\mathbb{P}}_h^{\mathsf{aug}}(s^{'} \mid s,a) = \begin{cases} \widehat{\mathbb{P}}_h(s^{'} \mid s,a), & \text{if } s^{'} \in \mathcal{S} \\ 1 - \sum_{s^{'}\in\mathcal{S}} \widehat{\mathbb{P}}_h(s^{'} \mid s,a), & \text{if } s^{'} = s_{\mathsf{aug}} \end{cases} \qquad \text{for all } (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]; \tag{22a}$$

$$\widehat{\mathbb{P}}_h^{\mathsf{aug}}(s^{'} \mid s_{\mathsf{aug}}, a) = \mathbb{1}(s^{'} = s_{\mathsf{aug}}) \qquad \text{for all } (a,h) \in \mathcal{A}\times[H]. \tag{22b}$$

It's evident that the augmented state behaves as an absorbing state, associated with zero immediate rewards.

D. Relationship with existing metrics

In Section D.1, we discuss the online IRL performance metric D^L proposed in Lindner et al. (2023), where we show that RLE is still efficient under this metric, yet D^L fails to distinguish certain pairs of reward mappings (or reward sets) that exhibit large distances under our metric. In Section D.2, we briefly discuss the existing IRL performance metric $d_{V^\star}^{G}$ used in the simulator setting (Metelli et al., 2021; 2023). In Section D.3, we provide a comparative analysis of our mapping-based metric in relation to Hausdorff-based metrics which is widely adopted by previous work (Metelli et al., 2021; 2023; Lindner et al., 2023). All proofs for this section can be found in Appendix D.4.

D.1. Discussion of existing metric for online IRL

Lindner et al. (2023) considers a performance metric between two IRL problems $\tau = (\mathcal{M}, \pi^{\mathsf{E}})$ and $\widehat{\tau} = (\widehat{\mathcal{M}}, \widehat{\pi}^{\mathsf{E}})$ instead of two reward mappings (or reward sets). Their metric D^L is defined as follows:

$$D^L(\tau,\widehat{\tau}) := \max\Bigg\{ \sup_{r\in\mathcal{R}_\tau} \inf_{\widehat{r}\in\mathcal{R}_{\widehat{\tau}}} \sup_{\widehat{\pi}^\star\in\Pi^\star_{\widehat{\mathcal{M}}\cup\widehat{r}}} \max_a \left| Q_1^{\pi^\star}(s_1,a;\mathcal{M}\cup r) - Q_1^{\widehat{\pi}^\star}(s_1,a;\mathcal{M}\cup r)\right|, \tag{23}$$

$$\sup_{\widehat{r}\in\mathcal{R}_{\widehat{\tau}}} \inf_{r\in\mathcal{R}_\tau} \sup_{\pi^\star\in\Pi^\star_{\mathcal{M}\cup r}} \max_a \left| Q_1^{\pi^\star}(s_1,a;\mathcal{M}\cup r) - Q_1^{\widehat{\pi}^\star}(s_1,a;\mathcal{M}\cup r)\right| \Bigg\}, \tag{24}$$

where the $\mathcal{R}_\tau, \mathcal{R}_{\widehat{\tau}}$ the set of all feasible rewards set for IRL problems $\tau, \widehat{\tau}$, respectively, $\pi^\star \in \Pi^\star_{\mathcal{M}\cup r}$, $\widehat{\pi}^\star \in \Pi^\star_{\widehat{\mathcal{M}}\cup\widehat{r}}$, and $Q_1^{\pi}(\cdot|\mathcal{M}\cup r)$ represent the Q-function induced by $\mathcal{M}\cup r$ and π. Since metric D^L is defined between two IRL problems, we can't directly compare D^L with our metrics. However, we can prove that our algorithm RLE is capable of achieving the goal of attaining a small D^L error.

Proposition D.1 (RLE achieves small D^L error)**.** *Denote the ground truth IRL problem as $\tau = (\mathcal{M}, \pi^{\mathsf{E}})$. Let $\widehat{\mathbb{P}}$ and $\widehat{\pi}^{\mathsf{E}}$ be the estimated expert policy and the estimated transition constructed by* RLE *(Algorithm 2), respectively. Define $\widehat{\tau} = \left(\widehat{\mathcal{M}}, \widehat{\pi}^{\mathsf{E}}\right)$, where $\widehat{\mathcal{M}}$ be the MDP\R equipped with the transition $\widehat{\mathbb{P}}$. Under the same assumptions and choice of parameters as in Theorem 11, for the online setting with both options in (10), for sufficiently small $\epsilon \leq H^{-9}(SA)^{-6}$, with probability at least $1-\delta$, we can ensure $D^L(\tau,\widehat{\tau}) \leq \epsilon$, as long as the total the number of episodes*

$$K + NH \geq \widetilde{\mathcal{O}}\left(\frac{H^4S^2A}{\epsilon^2} + \frac{H^2SA\eta}{\epsilon}\right).$$

Above, $\eta := \Delta^{-1}\mathbf{1}\{\text{option } 1\}$, and $\widetilde{\mathcal{O}}(\cdot)$ hides $\mathrm{polylog}(H,S,A,1/\delta)$ *factors.*

To achieve $D^L(\tau, \widehat{\tau}) \le \epsilon$, the sample complexity[4] of Lindner et al. (2023, Algorithm 1) is

$$\widetilde{\mathcal{O}}\bigg(\frac{H^4 S^2 A}{\epsilon^2} + \frac{H^2 S A \eta}{\epsilon}\bigg),$$

which exactly matches the sample complexity of RLE.

On the other hand, the following proposition shows that D^L cannot distinguish certain cases that our $D^{\mathsf{all}}_{\Theta}$ metric can.

Proposition D.2 (Example of problems distinguishable by $D^{\mathsf{all}}_{\Theta}$ but not D^L)**.** *Let $\Theta = \overline{\mathcal{V}} \times \overline{\mathcal{A}}$. There exist $\tau = (\mathcal{M}, \pi^{\mathsf{E}})$ and $\widehat{\tau} = (\widehat{\mathcal{M}}, \widehat{\pi^{\mathsf{E}}})$ such that $D^L(\tau, \widehat{\tau}) = 0$ but $D^{\mathsf{all}}_{\Theta}(\mathscr{R}^{\tau}, \mathscr{R}^{\widehat{\tau}}) \ge 1$ where $\mathscr{R}^{\tau}$ and $\mathscr{R}^{\widehat{\tau}}$ are reward mappings induced by τ and $\widehat{\tau}$ respectively using definition* (1)*.*

In fact, we also have $D^L(\tau, \widehat{\tau}) = 0$ whenever $\pi^{\mathsf{E}} = \widehat{\pi}^{\mathsf{E}}$ (but $\mathcal{M}$ and $\widehat{\mathcal{M}}$ may differ arbitrarily, which may induce arbitrary difference between $\mathscr{R}^{\tau}$ and $\mathscr{R}^{\widehat{\tau}}$).

D.2. Comparisons with existing metrics used in the simulator setting

Metelli et al. (2023) consider the following metric

$$d^{G}_{V^\star}(r, \widehat{r}) = \max_{\widehat{\pi} \in \Pi^\star_{\mathcal{M}\cup\widehat{r}}} \max_{(h,s)\in[H]\times\mathcal{S}} \Big| V^\star_h(s; r) - V^{\widehat{\pi}}_h(s; r) \Big|. \tag{25}$$

Notice the max over (h, s) in (25). In words, a small $d^{G}_{V^\star}(r, \widehat{r})$ requires r and $\widehat{r}$ to induce similar value functions *uniformly at all states*, which is achievable in their simulator setting and not achievable in standard offline/online settings where there may exist states that are not visitable at all by any policy in this particular MDP\R.

D.3. Comparison with Hausdorff-based metrics

Given a reward mapping $\mathscr{R} : \overline{\mathcal{V}} \times \overline{\mathcal{A}} \mapsto \mathcal{R}^{\mathsf{all}}$, we say a reward set $\mathcal{R} \subset \mathcal{R}^{\mathsf{all}}$ is a feasible reward set induced by $\mathscr{R}$, if $\mathcal{R} = \mathsf{image}(\mathscr{R})$. For any given base metric d between rewards, the Hausdorff (pre)metric D^{H} which is given by

$$D^{\mathsf{H}}(\mathcal{R}, \widehat{\mathcal{R}}) := \max\big\{ \sup_{r\in\mathcal{R}} \inf_{\widehat{r}\in\widehat{\mathcal{R}}} d(r, \widehat{r}), \sup_{\widehat{r}\in\widehat{\mathcal{R}}} \inf_{r\in\mathcal{R}} d(r, \widehat{r}) \big\}.$$

The works of Metelli et al. (2021; 2023) consider finding an estimated feasible set $\widehat{\mathcal{R}}$ that attains a small $D^{\mathsf{H}}(\mathcal{R}, \widehat{\mathcal{R}})$ using a certain base metric d.

Different from our mapping-based metric (Definition 5), the Hausdorff metric measures only the gap between the two sets $\mathcal{R}$ and $\widehat{\mathcal{R}}$, but cannot measure the gap between rewards for each parameter (V, A) in a *paired* fashion. Here we show that for any given base metric d, our mapping-based metric is stronger than the Hausdorff metric.

Lemma D.3 (D^{M} is stronger than D^{H})**.** *Given an IRL problem $(\mathcal{M}, \pi^{E})$ and a base metric $d : \mathcal{R}^{\mathsf{all}} \times \mathcal{R}^{\mathsf{all}} \mapsto \mathbb{R}_{\ge 0}$. We define the corresponding Hausdorff metric D^{H} for any reward set pair $(\mathcal{R}, \mathcal{R}')$ by*

$$D^{\mathsf{H}}(\mathcal{R}, \mathcal{R}') := \max\left\{ \sup_{r\in\mathcal{R}} \inf_{r'\in\mathcal{R}'} d(r, r'), \sup_{r'\in\mathcal{R}'} \inf_{r\in\mathcal{R}} d(r, r') \right\},$$

and the mapping-based metric D^{M} is defined for any reward mapping pair $(\mathscr{R}, \mathscr{R}')$ by

$$D^{\mathsf{M}}(\mathscr{R}, \mathscr{R}') := \sup_{V\in\overline{\mathcal{V}}, A\in\overline{\mathcal{A}}} d(\mathscr{R}(V, A), \mathscr{R}'(V, A)),$$

For any $(\mathscr{R}, \mathscr{R}')$, let $\mathcal{R} = \mathsf{image}(\mathscr{R})$ and $\mathcal{R}' = \mathsf{image}(\mathscr{R}')$, then we have

$$D^{\mathsf{H}}(\mathcal{R}, \mathcal{R}') \le D^{\mathsf{M}}(\mathscr{R}, \mathscr{R}').$$

[4] The original sample complexity given in Lindner et al. (2023) is $\widetilde{O}\Big(\frac{H^4 S A}{\epsilon^2}\Big)$. This is because, in the proof presented by Lindner et al. (2023), they didn't employ the uniform convergence argument. However, the uniform convergence result is necessary for proving the sample complexity of Lindner et al. (2023, Algorithm 1). As a result, an S factor was lost in the main term, and the burn-in term $\widetilde{O}\Big(\frac{H^2 S A \eta}{\epsilon}\Big)$ was neglected in their paper.

We then present the following lemma which demonstrates that for some d, D^{M} is *strictly* stronger than D^{H}.

Lemma D.4 (D^{M} is stronger than D^{H})**.** *There exists a base metric* d *defined on rewards such that for any IRL problem* $(\mathcal{M}, \pi^{\mathsf{E}})$*, there exists another IRL problem* $(\widehat{\mathcal{M}}, \widehat{\pi}^{\mathsf{E}})$ *such that* $D^{\mathsf{H}}(\mathcal{R}^\star, \widehat{\mathcal{R}}) = 0$*, but* $D^{\mathsf{M}}(\mathscr{R}^\star, \widehat{\mathscr{R}}) \geq 1/2$*, where* D^{H} *and* D^{M} *are the Hausdorff metric and mapping-based metric induced by* d*, respectively;* $\mathcal{R}^\star$ *and* $\widehat{\mathcal{R}}$ *are the feasible sets of* $(\mathcal{M}, \pi^{\mathsf{E}})$ *and* $(\widehat{\mathcal{M}}, \widehat{\pi}^{\mathsf{E}})$*, respectively;* $\mathscr{R}^\star$ *and* $\widehat{\mathscr{R}}$ *are the reward mappings induced by* $(\mathcal{M}, \pi^{\mathsf{E}})$ *and* $(\widehat{\mathcal{M}}, \widehat{\pi}^{\mathsf{E}})$*, respectively.*

D.4. Proofs for Section D

Proof of Proposition D.1. For any $\theta = (V, A)$, we first define the error $C_h^\theta(s,a)$ as follows:

$$C_h^\theta(s,a) := b_h^\theta(s,a) + \left|A_h(s,a) \cdot \left(\mathbf{1}\left\{a \in \mathrm{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\} - \mathbf{1}\left\{a \in \mathrm{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\}\right)\right|, \tag{26}$$

where $b_h^\theta(s,a)$ is defined in Eq.(8). Let $\mathscr{R}^\tau$ and $\mathscr{R}^{\widehat{\tau}}$ be the ground truth reward mappings induced by τ and $\widehat{\tau}$. We consider the concentration event $\mathcal{E}$ defined in Lemma F.2. Conditioning on $\mathcal{E}$, we next prove the following result:

$$C_h^\theta(s,a) \geq \max\left\{\left|r_h^{\tau,\theta}(s,a) - r^{\widehat{\tau},\theta}(s,a)\right|, \left|\left[\left(\mathbb{P}_h - \widehat{\mathbb{P}}_h\right)V_{h+1}\right](s,a)\right|\right\}, \tag{27}$$

holds for any $\theta \in \overline{\mathcal{V}} \times \overline{\mathcal{A}}$ and any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, where $r^{\tau,\theta} = \mathscr{R}^\tau(V,A)$, $r^{\widehat{\tau},\theta} = \mathscr{R}^{\widehat{\tau}}(V,A)$.

By Lemma F.2, under event $\mathcal{E}$, we directly have

$$C_h^\theta(s,a) \geq b_h^\theta(s,a) \geq \left|\left[\left(\mathbb{P}_h - \widehat{\mathbb{P}}_h\right)V_{h+1}\right](s,a)\right|, \tag{28}$$

holds for any $\theta \in \overline{\mathcal{V}} \times \overline{\mathcal{A}}$ and any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$. For the second part of Eq.(27), by definition of $r^{\tau,\theta}$ and $r^{\widehat{\tau},\theta}$, we have

$$\begin{aligned}
&\left|r_h^{\tau,\theta}(s,a) - r^{\widehat{\tau},\theta}(s,a)\right| = \Big| - A_h(s,a) \cdot \mathbf{1}\left\{a \notin \mathrm{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\} + V_h(s) - [\mathbb{P}_h V_{h+1}](s,a) \\
&+ A_h(s,a) \cdot \mathbf{1}\left\{a \notin \mathrm{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\} - V_h(s) + \left[\widehat{\mathbb{P}}_h V_{h+1}\right](s,a)\Big| \\
&\leq \left|A_h(s,a) \cdot \left(\mathbf{1}\left\{a \notin \mathrm{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\} - \mathbf{1}\left\{a \notin \mathrm{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\}\right)\right| + \left|\left(\mathbb{P}_h - \widehat{\mathbb{P}}_h\right)V_{h+1}(s,a)\right| \\
&\leq \left|A_h(s,a) \cdot \left(\mathbf{1}\left\{a \notin \mathrm{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\} - \mathbf{1}\left\{a \notin \mathrm{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\}\right)\right| + b_h^\theta(s,a) \\
&= C_h^\theta(s,a),
\end{aligned} \tag{29}$$

where the last line is by Lemma F.2. Combining Eq.(28) and Eq.(29), we compelete the proof for Eq.(27). When Eq.(27) holds, by Lindner et al. (2023, Lemma 20), there exists a policy π^L (see the proof of Lindner et al. (2023, Lemma 20) for the construction of π^L) such that

$$\begin{aligned}
D^L(\tau, \widehat{\tau}) &\lesssim \sup_{\theta \in \overline{\mathcal{V}} \times \overline{\mathcal{A}}} \sum_{h \in [H]} \sum_{s,a} d_h^{\pi^L}(s,a) \cdot C_h^\theta(s,a) \\
&= \sup_{\theta \in \overline{\mathcal{V}} \times \overline{\mathcal{A}}} \sum_{h \in [H]} \sum_{s,a} d_h^{\pi^L}(s,a) \cdot \left|A_h(s,a) \cdot \left(\mathbf{1}\left\{a \notin \mathrm{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\} - \mathbf{1}\left\{a \notin \mathrm{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\}\right)\right| \\
&\quad + \sup_{\theta \in \overline{\mathcal{V}} \times \overline{\mathcal{A}}} \sum_{h \in [H]} \sum_{s,a} d_h^{\pi^L}(s,a) \cdot b_h^\theta(s,a),
\end{aligned} \tag{30}$$

where $\left\{d_h^{\pi^L}(\cdot)\right\}_{h \in [H]}$ is the state-action visitation distribution induced by $\mathbb{P}$ and π^L. Following the proof of Theorem 11, we can prove that under $\mathcal{E}$

$$\begin{aligned}
&\sup_{\theta \in \overline{\mathcal{V}} \times \overline{\mathcal{A}}} \sum_{h \in [H]} \sum_{s,a} d_h^{\pi^L}(s,a) \cdot \left|A_h(s,a) \cdot \left(\mathbf{1}\left\{a \notin \mathrm{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\} - \mathbf{1}\left\{a \notin \mathrm{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\}\right)\right| \lesssim \epsilon, \\
&\sup_{\theta \in \overline{\mathcal{V}} \times \overline{\mathcal{A}}} \sum_{h \in [H]} \sum_{s,a} d_h^{\pi^L}(s,a) \cdot b_h^\theta(s,a) \lesssim \epsilon,
\end{aligned} \tag{31}$$

hold, provided that

$$K \geq \widetilde{\mathcal{O}}\Big(\frac{H^4 S^2 A}{\epsilon^2} + \frac{H^2 S A \eta}{\epsilon}\Big), \quad KH \geq N \geq \widetilde{\mathcal{O}}\Big(\sqrt{H^9 S^7 A^7 K}\Big).$$

Combining Eq.(30) and Eq.(31), we complete the proof.

□

Proof of Proposition D.2. To begin with, we prove a stronger result: for any $\tau = (\mathcal{M}, \pi^{\mathsf{E}})$ and any $\widehat{\tau} = (\widehat{\mathcal{M}}, \widehat{\pi}^{\mathsf{E}})$, if $\pi^{\mathsf{E}} = \widehat{\pi}^{\mathsf{E}}$, then $D^L(\tau, \widehat{\tau}) = 0$.

Let $\mathscr{R}^\tau$ and $\mathscr{R}^{\widehat{\tau}}$ be the reward mappings induced by τ and $\widehat{\tau}$, respectively. For any $\theta = (V, A)$, we define $r^{\tau,\theta} = \mathscr{R}^\tau(V, A)$ and $r^{\widehat{\tau},\theta} = \mathscr{R}^{\widehat{\tau}}(V, A)$. By the construction of reward mappings and the definition of optimal policies, we have that $\pi \in \Pi^\star_{\mathcal{M}\cup r}$ is equivalent to

$$A_h(s,a) \cdot \mathbf{1}\left\{\pi^{\mathsf{E}}_h(a|s) = 0\right\} = 0, \qquad \forall (h,s,a) \text{ s.t. } \pi_h(a|s) \neq 0. \tag{32}$$

Similarly, $\pi \in \Pi^\star_{\widehat{\mathcal{M}}\cup r^{\widehat{\tau},\theta}}$ is equivalent to

$$A_h(s,a) \cdot \mathbf{1}\left\{\widehat{\pi}^{\mathsf{E}}_h(a|s) = 0\right\} = A_h(s,a) \cdot \mathbf{1}\left\{\pi^{\mathsf{E}}_h(a|s) = 0\right\} = 0, \qquad \forall (h,s,a) \text{ s.t. } \pi_h(a|s) \neq 0. \tag{33}$$

Hence, we can conclude that $\Pi^\star_{\mathcal{M}\cup r^\theta} = \Pi^\star_{\widehat{\mathcal{M}}\cup r^{\widehat{\tau},\theta}}$. Notice that $r^{\tau,\theta} = \left\{r^{\tau,\theta} \,|\, \theta = (V,A)\right\}$ and $r^{\tau,\theta} = \left\{r^{\widehat{\tau},\theta} \,|\, \theta = (V,A)\right\}$, we then have

$$\begin{aligned}
&\sup_{r\in\mathcal{R}_\tau} \inf_{\widehat{r}\in\mathcal{R}_{\widehat{\tau}}} \sup_{\widehat{\pi}^\star\in\Pi^\star_{\widehat{\mathcal{M}}\cup\widehat{r}}} \max_a \left|Q_1^{\pi^\star}(s_1,a;\mathcal{M}\cup r) - Q_1^{\widehat{\pi}^\star}(s_1,a;\mathcal{M}\cup r)\right| \\
&= \sup_{\theta\in\Theta} \inf_{\theta'\in\Theta} \sup_{\widehat{\pi}^\star\in\Pi^\star_{\widehat{\mathcal{M}}\cup r^{\widehat{\tau},\theta'}}} \max_a \left|Q_1^{\pi^\star}(s_1,a;\mathcal{M}\cup r^{\tau,\theta}) - Q_1^{\widehat{\pi}^\star}(s_1,a;\mathcal{M}\cup r^{\tau,\theta})\right| \\
&= \sup_{\theta\in\Theta} \sup_{\widehat{\pi}^\star\in\Pi^\star_{\widehat{\mathcal{M}}\cup r^{\widehat{\tau},\theta}}} \max_a \left|Q_1^{\pi^\star}(s_1,a;\mathcal{M}\cup r^{\tau,\theta}) - Q_1^{\widehat{\pi}^\star}(s_1,a;\mathcal{M}\cup r^{\tau,\theta})\right| = 0,
\end{aligned} \tag{34}$$

where the last line is due to $\Pi^\star_{\mathcal{M}\cup r^\theta} = \Pi^\star_{\widehat{\mathcal{M}}\cup r^{\widehat{\tau},\theta}}$. Follow the same proof of Eq.(34), we have

$$\sup_{\widehat{r}\in\mathcal{R}_{\widehat{\tau}}} \inf_{r\in R_\tau} \sup_{\pi^\star\in\Pi^\star_{\mathcal{M}\cup r}} \max_a \left|Q_1^{\pi^\star}(s_1,a;\mathcal{M}\cup r) - Q_1^{\widehat{\pi}^\star}(s_1,a;\mathcal{M}\cup r)\right| = 0. \tag{35}$$

Combining Eq.(34) and Eq.(35), we conclude that $D^L(\tau, \widehat{\tau}) = 0$.

We then construct τ and $\widehat{\tau}$, respectively. We set $\mathcal{S} = \{1, 2, \ldots, S\}$, $\mathcal{A} = \{1, 2, \ldots, A\}$, and $H \geq 2$. Design the transitions $\mathbb{P}$ and $\widehat{\mathbb{P}}$ as follows:

$$\mathbb{P}_h(1|s,a) = 1, \qquad \widehat{\mathbb{P}}_h(2|s,a) = 1 \qquad \forall (h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}. \tag{36}$$

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P})$, $\widehat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, H, \widehat{\mathbb{P}})$. Define π^{E} and $(\bar{V}, \bar{A}) \in \overline{\mathcal{V}} \times \overline{\mathcal{A}}$ by

$$\begin{aligned}
&\pi^{\mathsf{E}}_h(1|s) = 1, \qquad \forall (h,s) \in [H]\times\mathcal{S}\times\mathcal{A} \\
&\bar{V}_h(s) = \mathbf{1}\{s = 1\}, \qquad \bar{A} \equiv \mathbf{0}, \qquad \forall h \in [H].
\end{aligned} \tag{37}$$

Set $\tau = (\mathcal{M}, \pi^{\mathsf{E}})$ and $\widehat{\tau} = (\widehat{\mathcal{M}}, \pi^{\mathsf{E}})$. By the result we proved at first, we have

$$D^L(\tau, \widehat{\tau}) = 0. \tag{38}$$

Let 1 be the initial state i.e., $\mathbb{P}(s_1 = 1) = 1$. By definition of D^π_Θ, we obtain that

$$D^{\mathsf{all}}_\Theta(\mathscr{R}^\tau, \mathscr{R}^{\widehat{\tau}}) \geq d^\pi\left(\mathscr{R}^\tau(\bar{V},\bar{A}), \mathscr{R}^{\widehat{\tau}}(\bar{V},\bar{A})\right) \geq \mathbb{E}_\pi\left[\left|V_2^\pi(s;\mathscr{R}^\tau(\bar{V},\bar{A})) - V_2^\pi\left(s;\mathscr{R}^{\widehat{\tau}}(\bar{V},\bar{A})\right)\right|\right] = |V_2(1) - V_2(2)| = 1, \tag{39}$$

where the send last equality is due to the construction of $\mathbb{P}$ and $\widehat{\mathbb{P}}$. □

Proof of Lemma D.3. Since $\mathcal{R}$ and $\mathcal{R}'$ are induced by $\mathscr{R}$ and $\mathscr{R}'$, then for any $r \in \mathcal{R}$ and $r' \in \mathcal{R}$, there exist V, $V' \in \overline{\mathcal{V}}$, A, $A \in \overline{\mathcal{A}}$ such that

$$r = \mathscr{R}(V, A), \qquad r' = \mathscr{R}'(V', A').$$

Then, we have

$$\begin{aligned}\sup_{r\in\mathcal{R}} \inf_{r'\in\mathcal{R}'} d(r, r') &= \sup_{V\in\overline{\mathcal{V}}, A\in\overline{\mathcal{A}}} \inf_{V'\in\overline{\mathcal{V}}, A'\in\overline{\mathcal{A}}} d(\mathscr{R}(V, A), \mathscr{R}'(V', A')) \\ &\leq \sup_{V\in\overline{\mathcal{V}}, A\in\overline{\mathcal{A}}} d(\mathscr{R}(V, A), \mathscr{R}'(V, A)) = D^{\mathsf{M}}(\mathscr{R}, \mathscr{R}').\end{aligned}$$

Similarly, we obtain

$$\sup_{r'\in\mathcal{R}'} \inf_{r\in\mathcal{R}} d(r, r') \leq D^{\mathsf{M}}(\mathscr{R}, \mathscr{R}').$$

Hence, we conclude that

$$D^{\mathsf{H}}(\mathcal{R}, \mathcal{R}') \leq D^{\mathsf{M}}(\mathscr{R}, \mathscr{R}').$$

□

Proof of Lemma D.4. Fix a $(\bar{s}, \bar{a}) \in \mathcal{S} \times \mathcal{A}$, we define metric d by

$$d(r, r') := |r_1(\bar{s}, \bar{a}) - r_1'(\bar{s}, \bar{a})|. \tag{40}$$

Give an IRL problem $(\mathcal{M}, \pi^{\mathsf{E}})$, let $\mathbb{P}$ be the transition dynamics of $(\mathcal{M}, \pi^{\mathsf{E}})$. Let $s^\star := \arg\min_{s\in\mathcal{S}} \mathbb{P}_1(s|\bar{s}, \bar{a})$. By the Pigeonhole Principle, we have $\mathbb{P}_1(s^\star|\bar{s}, \bar{a}) \leq 1/S \leq 1/2$. We construct transition $\mathbb{P}'$ by

$$\mathbb{P}_1'(s^\star|\bar{s}, \bar{a}) = 1. \tag{41}$$

Let $\widehat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}')$, $\widehat{\pi}^{\mathsf{E}} = \pi^{\mathsf{E}}$, and $\widehat{\mathscr{R}}$ be the reward mapping induced by $(\widehat{\mathcal{M}}, \widehat{\pi}^{\mathsf{E}})$. For any $(V, A) \in \bar{\mathcal{V}} \times \bar{\mathcal{A}}$, we define $(V', A') \in \bar{\mathcal{V}} \times \bar{\mathcal{A}}$ by

$$\begin{cases} V_2'(s^\star) = [\mathbb{P}_1 V_2](\bar{s}, \bar{a}) \\ V_h'(s) = V_h(s) & (h, s) \neq (2, s^\star), \end{cases} \quad \text{and } A' = A. \tag{42}$$

Then we have

$$\begin{aligned} d\Big(\mathscr{R}(V, A), \widehat{\mathscr{R}}(V', A')\Big) &= \Big|[\mathscr{R}(V, A)]_1(\bar{s}, \bar{a}) - \Big[\widehat{\mathscr{R}}(V, A)\Big]_1(\bar{s}, \bar{a})\Big| \qquad (43) \\ &= \Big| - A_1(\bar{s}, \bar{a}) \cdot \mathbf{1}\left\{\bar{a} \in \text{supp}\left(\pi_1^{\mathsf{E}}(\cdot|\bar{s})\right)\right\} + V_1(s) - [\mathbb{P}_1 V_2](\bar{s}, \bar{a}) \\ &\quad - \left\{-A_1'(\bar{s}, \bar{a}) \cdot \mathbf{1}\left\{\bar{a} \in \text{supp}\left(\widehat{\pi}_1^{\mathsf{E}}(\cdot|\bar{s})\right)\right\} + V_1'(s) - [\mathbb{P}_1' V_2'](\bar{s}, \bar{a})\right\}\Big| \\ &= |[\mathbb{P}_1' V_2'](\bar{s}, \bar{a}) - [\mathbb{P}_1 V_2](\bar{s}, \bar{a})| \\ &= |V_2'(s^\star) - [\mathbb{P}_1 V_2](\bar{s}, \bar{a})| = 0 \qquad (44) \end{aligned}$$

On the other hand, for any $(V', A') \in \bar{\mathcal{V}} \times \bar{\mathcal{A}}$, we set $(V, A) \in \bar{\mathcal{V}} \times \bar{\mathcal{A}}$ by

$$\begin{cases} V_2(s) = V_2'(s^\star), & s \in \mathcal{S}, \\ V_h(s) = V_h'(s) & h \neq 2, \end{cases} \tag{45}$$

which implies that

$$[\mathbb{P}_1 V_2](\bar{s}, \bar{a}) = V_2'(s^\star) = [\mathbb{P}_1' V_2'](\bar{s}, \bar{a}). \tag{46}$$

Hence, we have

$$\begin{aligned} d\Big(\mathscr{R}(V, A), \widehat{\mathscr{R}}(V', A')\Big) &= \Big|[\mathscr{R}(V, A)]_1(\bar{s}, \bar{a}) - \Big[\widehat{\mathscr{R}}(V, A)\Big]_1(\bar{s}, \bar{a})\Big| \qquad (47) \\ &= \Big| - A_1(\bar{s}, \bar{a}) \cdot \mathbf{1}\left\{\bar{a} \in \text{supp}\left(\pi_1^{\mathsf{E}}(\cdot|\bar{s})\right)\right\} + V_1(s) - [\mathbb{P}_1 V_2](\bar{s}, \bar{a}) \\ &\quad - \left\{-A_1'(\bar{s}, \bar{a}) \cdot \mathbf{1}\left\{\bar{a} \in \text{supp}\left(\widehat{\pi}_1^{\mathsf{E}}(\cdot|\bar{s})\right)\right\} + V_1'(s) - [\mathbb{P}_1' V_2'](\bar{s}, \bar{a})\right\}\Big| \\ &= |[\mathbb{P}_1' V_2'](\bar{s}, \bar{a}) - [\mathbb{P}_1 V_2](\bar{s}, \bar{a})| = 0 \qquad (48) \end{aligned}$$

Combining Eq.(43) and Eq.(47), we have $D^{\mathsf{H}}(\mathcal{R}, \widehat{\mathcal{R}}) = 0$.

Next, we lower bound $D^{\mathsf{M}}(\mathscr{R}, \widehat{\mathscr{R}})$. First, we define a parameter $(\widetilde{V}, \widetilde{A}) \in \bar{\mathcal{V}} \times \bar{\mathcal{A}}$ as follows:

$$\begin{cases} \widetilde{V}_2(s^\star) = H-1, & \\ \widetilde{V}_h(s) = 0, & (h,s) \neq (2, s^\star), \end{cases} \qquad \widetilde{A} \equiv \mathbf{0}. \tag{49}$$

Then we have

$$\begin{aligned} D^{\mathsf{M}}(\mathscr{R}, \widehat{\mathscr{R}}) \geq d\Big(\mathscr{R}(\widetilde{V}, \widetilde{A}), \widehat{\mathscr{R}}(\widetilde{V}, \widetilde{A})\Big) &= \Big|\Big[\mathbb{P}_1' \widetilde{V}_2\Big](\bar{s}, \bar{a}) - \Big[\mathbb{P}_1 \widetilde{V}_2\Big](\bar{s}, \bar{a})\Big| \\ &= |(H-1)(\mathbb{P}_1(s^\star|\bar{s}, \bar{a}) - 1)| \geq \frac{H-1}{2} \geq \frac{1}{2}, \end{aligned} \tag{50}$$

where the last line is due to $\mathbb{P}_1(s^\star|\bar{s}, \bar{a}) \leq 1/2$.

□

D.5. Proof of Proposition 6

Proof of Proposition 6. Since $\widehat{\pi}$ is an ϵ-optiaml policy in $\mathcal{M} \cup \widehat{r}$, we have

$$\epsilon' + V^{\widehat{\pi}}(s_1; \widehat{r}) \geq V^{\pi}(s_1; \widehat{r}). \tag{51}$$

In the same way, π is an $\bar{\epsilon}$-optiaml policy in $\mathcal{M} \cup r$, and therefore, we obtain that

$$\bar{\epsilon} + V^{\pi}(s_1; r) \geq V^{\widehat{\pi}}(s_1; r). \tag{52}$$

And by $\widehat{r}_h(s,a) \leq r_h(s,a)$ for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, we have

$$V^{\pi}(s_1; r) \geq V^{\pi}(s_1; \widehat{r}), \qquad V^{\widehat{\pi}}(s_1; r) \geq V^{\widehat{\pi}}(s_1; \widehat{r}). \tag{53}$$

Combining Eq.(51), Eq.(52) and Eq.(53), we conclude that

$$\epsilon' + \bar{\epsilon} + V^{\pi}(s_1; r) \geq \epsilon' + V^{\widehat{\pi}}(s_1; r) \geq \epsilon' + V^{\widehat{\pi}}(s_1; \widehat{r}) \geq V^{\pi}(s_1; \widehat{r}). \tag{54}$$

Hence, we have

$$\begin{aligned} |V^{\pi}(s_1; r) - V^{\widehat{\pi}}(s_1; r)| &\leq |\epsilon' + \bar{\epsilon} + V^{\pi}(s_1; r) - \Big(\epsilon' + V^{\widehat{\pi}}(s_1; r)\Big)| + \bar{\epsilon} \\ &\leq |\epsilon' + \bar{\epsilon} + V^{\pi}(s_1; r) - V^{\pi}(s_1; \widehat{r})| + \bar{\epsilon} \\ &\leq 2\bar{\epsilon} + \epsilon' + |V^{\pi}(s_1; r) - V^{\pi}(s_1; \widehat{r})| \leq \epsilon + \epsilon' + 2\bar{\epsilon}, \end{aligned} \tag{55}$$

where the first and last line is by triangle inequality and the second is by Eq.(54). □

E. Proofs for Section 4

E.1. Some lemmas

Lemma E.1 (Concentration event)**.** *Under the assumption of Theorem 8, there exists absolute constant C_1, C_2 such that the concentration event $\mathcal{E}$ holds with probability at least $1-\delta$, where*

$$\mathcal{E} := \Bigg\{ \textit{(i):} \ \Big|\Big[(\mathbb{P}_h - \widehat{\mathbb{P}}_h) V_{h+1}\Big](s,a)\Big| \leq b_h^{\theta}(s,a) \quad \forall\, \theta = (V, A) \in \Theta,\ (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}, \tag{56}$$

$$\textit{(ii):} \ \frac{1}{N_h(s,a) \vee 1} \leq \frac{C_1 \iota}{K d_h^{\pi^{\mathsf{b}}}(s,a)} \quad \forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}, \tag{57}$$

$$\textit{(iii):} \ N_h^e(s,a) \geq 1 \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \ \textit{s.t.}\ d_h^{\pi^{\mathsf{b}}}(s,a) \geq \frac{C_2 \eta \iota}{K}, a \in \operatorname{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right) \Bigg\}, \tag{58}$$

where $b_h(s,a)$ is defined in Eq.(8), $C^\star$ is specified in Definition B, and $N^e_h(s,a), \eta$ are given by

$$N^e_h(s,a) := \begin{cases} \sum_{(s_h,a_h,e_h)\in\mathcal{D}} \mathbf{1}\left\{(s_h,e_h)=(s,a)\right\} & \textit{in option } 1, \\ N^b_h(s,a) & \textit{in option } 2, \end{cases} \qquad \eta := \begin{cases} \frac{1}{\Delta} & \textit{in option } 1, \\ 1 & \textit{in option } 2. \end{cases}$$

Proof. When $N_h(s,a)=0$, then $\widehat{\mathbb{P}}_h(\cdot|s,a)=0$, as a result, claim (i) holds trivially. We then consider the case where $N_h(s,a)\geq 1$. For any $h\in[H]$, we define $\mathcal{N}_{\epsilon,h}$ as an ϵ/H-net with respect to $\|\cdot\|_\infty$ norm for $\overline{\mathcal{V}}^{\Theta}_h$. By definition of $\mathcal{N}(\Theta;\epsilon/H)$, we have

$$\log|\mathcal{N}_{\epsilon,h}| \leq \log \mathcal{N}(\Theta;\epsilon/H).$$

For fixed $\widetilde{V}_{h+1}\in\mathcal{N}_{\epsilon,h+1}$, $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$, by the empirical Bernstein inequality (Maurer & Pontil, 2009, Theorem 4), there exists some absolute constant $c>0$ such that

$$\begin{aligned} \left|\left[(\mathbb{P}_h-\widehat{\mathbb{P}}_h)\widetilde{V}_{h+1}\right](s,a)\right| &\leq \sqrt{\frac{c}{N^b_h(s,a)\vee 1}\left[\widehat{\mathbb{V}}_h\widetilde{V}_{h+1}\right](s,a)\log\frac{3HSA\cdot|\mathcal{N}_{\epsilon,h+1}|}{\delta}} \\ &\quad + \frac{cH}{N^b_h(s,a)\vee 1}\log\frac{3HSA\cdot|\mathcal{N}_{\epsilon,h+1}|}{\delta} \\ &\lesssim \sqrt{\frac{c\log\mathcal{N}(\Theta;\epsilon/H)\iota}{N^b_h(s,a)\vee 1}\left[\widehat{\mathbb{V}}_h\widetilde{V}_{h+1}\right](s,a)} + \frac{cH\log\mathcal{N}(\Theta;\epsilon/H)\iota}{N^b_h(s,a)\vee 1} \end{aligned}$$

with probability at least $1-\delta/(3HSA|\mathcal{N}_{\epsilon,h}|)$. Here $\lesssim$ hides absolute constants. Taking the union bound over all $\widetilde{V}_{h+1}\in\mathcal{N}_{\epsilon,h+1}$ and $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$, we know that with probability at least $1-\delta/3$,

$$\left|\left[(\mathbb{P}_h-\widehat{\mathbb{P}}_h)\widetilde{V}_{h+1}\right](s,a)\right| \lesssim \sqrt{\frac{c\log\mathcal{N}(\Theta;\epsilon/H)\iota}{N^b_h(s,a)\vee 1}\left[\widehat{\mathbb{V}}_h\widetilde{V}_{h+1}\right](s,a)} + \frac{cH\log\mathcal{N}(\Theta;\epsilon/H)\iota}{N^b_h(s,a)\vee 1}$$

holds simultaneously for all $\widetilde{V}\in\mathcal{N}_{\epsilon,h}$ and $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$.

For any $(V,A)\in\Theta$ and $h\in[H]$, there exists a $\widetilde{V}_h\in\overline{\mathcal{V}}^{\Theta}_h$ such that $\|V_h-\widetilde{V}_h\|_\infty\leq\epsilon/H$. Denote $(\widetilde{V}_1,\dots,\widetilde{V}_H)$ as $\widetilde{V}$. By applying the triangle inequality, we deduce that

$$\begin{aligned} &\left|\left[(\mathbb{P}_h-\widehat{\mathbb{P}}_h)V_{h+1}\right](s,a)\right| \leq \left|\left[\left(\widehat{\mathbb{P}}_h-\mathbb{P}_h\right)\widetilde{V}_{h+1}\right](s,a)\right| + 2\left\|\widetilde{V}-V\right\|_\infty \\ &\lesssim \sqrt{\frac{c\log\mathcal{N}(\Theta;\epsilon/H)\iota}{N^b_h(s,a)\vee 1}\left[\widehat{\mathbb{V}}_h\widetilde{V}_{h+1}\right](s,a)} + \frac{cH\log\mathcal{N}(\Theta;\epsilon/H)\iota}{N^b_h(s,a)\vee 1} + \frac{\epsilon}{H} \\ &\leq \sqrt{\frac{c\log\mathcal{N}(\Theta;\epsilon/H)\iota}{N^b_h(s,a)\vee 1}\left[\widehat{\mathbb{V}}_h V_{h+1}\right](s,a)} + \sqrt{\frac{c\log\mathcal{N}(\Theta;\epsilon/H)\iota}{N^b_h(s,a)\vee 1}\left[\widehat{\mathbb{V}}_h\left(\widetilde{V}_{h+1}-V_{h+1}\right)\right](s,a)} \\ &+ \frac{cH\log\mathcal{N}(\Theta;\epsilon/H)\iota}{N^b_h(s,a)\vee 1} + \frac{\epsilon}{H} \\ &\leq \sqrt{\frac{c\log\mathcal{N}(\Theta;\epsilon/H)\iota}{N^b_h(s,a)\vee 1}\left[\widehat{\mathbb{V}}_h V_{h+1}\right](s,a)} + \sqrt{\frac{c\log\mathcal{N}(\Theta;\epsilon/H)\iota\epsilon^2}{H^2\cdot N^b_h(s,a)\vee 1}} + \frac{cH\log\mathcal{N}(\Theta;\epsilon/H)\iota}{N^b_h(s,a)\vee 1} + \frac{\epsilon}{H} \\ &= \sqrt{\frac{c\log\mathcal{N}(\Theta;\epsilon/H)\iota}{N^b_h(s,a)\vee 1}\left[\widehat{\mathbb{V}}_h V_{h+1}\right](s,a)} + \frac{cH\log\mathcal{N}(\Theta;\epsilon/H)\iota}{N^b_h(s,a)\vee 1} + \frac{\epsilon}{H}\left(1+\sqrt{\frac{c\log\mathcal{N}(\Theta;\epsilon/H)\iota}{N^b_h(s,a)\vee 1}}\right) \end{aligned} \tag{59}$$

holds with probability at least $1-\delta/3$ for all $\theta=(V,A)\in\Theta$ and $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$. Here, the second inequality is by $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ and the last inequality in is by $\left[\widehat{\mathbb{V}}_h\left(\widetilde{V}_{h+1}-V_{h+1}\right)\right](s,a)\leq\frac{\epsilon^2}{H^2}$. On the other hand, by $|V_h|_\infty\leq H-h+1$, we obtain that

$$\left|\left[\left(\mathbb{P}_h-\widehat{\mathbb{P}}_h\right)V_{h+1}\right](s,a)\right| \leq 2(H-h+1)\leq 2H, \tag{60}$$

for all $V \in \Theta$ and $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$. Recall that, $b^{\theta}_h(s, a)$ is given by

$$\begin{aligned} b^{\theta}_h(s,a) = C \cdot \min \Bigg\{ & \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H)\iota}{N^b_h(s,a) \vee 1} \left[\widehat{\mathbb{V}}_h V_{h+1}\right](s,a)} + \frac{H \log \mathcal{N}(\Theta; \epsilon/H)\iota}{N^b_h(s,a) \vee 1} \\ & + \frac{\epsilon}{H}\left(1 + \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H)\iota}{N^b_h(s,a) \vee 1}}\right), H \Bigg\}, \end{aligned} \tag{61}$$

for some absolute constant C. Combining Eq.(59) and Eq.(60), it turns out that Claims (ii) holds.

For claim (ii), notice that $N_h(s, a) \sim \mathrm{Bin}(K, d^{\pi^{\mathsf{b}}}_h(s, a))$. Applying Lemma B.1 yields that

$$\frac{1}{N_h(s,a) \vee 1} \leq \frac{8}{K \cdot d^{\pi^{\mathsf{b}}}_h(s_h, a^E_h)} \cdot \log(\frac{3HSA}{\delta}) \leq \frac{C_1 \iota}{K d^{\pi^{\mathsf{b}}}_h(s,a)}$$

for some absolute constant C_1, with probability at least $1 - \delta/(3HSA)$. Taking the union bound yields claim (ii) over all (h, s, a) with probability at least $1 - \delta/3$.

For claim (iii), in option 2, for any $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$ such that $a \in \mathrm{supp}\left(\pi^{\mathsf{E}}_h(\cdot|s)\right)$ and $d^{\pi^{\mathsf{b}}}_h(s, a) \geq \frac{C_2 \eta \iota}{K}$, we have $N^e_h(s, a) \sim \mathrm{Bin}\Big(K, d^{\pi^{\mathsf{b}}}_h(s, a) \cdot \pi^{\mathsf{E}}_h(a|s)\Big)$. By direct computing, we obtain that

$$\begin{aligned} \mathbb{P}[N^e_h(s,a) = 0] &= (1 - d^{\pi^{\mathsf{b}}}_h(s,a) \cdot \pi^{\mathsf{E}}_h(a|s))^K \leq \Big(1 - \Delta \cdot d^{\pi^{\mathsf{b}}}_h(s,a)\Big)^K \\ &= \left[1 - \left(\frac{\delta}{3HSA}\right)^{1/K} + \left(\frac{\delta}{3HSA}\right)^{1/K} - \Delta \cdot d^{\pi^{\mathsf{b}}}_h(s,a)\right]^K \\ &\leq \left[\left(\frac{\delta}{3HSA}\right)^{1/K} + \underbrace{1 - \left(\frac{\delta}{3HSA}\right)^{1/K} - \Delta \cdot d^{\pi^{\mathsf{b}}}_h(s,a)}_{\leq 0}\right]^K \\ &\leq \left(\frac{\delta}{3HSA}\right)^{1/K \cdot K} = \frac{\delta}{3HSA}, \end{aligned}$$

where the second line follows from the well-posedness condition: $\pi^{\mathsf{E}}(a|s) \geq \Delta$ and the last inequality is valid since

$$1 - \left(\frac{\delta}{3HSA}\right)^{1/K} = 1 - \exp(-\frac{1}{K}\log\frac{\delta}{3HSA}) \leq -\frac{\widetilde{C}_2}{K}\log\frac{\delta}{3HSA} \leq \frac{C_2 \iota}{K} \leq \Delta \cdot d^{\pi^{\mathsf{b}}}_h(s,a),$$

where $\widetilde{C}_2$ and C_2 are absolute constants and the last inequality comes from $d^{\pi^{\mathsf{b}}}_h(s, a) \geq \frac{C_2 \eta \iota}{K} = \frac{C_2 \iota}{\Delta \cdot K}$. Hence, it holds that

$$N^e_h(s,a) \geq 1,$$

with probability at least $1 - \delta/(3HSA)$. Taking the union bound over all $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$ yields that

$$N^e_h(s,a) \geq 1$$

holds with probability at least $1 - \delta/3$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ s.t. $d^{\pi^{\mathsf{b}}}_h(s, a) \geq \frac{C_2 \eta \iota}{K}$, $a \in \mathrm{supp}\left(\pi^{\mathsf{E}}_h(\cdot|s)\right)$, which implies that claim (iii) holds.

In option 1, notice that $\mathbb{P}[N^e_h(s, a) = 0] = \Big(1 - d^{\pi^{\mathsf{b}}(s,a)}_h(s, a)\Big)^K$, with a similar argument, we can prove the claim (iii) in option 1.

Further, we can conclude that the concentration event $\mathcal{E}$ holds with probability at least $1 - \delta$. □

Lemma E.2 (Performance decomposition for RLE)**.** *For any $\theta = (V, A) \in \Theta$, let $r^{\theta} = \mathscr{R}^{\star}(V, A)$ and $\widehat{r}^{\theta} = \widehat{\mathscr{R}}(V, A)$, where $\mathscr{R}^{\star}$ is the ground truth reward mapping and $\widehat{\mathscr{R}}$ is the estimated reward mapping outputted by* RLP. *On the event defined in Lemma E.1, for any $\theta \in \Theta$ and any $h \in [H]$, we have*

$$d^{\pi^{\mathsf{eval}}}\left(r^{\theta}, \widehat{r}^{\theta}\right) \lesssim \frac{C^{\star} H^2 S \eta \iota}{K} + \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) b_h^{\theta}(s,a),$$

where $C^{\star}$ is defined in Assumption B and η is specified in Lemma E.1

Proof. Fix a tuple $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$. When $a \notin \text{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)$, by definition of $N_h^e(s,a)$, we have $N_h^e(s,a) = 0$ By construction of $\widehat{\pi}_h^{\mathsf{E}}(a|s)$ in Algorithm 1, we deduce that $\widehat{\pi}_h^{\mathsf{E}}(a|s) = 0$, and therefore

$$\left|\mathbf{1}\left\{a \notin \text{supp}\, \widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right\} - \mathbf{1}\left\{a \notin \text{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\}\right| = |0 - 0| = 0.$$

When $a \in \text{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)$ and $d_h^{\pi^{\mathsf{b}}}(s,a) < \frac{C_2\eta\iota}{K}$, then

$$\left|\mathbf{1}\left\{a \notin \text{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\} - \mathbf{1}\left\{a \notin \text{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\}\right| \le 2.$$

If $a \in \text{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)$ and $d_h^{\pi^{\mathsf{b}}}(s,a) \ge \frac{C_2\eta\iota}{K}$, then by concentration event $\mathcal{E}$ (iii), $N_h^e(s,a) \ge 1$ which implies that $\widehat{\pi}_h^{\mathsf{E}}(a|s) > 0$. Hence, we obtain that

$$\left|\mathbf{1}\left\{a \notin \text{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\} - \mathbf{1}\left\{a \notin \text{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\}\right| = |1 - 1| = 0.$$

Thus we can conclude that

$$\left|\mathbf{1}\left\{a \notin \text{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\} - \mathbf{1}\left\{a \notin \text{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\}\right| \le 2 \cdot \mathbf{1}\left\{d_h^{\pi^{\mathsf{b}}}(s,a) < \frac{C_2\eta\iota}{K}, a \in \text{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\}. \tag{62}$$

We then bound the $\left|\left[r_h^{\theta} - \widehat{r}_h^{\theta}\right](s,a)\right|$ for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$:

$$\begin{aligned}
&\left|\left[r_h^{\theta} - \widehat{r}_h^{\theta}\right](s,a)\right| = \Big| - A_h(s,a) \cdot \mathbf{1}\left\{a \notin \text{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\} + V_h(s) - [\mathbb{P}_h V_{h+1}](s,a) \\
&+ A_h(s,a) \cdot \mathbf{1}\left\{a \notin \text{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\} - V_h(s) + \left[\widehat{\mathbb{P}}_h V_{h+1}\right](s,a) + b_h^{\theta}(s,a)\Big| \\
&\le A_h(s,a) \cdot \left|\mathbf{1}\left\{a \notin \text{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\} - \mathbf{1}\left\{a \notin \text{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\}\right| + \left|\left[(\mathbb{P}_h - \widehat{\mathbb{P}}_h)V_{h+1}\right](s,a)\right| + b_h^{\theta}(s,a) \\
&\le 2H \cdot \mathbf{1}\left\{d_h^{\pi^{\mathsf{b}}}(s,a) < \frac{C_2\eta\iota}{K}, a \in \text{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\} + \left|\left[(\mathbb{P}_h - \widehat{\mathbb{P}}_h)V_{h+1}\right](s,a)\right| + b_h^{\theta}(s,a) \\
&\le 2H \cdot \mathbf{1}\left\{d_h^{\pi^{\mathsf{b}}}(s,a) < \frac{C_2\eta\iota}{K}, a \in \text{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\} + 2b_h^{\theta}(s,a),
\end{aligned} \tag{63}$$

where the second line follows from the triangle inequality, the third line comes from Eq.(62), the second last line follows from $\|A_h\|_\infty \le H$, the last line comes from the concentration event $\mathcal{E}$ (i). Finally, we give the bound of $\mathbb{E}_{\pi^{\mathsf{eval}}}|V_h^{\pi^{\mathsf{eval}}}(s; r^{\theta}) -$

$V_h^{\pi^{\mathsf{eval}}}(s;\widehat{r}^\theta)|$. By definition of the V function, we have

$$
\begin{aligned}
&\mathbb{E}_{\pi^{\mathsf{eval}}}\left|V_h^{\pi^{\mathsf{eval}}}(s;r^\theta)-V_h^{\pi^{\mathsf{eval}}}(s;\widehat{r}^\theta)\right| = \sum_{s\in\mathcal{S}} d_h^{\pi^{\mathsf{eval}}}(s)\cdot\left|V_h^{\pi^{\mathsf{eval}}}(s;r^\theta)-V_h^{\pi^{\mathsf{eval}}}(s;\widehat{r}^\theta)\right|\\
&= \sum_{s'\in\mathcal{S}} d_h^{\pi^{\mathsf{eval}}}(s')\cdot\left|\sum_{h'\geq h}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{h'}^{\pi^{\mathsf{eval}}}(s_{h'}=s,a_{h'}=a|s_h=s')\cdot\big[r_{h'}^\theta-\widehat{r}_{h'}^\theta\big](s,a)\right|\\
&\leq \sum_{h'\geq h}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left\{\sum_{s\in\mathcal{S}} d_h^{\pi^{\mathsf{eval}}}(s)\cdot d_{h'}^{\pi^{\mathsf{eval}}}(s,a|s_h=s)\right\}\cdot\big|\big[r_{h'}^\theta-\widehat{r}_{h'}^\theta\big](s,a)\big|\\
&\overset{(i)}{\leq} \sum_{h'\geq h}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{h'}^{\pi^{\mathsf{eval}}}(s,a)\cdot\big|\big[r_{h'}^\theta-\widehat{r}_{h'}^\theta\big](s,a)\big|\\
&\overset{(ii)}{\leq} \sum_{h'\geq h}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{h'}^{\pi^{\mathsf{eval}}}(s,a)\cdot\left[2H\cdot\mathbf{1}\left\{d_h^{\pi^{\mathsf{b}}}(s,a)<\frac{C_2\eta\iota}{K}, a\in\mathrm{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\}+2b_h^\theta(s,a)\right]\\
&\leq \sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{2Hd_h^{\pi^{\mathsf{eval}}}(s,a)}{d_h^{\pi^{\mathsf{b}}}(s,a)}d_h^{\pi^{\mathsf{b}}}(s,a)\cdot\mathbf{1}\left\{d_h^{\pi^{\mathsf{b}}}(s,a)<\frac{C_2\eta\iota}{K}, a\in\mathrm{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\}\\
&+\sum_{h\geq 1}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} 2d_h^{\pi^{\mathsf{eval}}}(s,a)b_h^\theta(s,a)\\
&\leq 2H\cdot\frac{C_2\eta\iota}{K}\cdot\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{d_h^{\pi^{\mathsf{eval}}}(s,a)}{d_h^{\pi^{\mathsf{b}}}(s,a)}+\sum_{h\geq 1}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} 2d_h^{\pi^{\mathsf{eval}}}(s,a)b_h^\theta(s,a)\\
&\overset{(iii)}{\lesssim} \frac{C^\star H^2 S\eta\iota}{K}+\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a)b_h^\theta(s,a),
\end{aligned}
$$

where $d_{h'}^{\pi^{\mathsf{eval}}}(s_{h'}=s,a_{h'}=a|s_h=s)=\mathbb{P}_h(s_{h'}=s,a_{h'}=a|s_h=s)$, (i) is due to $d_{h'}^{\pi^{\mathsf{eval}}}(s,a)=\sum_{s\in\mathcal{S}} d_h^{\pi^{\mathsf{eval}}}(s)\cdot d_{h'}^{\pi^{\mathsf{eval}}}(s_{h'}=s,a_{h'}=a|s_h=s)$, (ii) follows from Eq.(63) and (iii) comes from definition of $C^\star$-concentrability. This completes the proof. □

E.2. Proof of Theorem 8

Proof. By Lemma E.2, we have

$$
D_\Theta^{\pi^{\mathsf{eval}}}(\mathscr{R}^\star,\widehat{\mathscr{R}})\lesssim\frac{C^\star H^2 S\eta\iota}{K}+\underbrace{\sup_{\theta\in\Theta}\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a)b_h^\theta(s,a)}_{\text{(I)}}\,. \tag{64}
$$

Plugging Eq.(61) into Eq.(64), we obtain that

$$
\begin{aligned}
\text{(I)} &= \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) b_h(s,a) \\
&\lesssim \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \Bigg\{ \sqrt{\frac{\log \mathcal{N}(\Theta;\epsilon/H)\iota}{N_h^b(s,a)\vee 1}\Big[\widehat{\mathbb{V}}_h V_{h+1}\Big](s,a)} \\
&\quad + \frac{H\log \mathcal{N}(\Theta;\epsilon/H)\iota}{N_h^b(s,a)\vee 1} + \frac{\epsilon}{H}\left(1+\sqrt{\frac{\log \mathcal{N}(\Theta;\epsilon/H)\iota}{N_h^b(s,a)\vee 1}}\right)\Bigg\}
\end{aligned}
\tag{65}
$$

$$
\begin{aligned}
&\le \underbrace{\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \sqrt{\frac{\log \mathcal{N}(\Theta;\epsilon/H)\iota}{N_h^b(s,a)\vee 1}[\mathbb{V}_h V_{h+1}](s,a)}}_{\text{(I.a)}} \\
&\quad + \underbrace{\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \sqrt{\frac{\log \mathcal{N}(\Theta;\epsilon/H)\iota}{N_h^b(s,a)\vee 1}\Big[\Big(\widehat{\mathbb{V}}_h - \mathbb{V}_h\Big)V_{h+1}\Big](s,a)}}_{\text{(I.b)}} \\
&\quad + \underbrace{\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \frac{H\log \mathcal{N}(\Theta;\epsilon/H)\iota}{N_h^b(s,a)\vee 1}}_{\text{(I.c)}} \\
&\quad + \underbrace{\epsilon \cdot \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \left(\frac{1}{H} + \sqrt{\frac{\log \mathcal{N}(\Theta;\epsilon/H)\iota}{H^2\cdot N_h^b(s,a)\vee 1}}\right)}_{\text{(I.d)}}
\end{aligned}
\tag{66}
$$

where the last inequality comes from the triangle inequality. We study the four terms separately.
For the term (I.a), on the concentration event $\mathcal{E}$, we have

$$
\begin{aligned}
\text{(I.a)} &= \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \sqrt{\frac{\log \mathcal{N}(\Theta;\epsilon/H)\iota}{N_h^b(s,a)\vee 1}[\mathbb{V}_h V_{h+1}](s,a)} \\
&\lesssim \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \sqrt{\frac{\log \mathcal{N}(\Theta;\epsilon/H)\iota}{K d^{\pi^{\mathsf{b}}}(s,a)}[\mathbb{V}_h V_{h+1}](s,a)} \\
&\le \sqrt{\frac{H^2\log \mathcal{N}(\Theta;\epsilon/H)\iota}{K}} \cdot \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sqrt{d_h^{\pi^{\mathsf{eval}}}(s,a)} \cdot \sqrt{\frac{d_h^{\pi^{\mathsf{eval}}}(s,a)}{d_h^{\pi^{\mathsf{b}}}(s,a)}} \\
&\le \sqrt{\frac{H^2\log \mathcal{N}(\Theta;\epsilon/H)\iota}{K}} \cdot \sqrt{\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{d_h^{\pi^{\mathsf{eval}}}(s,a)}{d_h^{\pi^{\mathsf{b}}}(s,a)}} \cdot \sqrt{\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a)} \\
&\hspace{20em}\text{(by Cauchy-Schwarz inequality)} \\
&\le \sqrt{\frac{C^\star H^4 S\log \mathcal{N}(\Theta;\epsilon/H)\iota}{K}},
\end{aligned}
\tag{67}
$$

where the second line comes from concentration event $\mathcal{E}$(ii), the third line is valid since $\|V_{h+1}\|_\infty \le H$ and the last is by thw definition of $C^\star$-concentrability.

Next, we study the term (I.b). For any (h, s, a), we have

$$\begin{aligned}
&\Big|\Big[\Big(\widehat{\mathbb{V}}_h - \mathbb{V}_h\Big)V_{h+1}\Big](s,a)\Big| \\
&= \Big[(\widehat{\mathbb{P}}_h V_{h+1})^2 - (\widehat{\mathbb{P}}_h V_{h+1})^2 - \big(\mathbb{P}_h (V_{h+1})^2 - (\mathbb{P}_h V_{h+1})^2\big)\Big](s,a) \\
&\le \Big|\Big[(\widehat{\mathbb{P}}_h - \mathbb{P}_h)(V_{h+1})^2\Big](s,a)\Big| + \Big|\Big[(\widehat{\mathbb{P}}_h + \mathbb{P}_h)V_{h+1} \cdot (\widehat{\mathbb{P}}_h - \mathbb{P}_h)V_{h+1}\Big](s,a)\Big| \\
&\le 2H\Big|\Big[(\widehat{\mathbb{P}}_h - \mathbb{P}_h)(V_{h+1})\Big](s,a)\Big| + 2H\Big|\Big[(\widehat{\mathbb{P}}_h - \mathbb{P}_h)V_{h+1}\Big](s,a)\Big| \\
&\lesssim c\sqrt{\frac{H^4\iota}{N_h^b(s,a) \vee 1}},
\end{aligned} \tag{68}$$

where the second last inequality is by $\|V_{h+1}\|_\infty \le H$ and the last inequality follows from the Azuma-Hoeffding inequality. By applying Eq.(68), we can obtain the bound for the term (I.b):

$$\begin{aligned}
\text{(I.b)} &= \sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota}{N_h^b(s,a)\vee 1}\Big[\Big(\widehat{\mathbb{V}}_h - \mathbb{V}_h\Big)V_{h+1}\Big](s,a)} \\
&\le \sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota}{N_h^b(s,a)\vee 1} \cdot \sqrt{\frac{H^4\iota}{\widehat{N}_h^b(s,a)\vee 1}}} \\
&= (\log\mathcal{N}(\Theta;\epsilon/H))^{1/2} \cdot \sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \frac{H\iota^{3/4}}{\big\{N_h^b(s,a)\vee 1\big\}^{3/4}} \\
&\le \underbrace{(\log\mathcal{N}(\Theta;\epsilon/H))^{1/2} \cdot \sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \sqrt{\frac{1}{N_h^b(s,a)\vee 1}}}_{\text{(I.b.1)}} \\
&\quad + \underbrace{(\log\mathcal{N}(\Theta;\epsilon/H))^{1/2} \cdot \sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \frac{H^2\iota^{3/2}}{N_h^b(s,a)\vee 1}}_{\text{(I.b.2)}},
\end{aligned} \tag{69}$$

where the last line is from AM-GM inequality. For the term (I.b.1), on the concentration event $\mathcal{E}$, we have

$$\begin{aligned}
\text{(I.b.1)} &= (\log\mathcal{N}(\Theta;\epsilon/H))^{1/2} \cdot \sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \sqrt{\frac{1}{N_h^b(s,a)\vee 1}} \\
&\le (\log\mathcal{N}(\Theta;\epsilon/H))^{1/2} \cdot \sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sqrt{d_h^{\pi^{\mathsf{eval}}}(s,a)} \cdot \sqrt{\frac{d_h^{\pi^{\mathsf{eval}}}(s,a)}{K d_h^{\pi^{\mathsf{b}}}(s,a)}} \\
&\le (\log\mathcal{N}(\Theta;\epsilon/H))^{1/2} \cdot \sqrt{\frac{1}{K}} \cdot \sqrt{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{d_h^{\pi^{\mathsf{eval}}}(s,a)}{d_h^{\pi^{\mathsf{b}}}(s,a)}} \cdot \sqrt{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a)} \\
&\le \sqrt{\frac{C^\star HS\log\mathcal{N}(\Theta;\epsilon/H)}{K}} \cdot \sqrt{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a)} \\
&= \sqrt{\frac{C^\star H^2 S\log\mathcal{N}(\Theta;\epsilon/H)}{K}},
\end{aligned} \tag{70}$$

where the second line is directly from concentration event $\mathcal{E}(ii)$, the third line follows from Cauchy-Schwarz inequality and the second last line comes from the definition of $C^\star$-concentrability. For the term (I.b.2), on the concentration event $\mathcal{E}$, we obtain

$$\begin{aligned}
\text{(I.b.2)} &= (\log \mathcal{N}(\Theta; \epsilon/H))^{1/2} \cdot \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \frac{H^2\iota^{3/2}}{N_h^b(s,a)\vee 1} \\
&\le (\log \mathcal{N}(\Theta; \epsilon/H))^{1/2} \cdot \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \frac{H^2\iota^{5/2}}{K d_h^{\pi^{\mathsf{b}}}(s,a)} \\
&\le (\log \mathcal{N}(\Theta; \epsilon/H))^{1/2} \cdot \frac{H^2\iota^{5/2}}{K} \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{d_h^{\pi^{\mathsf{eval}}}(s,a)}{d_h^{\pi^{\mathsf{b}}}(s,a)} \\
&= \frac{C^\star H^3 S \log \mathcal{N}(\Theta; \epsilon/H)\iota^{5/2}}{K},
\end{aligned} \tag{71}$$

where the second line comes from concentration event $\mathcal{E}$(ii), the third line follows from the definition of $C^\star$-concentrability. Combining Eq.(70) and Eq.(71), the term (I.b) can be bounded as follows:

$$\text{(I.b)} \lesssim \sqrt{\frac{C^\star H^2 S \log \mathcal{N}(\Theta; \epsilon/H)}{K}} + \frac{C^\star H^3 S \log \mathcal{N}(\Theta; \epsilon/H)\iota^{5/2}}{K}. \tag{72}$$

For the term (I.c), observe that

$$\text{(I.c)} = \text{(I.b.2)}/(H\iota^{3/2}).$$

Hence, by Eq.(71), we deduce that

$$\text{(I.c)} \le \frac{C^\star H^2 S \log \mathcal{N}(\Theta; \epsilon/H)\iota}{K} \tag{73}$$

For the term (I.d),

$$\begin{aligned}
\text{(I.d)} &= \epsilon \cdot \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \left(\frac{1}{H} + \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H)\iota}{H^2 \cdot N_h^b(s,a)\vee 1}} \right) \\
&= \epsilon + \epsilon \cdot \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H)\iota}{H^2 \cdot N_h^b(s,a)\vee 1}} \\
&= \epsilon + \epsilon \cdot \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) \cdot \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H)\iota}{H^2 K d_h^{\pi^{\mathsf{b}}}(s,a)}} \\
&= \epsilon + \epsilon \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H)\iota}{H^2 K}} \cdot \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sqrt{d_h^{\pi^{\mathsf{eval}}}(s,a)} \sqrt{\frac{d_h^{\pi^{\mathsf{eval}}}(s,a)}{d_h^{\pi^{\mathsf{b}}}(s,a)}} \\
&\le \epsilon + \epsilon \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H)\iota}{H^2 K}} \sqrt{\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a)} \cdot \sqrt{\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{d_h^{\pi^{\mathsf{eval}}}(s,a)}{d_h^{\pi^{\mathsf{b}}}(s,a)}} \\
&\le \epsilon \cdot (1 + \sqrt{\frac{C^\star S \log \mathcal{N}(\Theta; \epsilon/H)\iota}{K}}),
\end{aligned} \tag{74}$$

where the second last line is by Cauchy-Schwarz inequality and the last line is by definition of $C^\star$-concentrablity. Combining Eq.(67), Eq.(72), Eq.(73) and Eq.(74), we deduce that

$$\begin{aligned}
\text{(I)} &\lesssim \text{(I.a)} + \text{(I.b)} + \text{(I.c)} + \text{(I.d)} \\
&\lesssim \sqrt{\frac{C^\star H^4 S \log \mathcal{N}(\Theta; \epsilon/H)\iota}{K}} + \sqrt{\frac{C^\star H^2 S \log \mathcal{N}(\Theta; \epsilon/H)}{K}} + \frac{C^\star H^3 S \log \mathcal{N}(\Theta; \epsilon/H)\iota^{5/2}}{K} \\
&+ \frac{C^\star H^2 S \log \mathcal{N}(\Theta; \epsilon/H)\iota}{K} + \epsilon \cdot (1 + \sqrt{\frac{C^\star S \log \mathcal{N}(\Theta; \epsilon/H)\iota}{K}}) \\
&\lesssim \sqrt{\frac{C^\star H^4 S \log \mathcal{N}(\Theta; \epsilon/H)\iota}{K}} + \frac{C^\star H^3 S \log \mathcal{N}(\Theta; \epsilon/H)\iota^{5/2}}{K} + \epsilon.
\end{aligned}$$

Finally, plugging into Eq.(64), the final bound is given by

$$D_{\Theta}^{\pi^{\mathsf{eval}}}(\mathscr{R}^\star,\widehat{\mathscr{R}}) \lesssim \frac{C^\star H^2 S\eta\iota}{K} + \sqrt{\frac{C^\star H^4 S\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K}} + \frac{C^\star H^3 S\log\mathcal{N}(\Theta;\epsilon/H)\iota^{5/2}}{K} + \epsilon$$

The right-hand-side is upper bounded by 2ϵ as long as

$$K \geq \widetilde{\mathcal{O}}\bigg(\frac{C^\star H^4 S\log\mathcal{N}(\Theta;\epsilon/H)}{\epsilon^2} + \frac{C^\star H^2 S\eta}{\epsilon}\bigg).$$

Here $poly\log(H,S,A,1/\delta)$ are omitted.

□

E.3. Proof of Theorem 9

In this section, we will consider the case that $\pi^{\mathsf{eval}}=\pi^{\mathsf{E}}$. We first introduce the following concentration event which is slightly different from the concentration event defined in Lemma E.1.

Lemma E.3 (Concentration event). *Under the setting of Theorem 8, there exists an absolute constant $C_1,\ C_2$ such that the concentration event $\mathcal{E}$ holds with probability at least $1-\delta$, where*

$$\mathcal{E} := \bigg\{\textit{(i):}\ \Big|\Big[(\mathbb{P}_h-\widehat{\mathbb{P}}_h)V_{h+1}\Big](s,a)\Big| \leq b_h^{\theta}(s,a)\quad \forall\,\theta=(V,A)\in\Theta,\ (h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}, \tag{75}$$

$$\textit{(ii):}\ \frac{1}{N_h(s,a)\vee 1} \leq \frac{C_1\iota}{Kd_h^{\pi^{\mathsf{b}}}(s,a)},\quad \forall (h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}, \tag{76}$$

$$\textit{(iii):}\ N_h^e(s,a)\geq 1\quad \forall (h,s,a)\in\mathcal{S}\times\mathcal{A}\ \textit{s.t.}\ \bar{d}_h(s,a)\geq \frac{C_2\iota}{K}, a\in \text{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\bigg\}, \tag{77}$$

where $b_h^\theta(s,a)$ is defined in Eq.(8), $C^\star$ is specified in Definition B, and $N_h^e(s)$ is given by

$$N_h^e(s,a) := \begin{cases}\sum_{(s_h,a_h,e_h)\in\mathcal{D}}\mathbf{1}\left\{(s_h,e_h)=(s,a)\right\} & \textit{in option } 1,\\ N_h^b(s,a) & \textit{in option } 2,\end{cases}$$

$$\bar{d}_h(s,a) := \begin{cases} d_h^{\pi^{\mathsf{b}}}(s)\cdot\pi_h^{\mathsf{E}}(a|s) & \textit{in option } 1,\\ d_h^{\pi^{\mathsf{b}}}(s,a) & \textit{in option } 2.\end{cases} \tag{78}$$

Proof. Repeating the arguments in the proof of Lemma E.1, we can prove that claim (i), (ii) holds with probability at least $1-\frac{2\delta}{3}$.
For claim (iii), for any $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$ such that $a\in\text{supp}\left(\pi_h^{\mathsf{E}}\cdot|s\right)$ and $\bar{d}_h(s,a)\geq\frac{C_2\iota}{K}$, $N_h^e(s,a)\sim\text{Bin}\big(K,\bar{d}_h(s,a)\big)$.
By direct computing, we obtain that

$$\begin{aligned}
\mathbb{P}[N_h^e(s,a)=0] = \big(1-\bar{d}_h(s,a)\big)^K &= \left[1-\left(\frac{\delta}{3HSA}\right)^{1/K}+\left(\frac{\delta}{3HSA}\right)^{1/K}-\bar{d}_h(s,a)\right]^K\\
&\leq \left[\left(\frac{\delta}{3HSA}\right)^{1/K} + \underbrace{1-\left(\frac{\delta}{3HSA}\right)^{1/K}-\bar{d}_h(s,a)}_{\leq 0}\right]^K\\
&\leq \left(\frac{\delta}{3HSA}\right)^{1/K\cdot K} = \frac{\delta}{3HSA},
\end{aligned}$$

where the last inequality is valid since

$$1-\left(\frac{\delta}{3HSA}\right)^{1/K}=1-\exp(-\frac{1}{K}\log\frac{\delta}{3HSA})\leq-\frac{\widetilde{C}_2}{K}\log\frac{\delta}{3HSA}\leq\frac{C_2\iota}{K}\leq\bar{d}_h(s,a),$$

where $\widetilde{C}_2$ and C_2 are absolute constants. Hence, it holds that

$$N_h^e(s,a)\geq 1,$$

with probability at least $1-\delta/(3HSA)$. Taking the union bound over all $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$ yields that

$$N_h^e(s,a)\geq 1$$

holds with probability at least $1-\delta/3$ for all $(h,s,a)\in\mathcal{S}\times\mathcal{A}$ s.t. $\bar{d}_h(s,a)\geq\frac{C_2\iota}{K}$, $a\in\operatorname{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)$, which implies that claim (iii) holds. Further, we conclude that the concentration event $\mathcal{E}$ holds with probability at least $1-\delta$. □

Proof of Corollary 9. Recall that $r^\theta=\mathscr{R}^\star(V,A)$ and $\widehat{r}^\theta=\widehat{\mathscr{R}}(V,A)$ for any $\theta=(V,A)\in\Theta$. When $\pi^{\mathsf{eval}}=\pi^{\mathsf{E}}$, repeating the arguments in Lemma E.2, we have following decomposition:

$$\begin{aligned}
d^{\pi^{\mathsf{eval}}}\left(r^\theta,\widehat{r}^\theta\right)&\leq 2H\cdot\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\pi^{\mathsf{E}}}(s,a)\cdot\mathbf{1}\left\{\bar{d}_h(s,a)<\frac{C_2\iota}{K},a\in\operatorname{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\}\\
&\quad+\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}2d_h^{\pi^{\mathsf{eval}}}(s,a)b_h^\theta(s,a)\\
&\leq 2H\cdot\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{d_h^{\pi^{\mathsf{E}}}(s,a)}{\bar{d}_h(s,a)}\cdot\bar{d}_h(s,a)\cdot\mathbf{1}\left\{\bar{d}_h(s,a)<\frac{C_2\iota}{K},a\in\operatorname{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\}\\
&\quad+\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}2d_h^{\pi^{\mathsf{eval}}}(s,a)b_h^\theta(s,a)\\
&\lesssim\frac{H\iota}{K}\cdot\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{d_h^{\pi^{\mathsf{E}}}(s,a)}{\bar{d}_h(s,a)}+\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\pi^{\mathsf{eval}}}(s,a)b_h^\theta(s,a)\\
&\leq\frac{C^\star H^2SA\iota}{K}+\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\pi^{\mathsf{eval}}}(s,a)b_h^\theta(s,a).
\end{aligned}$$

where the second last line is valid since

$$\begin{aligned}
\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{d_h^{\pi^{\mathsf{E}}}(s,a)}{\bar{d}_h(s,a)}&=\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{d_h^{\pi^{\mathsf{E}}}(s)\cdot\pi_h^{\mathsf{E}}(a|s)}{d_h^{\pi^{\mathsf{b}}}(s)\cdot\pi_h^{\mathsf{E}}(a|s)}\\
&=A\cdot\sum_{h\in[H]}\sum_{s\in\mathcal{S}}\frac{d_h^{\pi^{\mathsf{E}}}(s)}{d_h^{\pi^{\mathsf{b}}}(s)}\\
&=A\cdot\sum_{h\in[H]}\sum_{s\in\mathcal{S}}\frac{\sum_{a\in\mathcal{A}}d_h^{\pi^{\mathsf{E}}}(s,a)}{\sum_{a\in\mathcal{A}}d_h^{\pi^{\mathsf{b}}}(s,a)}\\
&\leq A\cdot\sum_{h\in[H]}\sum_{s\in\mathcal{S}}\max_{a\in\mathcal{A}}\frac{d_h^{\pi^{\mathsf{E}}}(s,a)}{d_h^{\pi^{\mathsf{b}}}(s,a)}\\
&\leq A\cdot\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{d_h^{\pi^{\mathsf{E}}}(s,a)}{d_h^{\pi^{\mathsf{b}}}(s,a)}\\
&\leq C^\star HSA.
\end{aligned}$$

Similar as Eq.(64), we can decompose $D_{\Theta}^{\pi^{\mathsf{eval}}}(\mathscr{R}^\star, \widehat{\mathscr{R}})$ as follows:

$$D_{\Theta}^{\pi^{\mathsf{E}}}(\mathscr{R}^\star, \widehat{\mathscr{R}}) \lesssim \frac{C^\star H^2 S \eta \iota}{K} + \underbrace{\sup_{\theta\in\Theta} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s,a) b_h^{\theta}(s,a)}_{\text{(I)}} \tag{79}$$

We can decompose terms (I) into four terms (I.a), (I.b), (I.c), and (I.d) as in Eq.(66). Since we don't use claim (iii) in the proof of bounding (I.b), (I.c), and (I.d), Eq.(72), Eq.(73) and Eq.(74) still holds on the concentration event $\mathcal{E}$ defined in Lemma E.3. In the following, we will prove an improved bound of the term (I.a):

$$\begin{aligned}
\text{(I.a)} &= \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{E}}}(s,a) \cdot \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H)\iota}{N_h^b(s,a) \vee 1} [\mathbb{V}_h V_{h+1}](s,a)} \\
&\leq \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{E}}}(s,a) \cdot \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H)\iota}{K d^{\pi^{\mathsf{b}}}(s,a)} [\mathbb{V}_h V_{h+1}](s,a)} \\
&= \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H)\iota}{K}} \cdot \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sqrt{d_h^{\pi^{\mathsf{E}}}(s,a) \cdot [\mathbb{V}_h V_{h+1}](s,a)} \cdot \sqrt{\frac{d_h^{\pi^{\mathsf{E}}}(s,a)}{d_h^{\pi^{\mathsf{b}}}(s,a)}} \\
&\leq \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H)\iota}{K}} \cdot \sqrt{\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{E}}}(s,a) \cdot [\mathbb{V}_h V_{h+1}](s,a)} \cdot \sqrt{\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{d_h^{\pi^{\mathsf{E}}}(s,a)}{d_h^{\pi^{\mathsf{b}}}(s,a)}} \\
&\sqrt{\frac{C^\star H S \log \mathcal{N}(\Theta; \epsilon/H)\iota}{K}} \cdot \sqrt{\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{E}}}(s,a) \cdot [\mathbb{V}_h V_{h+1}](s,a)}
\end{aligned} \tag{80}$$

We then give a sharp bound of $\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{E}}}(s,a)\cdot\left[\mathbb{V}_h V_{h+1}\right](s,a)$.

$$
\begin{aligned}
&\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathsf{E}}}(s,a)\cdot\left[\mathbb{V}_h V_{h+1}\right](s,a)\\
&=\sum_{h=1}^{H}\mathbb{E}_{\pi^{\mathsf{E}}}\left[\mathrm{Var}_{\pi^{\mathsf{E}}}\left[V_{h+1}(s_{h+1})|s_h,a_h\right]\right]\\
&\overset{\text{(i)}}{=}\sum_{h=1}^{H}\mathbb{E}_{\pi^{\mathsf{E}}}\left[\mathbb{E}\left[\left(V_{h+1}(s_{h+1})+A_h(s_h,a_h)\cdot\mathbf{1}\left\{a_h\notin\operatorname{supp}\left(\pi^{\mathsf{E}}(\cdot|s_h)\right)\right\}+r_h^{\theta}(s_h,a_h)-V_h(s_h)\right)^2\Big|s_h,a_h\right]\right]\\
&=\sum_{h=1}^{H}\mathbb{E}_{\pi^{\mathsf{E}}}\left[\left(V_{h+1}(s_{h+1})+A_h(s_h,a_h)\cdot\mathbf{1}\left\{a_h\notin\operatorname{supp}\left(\pi^{\mathsf{E}}(\cdot|s_h)\right)\right\}+r_h^{\theta}(s_h,a_h)-V_h(s_h)\right)^2\right]\\
&\overset{\text{(ii)}}{=}\sum_{h=1}^{H}\mathbb{E}_{\pi^{\mathsf{E}}}\left[\left(V_{h+1}(s_{h+1})+r_h^{\theta}(s_h,a_h)-V_h(s_h)\right)^2\right]\\
&=\mathbb{E}_{\pi^{\mathsf{E}}}\left[\left(\sum_{h=1}^{H}\left(V_{h+1}(s_{h+1})+r_h^{\theta}(s_h,a_h)-V_h(s_h)\right)\right)^2\right]\\
&+2\sum_{1\le h<h'\le H}\mathbb{E}_{\pi^E}\left[\left(V_{h+1}(s_{h+1})+r_h^{\theta}(s_h,a_h)-V_h(s_h)\right)\cdot\left(V_{h'+1}(s_{h'+1})+r^{\theta}(s'_h,a'_h)-V_h(s'_h)\right)\right]\\
&\overset{\text{(iii)}}{=}\mathbb{E}_{\pi^{\mathsf{E}}}\left[\left(\sum_{h=1}^{H}\left(V_{h+1}(s_{h+1})+r_h^{\theta}(s_h,a_h)-V_h(s_h)\right)\right)^2\right]\\
&=\mathbb{E}_{\pi^{\mathsf{E}}}\left[\left(\sum_{h=1}^{H}r_h^{\theta}(s_h,a_h)+\sum_{h=1}^{H}\left(V_{h+1}(s_{h+1})-V_h(s_h)\right)\right)^2\right]\\
&=\mathbb{E}_{\pi^{\mathsf{E}}}\left[\left(\sum_{h=1}^{H}r_h^{\theta}(s_h,a_h)-V_1(s_1)\right)^2\right]\\
&\overset{\text{(iv)}}{=}\mathrm{Var}_{\pi^{\mathsf{E}}}\left(\sum_{h=1}^{H}r_h^{\theta}(s_h,a_h)\right)\le H^2.,
\end{aligned}
\tag{81}
$$

where (i) is by definition of reward mapping $r_h^{\theta}(s,a)=-A_h(s,a)\cdot\mathbf{1}\left\{a\in\operatorname{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\}+V_h(s)-[\mathbb{P}_hV_{h+1}](s,a)$, (ii) comes from

$$
\mathbf{1}\left\{a_h\notin\operatorname{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s_h)\right)\right\}=0
$$

for any $(s_h,a_h)\in\operatorname{supp}\left(d_h^{\pi^{\mathsf{E}}}(\cdot)\right)$, (iii) is valid since

$$
\begin{aligned}
&\mathbb{E}_{\pi^E}\left[\left(V_{h+1}(s_{h+1})+r_h^{\theta}(s_h,a_h)-V_h(s_h)\right)\cdot\left(V_{h'+1}(s_{h'+1})+r^{\theta}(s'_h,a'_h)-V_h(s'_h)\right)\right]\\
&=\mathbb{E}_{\pi^{\mathsf{E}}}\left[\left(V_{h+1}(s_{h+1})+r_h^{\theta}(s_h,a_h)-V_h(s_h)\right)\mathbb{E}_{d^{\pi^{\mathsf{E}}}}\left[V_{h'+1}(s_{h'+1})-V_{h'}(s_{h'})+r_{h'}^{\theta}(s_{h'},a_{h'})|\mathcal{F}_{h+1}\right]\right]=0,
\end{aligned}
\tag{82}
$$

$(\mathcal{F}_{h+1})$ and (iv) is by $\Theta\in\overline{\Theta}$. Plugging Eq.(81) into Eq.(80), we deduce that

$$
\text{(I.a)}\le\sqrt{\frac{C^{\star}H^3S\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K}}.
\tag{83}
$$

Combining Eq.(83), Eq.(72), Eq.(73) and Eq.(74), we have

Algorithm 6 FRAMEWORK FOR OFFLINE INVERSE REINFORCEMENT LEARNING

```
Input: Dataset $\mathcal{D}$ collected by executing $\pi^{\mathsf{b}}$ in $\mathcal{M}$.
Recover the transition dynamics $\widehat{\mathbb{P}} : [H] \times \mathcal{S} \times \mathcal{A} \rightarrow \Delta\mathcal{S}$ and expert policy $\widehat{\pi}^{\mathsf{E}} = \left\{\widehat{\pi}^{\mathsf{E}}_h : \mathcal{S} \times \Delta(\mathcal{S})\right\}$ and design the bonus $b : [H] \times \mathcal{S} \times \mathcal{A} \times \Theta \rightarrow \mathbb{R}_{\geq 0}$ .
Compute $\widehat{\mathscr{R}}$ by
    $[\widehat{\mathscr{R}}(V, A)]_h(s, a) = -A_h(s, a) \cdot \mathbf{1}\left\{a \notin \mathrm{supp}\Big(\widehat{\pi}^{\mathsf{E}}_h(\cdot|s)\Big)\right\} + V_h(s) - [\widehat{\mathbb{P}}_h V_{h+1}](s, a) - b^{\theta}_h(s, a) \quad (84)$
Output: Estimated reward mapping $\widehat{\mathscr{R}}$.
```

$$
\begin{aligned}
\text{(I)} &\lesssim \text{(I.a)} + \text{(I.b)} + \text{(I.c)} + \text{(I.d)} \\
&\lesssim \sqrt{\frac{C^{\star} H^3 S \log \mathcal{N}(\Theta; \epsilon/H)\iota}{K}} + \sqrt{\frac{C^{\star} H^2 S \log \mathcal{N}(\Theta; \epsilon/H)}{K}} + \frac{C^{\star} H^3 S \log \mathcal{N}(\Theta; \epsilon/H)\iota^{5/2}}{K} \\
&\quad + \frac{C^{\star} H^2 S \log \mathcal{N}(\Theta; \epsilon/H)\iota}{K} + \epsilon \cdot (1 + \epsilon\sqrt{\frac{C^{\star} S \log \mathcal{N}(\Theta; \epsilon/H)\iota}{K}}) \\
&\lesssim \sqrt{\frac{C^{\star} H^3 S \log \mathcal{N}(\Theta; \epsilon/H)\iota}{K}} + \frac{C^{\star} H^3 S \log \mathcal{N}(\Theta; \epsilon/H)\iota^{5/2}}{K} + \epsilon.
\end{aligned}
$$

Pligging into Eq.(79), the final bound is given by

$$
D^{\pi^{\mathsf{eval}}}_{\Theta}(\mathscr{R}^{\star}, \widehat{\mathscr{R}}) \lesssim \frac{C^{\star} H^2 S A \iota}{K} + \sqrt{\frac{C^{\star} H^3 S \log \mathcal{N}(\Theta; \epsilon/H)\iota}{K}} + \frac{C^{\star} H^3 S \log \mathcal{N}(\Theta; \epsilon/H)\iota^{5/2}}{K} + \epsilon
$$

The right-hand-side is upper bounded by 2ϵ as long as

$$
K \geq \widetilde{\mathcal{O}}\bigg(\frac{C^{\star} H^3 S \log \mathcal{N}(\Theta; \epsilon/H)}{\epsilon^2} + \frac{C^{\star} H^2 S A}{\epsilon}\bigg).
$$

Here $poly \log(H, S, A, 1/\delta)$ are omitted.

□

E.4. Framework for offline inverse reinforcement learning

Pessimism As shown in Eq.(84), that estimator reward mapping involves a penalty term $b^{\theta}_h(s, a)$. The reason for introducing the penalty term $b^{\theta}_h(s, a)$ is to ensure that our reward satisfies the monotonicity condition: $\Big[\widehat{\mathscr{R}}(V, A)\Big]_h(s, a) \leq \Big[\widehat{\mathscr{R}}(V, A)\Big]_h(s, a)$, which is crucial for the guarantee of the performance of RL algorithms with learned rewards, as demonstrated in Proposition 6 and Corollary I.5.

Condition E.4 . With probability at least $1 - \delta$, we have $\sup_{(V,A)\in\Theta} \Big|\Big[(\mathbb{P}_h - \widehat{\mathbb{P}}_h)V_{h+1}\Big](s, a)\Big| \leq b^{\theta}_h(s, a)$ and $\mathrm{supp}\left(\widehat{\pi}^{\mathsf{E}}_h(\cdot|s)\right) \subset \mathrm{supp}\left(\pi^{\mathsf{E}}_h(\cdot|s)\right)$ for all $(h, s) \in [H] \times \mathcal{S}$ and all $(V, A) \in \Theta$.

Theorem E.5 (Learning bound for Algorithm 6)**.** *Suppose that Condition E.4 holds. With probability at least* $1 - \delta$*, we have* $\Big[\widehat{\mathscr{R}}(V, A)\Big]_h(s, a) \leq [\mathscr{R}^{\star}(V, A)]_h(s, a)$ *for all* $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$*, and*

$$
\begin{aligned}
D^{\pi^{\mathsf{eval}}}_{\Theta}\Big(\mathscr{R}^{\star}, \widehat{\mathscr{R}}\Big) \leq \sup_{\theta\in\Theta} \Bigg\{ & H \cdot \sum_{h\in[H]} \mathbb{E}_{(s,a)\sim d^{\pi^{\mathsf{eval}}}_h}\left[\mathbf{1}\left\{a \in \mathrm{supp}\left(\pi^{\mathsf{E}}_h(\cdot|s)\right), a \notin \mathrm{supp}\left(\widehat{\pi}^{\mathsf{E}}_h(\cdot|s)\right)\right\}\right] \\
& + 2 \sum_{h\in[H]} \mathbb{E}_{(s,a)\sim d^{\pi^{\mathsf{eval}}}_h}\left[b^{\theta}_h(s, a)\right] \Bigg\}. \qquad (85)
\end{aligned}
$$

Proof. When $\sup_{(V,A)\in\Theta}\Big|\Big[(\mathbb{P}_h - \widehat{\mathbb{P}}_h)V_{h+1}\Big](s,a)\Big| \leq b_h^\theta(s,a)$ and $\mathrm{supp}\,\widehat{\pi}_h^{\mathsf{E}}(\cdot|s) \subset \mathrm{supp}\,\pi_h^{\mathsf{E}}(\cdot|s)$ holds for all $(h,s) \in [H]\times\mathcal{S}$ and all $(V,A)\in\Theta$ hold, we have

$$
\begin{aligned}
&\Big[\widehat{\mathscr{R}}(V,h)\Big]_h(s,a) - [\mathscr{R}^\star(V,h)]_h(s,a)\\
&= -A_h(s,a)\cdot\Big[\mathbf{1}\Big\{a\notin\mathrm{supp}\big(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\big)\Big\} - \mathbf{1}\Big\{a\notin\mathrm{supp}\big(\pi_h^{\mathsf{E}}(\cdot|s)\big)\Big\}\Big] - \Big[\Big(\widehat{\mathbb{P}}_h - \mathbb{P}_h\Big)V_{h+1}\Big](s,a) - b_h^\theta(s,a)\\
&= \underbrace{-A_h(s,a)\cdot\mathbf{1}\Big\{a\in\mathrm{supp}\,\big(\pi_h^{\mathsf{E}}(\cdot|s)\big), a\notin\mathrm{supp}\,\big(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\big)\Big\}}_{\leq 0}\,\underbrace{-\Big[\Big(\widehat{\mathbb{P}}_h - \mathbb{P}_h\Big)V_{h+1}\Big](s,a) - b_h^\theta(s,a)}_{\leq 0} \leq 0,
\end{aligned}
\tag{86}
$$

where the second line is by $\mathrm{supp}\,\big(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\big) \subset \mathrm{supp}\,\big(\pi_h^{\mathsf{E}}(\cdot|s)\big)$ and $\sup_{(V,A)\in\Theta}\Big|\Big[(\mathbb{P}_h - \widehat{\mathbb{P}}_h)V_{h+1}\Big](s,a)\Big| \leq b_h^\theta(s,a)$. Further, by triangle inequality, we obtain that

$$
\begin{aligned}
&\Big|\Big[\widehat{\mathscr{R}}(V,h)\Big]_h(s,a) - [\mathscr{R}^\star(V,h)]_h(s,a)\Big|\\
&\leq A_h(s,a)\cdot\mathbf{1}\Big\{a\in\mathrm{supp}\,\big(\pi_h^{\mathsf{E}}(\cdot|s)\big), a\notin\mathrm{supp}\,\big(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\big)\Big\} + \Big|\Big[\Big(\widehat{\mathbb{P}}_h - \mathbb{P}_h\Big)V_{h+1}\Big](s,a)\Big| + b_h^\theta(s,a)\\
&\leq H\cdot\mathbf{1}\Big\{a\in\mathrm{supp}\,\big(\pi_h^{\mathsf{E}}(\cdot|s)\big), a\notin\mathrm{supp}\,\big(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\big)\Big\} + 2b_h^\theta(s,a),
\end{aligned}
\tag{87}
$$

where the last line is due to $A_h(s,a)\leq H$ and $\Big|\Big[\Big(\widehat{\mathbb{P}}_h - \mathbb{P}_h\Big)V_{h+1}\Big](s,a)\Big| \leq b_h^\theta(s,a)$. Similar to the proof of Lemma E.2, we have

$$
d^{\pi^{\mathsf{eval}}}\Big(\widehat{\mathscr{R}}(V,A), \mathscr{R}^\star(V,A)\Big) \leq \sum_{h\in[H]}\mathbb{E}_{(s,a)\sim d_h^{\pi^{\mathsf{eval}}}}\Big[\Big|\Big[\widehat{\mathscr{R}}(V,h)\Big]_h(s,a) - [\mathscr{R}^\star(V,h)]_h(s,a)\Big|\Big]. \tag{88}
$$

Combining Eq.(87) and Eq.(88), we obtain that

$$
\begin{aligned}
d^{\pi^{\mathsf{eval}}}\Big(\widehat{\mathscr{R}}(V,A), \mathscr{R}^\star(V,A)\Big) \leq H\cdot&\sum_{h\in[H]}\mathbb{E}_{(s,a)\sim d_h^{\pi^{\mathsf{eval}}}}\Big[\mathbf{1}\Big\{a\in\mathrm{supp}\,\big(\pi_h^{\mathsf{E}}(\cdot|s)\big), a\notin\mathrm{supp}\,\big(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\big)\Big\}\Big]\\
&+ 2\sum_{h\in[H]}\mathbb{E}_{(s,a)\sim d_h^{\pi^{\mathsf{eval}}}}\big[b_h^\theta(s,a)\big].
\end{aligned}
\tag{89}
$$

By the definition of $D_\Theta^{\pi^{\mathsf{eval}}}$: $D_\Theta^{\pi^{\mathsf{eval}}}\Big(\mathscr{R}^\star, \widehat{\mathscr{R}}\Big) = \sup_{\theta\in\Theta} d^{\pi^{\mathsf{eval}}}\Big(\widehat{\mathscr{R}}(V,A), \mathscr{R}^\star(V,A)\Big)$, we complete the proof. □

By Theorem E.5, all we need to do is design b_h^θ and learn $\widehat{\mathbb{P}}$, $\widehat{\pi}^{\mathsf{E}}$ from the data to satisfy Condition E.4, thereby obtaining an IRL algorithm. The crux of the problem lies in the design of $b_h^\theta, \widehat{\mathbb{P}}$ and $\widehat{\pi}^{\mathsf{E}}$. In RLP, we employ the pessimism technique from offline RL, and the construction of b_h^θ and $\widehat{\pi}^{\mathsf{E}}$ using pessimism in RLP satisfies Condition E.4, as illustrated in the proof of Theorem 8.

F. Proofs for Section 5

F.1. Full description of Reward Learning with Exploration

We propose a meta-algorithm, named Reward Learning with Exploration (RLE). The pseudocode of RLE is presented in Algorithm 2, where the algorithm contains the following three main components:

- Exploring the unknown environment: This segment involves computing a desired behavior policy $\pi^{\mathsf{b}} = \mathbb{E}_{\pi\sim\mu^{\mathsf{b}}}[\pi]$, which takes the form of a finite mixture of deterministic policies. To achieve this, we need to collect NH episodes of samples. We then execute this policy to gather a total of K episodes worth of samples. Our exploration approach is based on leveraging the exploration scheme outlined in Li et al. (2023, Algorithm 1). A comprehensive description of this exploration method is postponed and will be provided in Section C.
- Subsampling: For the sake of theoretical simplicity, we apply subsampling. For each $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$, we populate the new dataset with $\min\Big\{\widehat{N}_h^{\mathsf{b}}(s,a), N_h(s,a)\Big\}$ sample transitions. Here, $\widehat{N}_h^{\mathsf{b}}(s,a)$, as defined in Eq.(13), acts as a lower bound on the total number of visits to (h,s,a) among these K sample episodes, with high probability.

- Computing estimated reward mapping: With the previously collected dataset at hand, we then utilize the offline IRL algorithm RLP to compute the desired reward mapping.

We remark that our algorithm RLE follows a similar approach to that of Li et al. (2023, Algorithm 1). We begin by computing a desired behavior policy π^{b}, then proceed to collect data, and finally compute results through the invocation of an offline algorithm. In contrast to the offline setting, we have the flexibility to select the desired behavior. In the following, we will observe that the behavior policy π^{b} exhibits concentrability with *any* deterministic policy, as shown in Eq.(16). This property enables us to achieve our learning goal within the online setting.

F.2. Proof of Theorem 11

Lemma F.1 (Li et al. (2023))**.** *Recall that $\xi = c_\xi H^3S^3A^3\log\frac{10HSA}{\delta}$ for some large enough constant $c_\xi > 0$ (see line 3 in Algorithm 3). Then, with probability at least $1-\delta$, the estimated occupancy distributions specified in Eq.*(14) *and* (15) *of Algorithm 3 satisfy*

$$\frac{1}{2}\widehat{d}_h^\pi(s,a) - \frac{\xi}{4N} \le d_h^\pi(s,a) \le 2\widehat{d}_h^\pi(s,a) + 2e_h^\pi(s,a) + \frac{\xi}{4N} \tag{90}$$

simultaneously for all $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$ and all deterministic Markov policy $\pi\in\Pi^{\mathrm{det}}$, provided that

$$KH \ge N \ge C_N\sqrt{H^9S^7A^7K}\log\frac{10HSA}{\delta} \qquad \textit{and} \qquad K \ge C_K HSA \tag{91}$$

for some large enough constants $C_N, C_K > 0$, where, $\{e_h^\pi(s,a)\in\mathbb{R}_+\}$ satisfies that

$$\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} e_h^\pi(s,a) \le \frac{2SA}{K} + \frac{13SAH\xi}{N} \lesssim \sqrt{\frac{SA}{HK}} \qquad \forall h\in[H], \pi\in\Pi^{\mathrm{det}} \tag{92}$$

Notice that Eq.(90) only holds for $\pi\in\Pi^{\mathrm{det}}$, however, we will show a similar result also holds for any stochastic policy. For any stochastic policy $\pi = \mathbb{E}_{\pi'\sim\mu}[\pi']$ ($\mu\in\Delta(\Pi^{\mathrm{det}})$), the visitation distribution $\{d_h^\pi\}$ can be expressed as

$$d_h^\pi(s,a) = \mathbb{E}_{\pi'\sim\mu}\Big[d_h^{\pi'}(s,a)\Big], \qquad \forall (h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A},$$

. We can define $\widehat{d}^\pi$ as

$$\widehat{d}_h^\pi(s,a) = \mathbb{E}_{\pi'\sim\mu}\Big[\widehat{d}_h^{\pi'}(s,a)\Big], \qquad \forall (h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A},$$

where $\Big\{d_h^{\pi'}\Big\}$ are the estimated occupancy distributions in Algorithm 3.

By Eq.(90), we have

$$\begin{aligned}\frac{1}{2}\widehat{d}_h^\pi(s,a) - \frac{\xi}{4N} \le d_h^\pi(s,a) = \mathbb{E}_{\pi'\sim\mu}\Big[d_h^{\pi'}(s,a)\Big] \le 2\widehat{d}_h^\pi(s,a) + 2\mathbb{E}_{\pi'\sim\mu}\Big[e_h^{\pi'}(s,a)\Big] + \frac{\xi}{4N}\\ \frac{1}{2}\widehat{d}_h^\pi(s,a) - \frac{\xi}{4N} \le d_h^\pi(s,a) = \mathbb{E}_{\pi'\sim\mu}\Big[d_h^{\pi'}(s,a)\Big] \le 2\widehat{d}_h^\pi(s,a) + 2\mathbb{E}_{\pi'\sim\mu}\Big[e_h^{\pi'}(s,a)\Big] + \frac{\xi}{4N}.\end{aligned}$$

$$\Big(e_h^\pi(s,a) := \mathbb{E}_{\pi'\sim\mu}\Big[e_h^{\pi'}(s,a)\Big]\Big)$$

We also have

$$\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} e_h^\pi(s,a) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbb{E}_{\pi'\sim\mu}\Big[e_h^{\pi'}(s,a)\Big] \le \frac{2SA}{K} + \frac{13SAH\xi}{N} \lesssim \sqrt{\frac{SA}{HK}},$$

provided Eq.(91).

Different from previous sections, we set $\iota = \log\frac{10HSA}{\delta}$.

Lemma F.2 (Concentration event)**.** *Suppose Eq.*(91)*. Under the setting of Theorem 11, there exists an absolute constants* C_1*,* $C_2 \geq 2$ *such that the concentration event* $\mathcal{E}$ *holds with probability at least* $1-\delta$*, where*

$$\mathcal{E} := \Bigg\{ \textit{(i):} \left|\left[(\mathbb{P}_h - \widehat{\mathbb{P}}_h)V_{h+1}\right](s,a)\right| \leq b_h^{\theta}(s,a) \quad \forall\, \theta = (V,A) \in \Theta,\ (h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A},$$
$$\textit{(ii):}\ \frac{1}{2}\widehat{d}_h^{\pi}(s,a) - \frac{\xi}{4N} \leq d_h^{\pi}(s,a) \leq 2\widehat{d}_h^{\pi}(s,a) + 2e_h^{\pi}(s,a) + \frac{\xi}{4N} \quad \forall (h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}, \pi \in \Pi,$$
$$\textit{(iii):}\ \widehat{N}_h^b(s,a) \leq N_h^b(s,a) \quad \forall (h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A},$$
$$\textit{(iv):}\ \widehat{N}_h^e(s,a) \geq 1 \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A} \textit{ s.t. } \widehat{N}_h^b(s,a) \geq \max\{C_2\eta\iota, 1\} \Bigg\}$$

where $b_h^{\theta}(s,a)$ *is defined in Eq.*(8)*,* $N_h^b(s,a)$ $\widehat{N}_h^b(s,a)$ *is defined in Eq.*(13)*,* η *are specified in Lemma E.1 and* $\widehat{N}_h^e(s,a)$ *is given by*

$$\widehat{N}_h^e(s,a) := \begin{cases} \sum_{(s_h,a_h,e_h)\in\mathcal{D}^{\mathsf{trim}}} \mathbf{1}\{(s_h,e_h) = (s,a)\} & \textit{in option } 1, \\ \widehat{N}_h^b(s,a) & \textit{in option } 2, \end{cases}$$

Proof. First, we observe that Claim (i) can be proved to hold with probability at least $1-\delta/10$ by repeating a similar argument as in Lemma E.1. By Lemma F.1, Claim (ii) holds with probability at least $1-\delta/10$. Claim (iii) has been shown to hold with probability at least $1-\delta/10$ in the proof of Li et al. (2023, Theorem 2).

Next, we focus on (iv). For claim (iv), in option 1, we have

$$\begin{aligned} \mathbb{P}\Big(\widehat{N}_h^e(s,a) = 0\Big) = \big(1-\pi_h^{\mathsf{E}}(a|s)\big)^{\widehat{N}_h^b(s,a)} &\leq \exp\Big(\widehat{N}_h^b(s,a)\log(1-\eta)\Big) \\ &\leq \exp\left(\log\frac{\delta}{4HSA}\right) = \frac{\delta}{4HSA}, \end{aligned}$$

for all $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$. The last line is valid since

$$\widehat{N}_h^b(s,a)\log(1-\eta) \leq C_2\log\frac{\delta}{HSA}\cdot\frac{\log(1-\eta)}{\eta} \leq \log\frac{\delta}{4HSA},$$

holds for sufficiently large constant C_2. In option 2, we have

$$\widehat{N}_h^e(s,a) = \widehat{N}_h^b(s,a) \geq \max\{C_2\eta\iota, 1\} \geq 1,$$

for all $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$. This completes the proof.

□

F.3. Proof of Theorem 11

Define

$$\mathcal{I}_h = \left\{(s,a) \in \mathcal{S}\times\mathcal{A} \,|\, \mathbb{E}_{\pi'\sim\mu_{\mathsf{b}}}\left[\widehat{d}_h^{\pi'}(s,a)\right] \geq \frac{\xi}{N} + \frac{4(C_2\eta+3)\iota}{K}\right\}, \tag{93}$$

for all $h \in [H]$. Then for $(s,a) \in \mathcal{I}_h$, we have

$$\widehat{N}_h^b(s,a) \geq \frac{K}{4}\mathbb{E}_{\pi'\sim\mu_{\mathsf{b}}}\left[\widehat{d}_h^{\pi'}(s,a)\right] - \frac{K\xi}{8N} - 3\iota \geq C_2\eta\iota. \tag{94}$$

By concentration event $\mathcal{E}$ (iv), we have

$$\widehat{N}_h^e(s,a) \geq 1,$$

By construction of $\widehat{\pi}^{\mathsf{E}}$ in Algorithm 1, we deduce that

$$\left|\mathbf{1}\left\{a \in \operatorname{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\} - \mathbf{1}\left\{a \in \operatorname{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\}\right| = 0. \tag{95}$$

for all $(s, a) \in \mathcal{I}_h$.
With $\mathcal{I}_h$ at hand, we can decompose the $d^\pi\left(r^\theta, \widehat{r}^\theta\right)$ for any π and $\theta \in \Theta$ as follows:

$$
\begin{aligned}
d^\pi\left(r_h^\theta, \widehat{r}_h^\theta\right) &\leq \sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} d_h^\pi(s,a)\cdot\left|r_h^\theta(s,a)-\widehat{r}_h^\theta(s,a)\right| \\
&\leq \underbrace{\sum_{h\in[H]}\sum_{(s,a)\notin\mathcal{I}_h} d_h^\pi(s,a)\cdot\left|r_h^\theta(s,a)-\widehat{r}_h^\theta(s,a)\right|}_{\text{(I)}} + \underbrace{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h} d_h^\pi(s,a)\cdot\left|r^\theta(s,a)-\widehat{r}_h^\theta(s,a)\right|}_{\text{(II)}}, \qquad (96)
\end{aligned}
$$

where the first line follows the same argument in the proof of Lemma E.2. We then study the terms (I) and (II) separately. For the term (I), by the construction of Algorithm 1, we obtain that

$$
\begin{aligned}
\text{(I)} &= \sum_{h\in[H]}\sum_{(s,a)\notin\mathcal{I}_h} d_h^\pi(s,a)\cdot\left|r^\theta(s,a)-\widehat{r}^\theta(s,a)\right| \\
&= \sum_{h\in[H]}\sum_{(s,a)\notin\mathcal{I}_h} d_h^\pi(s,a)\cdot\Bigg| -A_h(s,a)\Big(\mathbf{1}\left\{a\in\text{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\}-\mathbf{1}\left\{a\in\text{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\}\Big) \\
&\quad -\left[\left(\mathbb{P}_h-\widehat{\mathbb{P}}_h\right)V_{h+1}\right](s,a)-b_h^\theta(s,a)\Bigg\| \\
&\leq \sum_{h\in[H]}\sum_{(s,a)\notin\mathcal{I}_h} d_h^\pi(s,a)\cdot\Big\{\big|A_h(s,a)\cdot\big(\mathbf{1}\left\{a\in\text{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\}-\cdot\mathbf{1}\left\{a\in\text{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\}\big)\big| \\
&\quad +\left|\left[(\mathbb{P}_h-\widehat{\mathbb{P}}_h)V_{h+1}\right](s,a)\right|+b_h^\theta(s,a)\Big\} \qquad \text{(by triangle inequality)} \\
&\overset{\text{(i)}}{\lesssim} H\cdot\sum_{h\in[H]}\sum_{(s,a)\notin\mathcal{I}_h} d_h^\pi(s,a) \\
&\overset{\text{(ii)}}{\lesssim} H\cdot\sum_{h\in[H]}\sum_{(s,a)\notin\mathcal{I}_h}\left(2\widehat{d}_h^\pi(s,a)+2e_h^\pi(s,a)+\frac{\xi}{4N}\right) \\
&\overset{\text{(iii)}}{\lesssim} H\cdot\sum_{h\in[H]}\sum_{(s,a)\notin\mathcal{I}_h}\widehat{d}_h^\pi(s,a)+\frac{\xi H^2SA}{N}+\sqrt{\frac{HSA}{K}} \\
&= H\cdot\sum_{h\in[H]}\sum_{(s,a)\notin\mathcal{I}_h}\frac{\widehat{d}_h^\pi(s,a)}{\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\left[\widehat{d}_h^{\pi'}(s,a)\right]+\frac{1}{KH}}\cdot\left(\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\left[\widehat{d}_h^{\pi'}(s,a)\right]+\frac{1}{KH}\right)+\frac{\xi H^2SA}{N}+\sqrt{\frac{HSA}{K}} \\
&\overset{\text{(iv)}}{\lesssim}\left(\frac{\xi H}{N}+\frac{4H(C_2\eta+3)\iota}{K}+\frac{1}{K}\right)\sum_{h\in[H]}\sum_{(s,a)\notin\mathcal{I}_h}\frac{\widehat{d}_h^\pi(s,a)}{\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\left[\widehat{d}_h^{\pi'}(s,a)\right]+\frac{1}{KH}}+\frac{\xi H^2SA}{N}+\sqrt{\frac{HSA}{K}} \\
&\lesssim\left(\frac{H\xi}{N}+\frac{4H(C_2\eta+3)\iota}{K}+\frac{1}{K}\right)\cdot HSA+\frac{\xi H^2SA}{N}+\sqrt{\frac{HSA}{K}} \\
&\asymp\frac{\xi H^2SA}{N}+\frac{H^2SA\eta\iota}{K}+\frac{HSA}{K}+\sqrt{\frac{HSA}{K}}, \qquad (97)
\end{aligned}
$$

where (i) is by $\|A_h\|_\infty$, $\|V_{h+1}\|_\infty, b_h^\theta(s,a) \le H$, (ii) comes from concentration $\mathcal{E}$(ii), (iii) comes from Eq.(90), and (iv) is by definition of $\mathcal{I}_h$. For the term (I), conditioning on the concentration event $\mathcal{E}$, we have

$$\begin{aligned}
\text{(II)} &= \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{I}_h} d_h^\pi(s,a) \cdot \left| r_h^\theta(s,a) - \widehat{r}_h^\theta(s,a) \right| \\
&\le \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{I}_h} d_h^\pi(s,a) \cdot \left| \Big[(\mathbb{P}_h - \widehat{\mathbb{P}}_h) V_{h+1} \Big](s,a) - b_h^\theta(s,a) \right| \\
&\le 2 \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{I}_h} d_h^\pi(s,a) \cdot b_h^\theta(s,a) \\
&\le \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{I}_h} \left(4\widehat{d}_h^\pi(s,a) + 4e_h^\pi(s,a) + \frac{\xi}{2N} \right) \cdot b_h^\theta(s,a) \\
&\lesssim \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{I}_h} \widehat{d}_h^\pi(s,a) \cdot b_h^\theta(s,a) + H \cdot \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{I}_h} \left(\frac{\xi}{N} + e_h^\pi(s,a) \right) \\
&\lesssim \frac{\xi H^2 SA}{N} + \sqrt{\frac{HSA}{K}} + \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{I}_h} \widehat{d}_h^\pi(s,a) \cdot b_h^\theta(s,a), \qquad (98)
\end{aligned}$$

where the second line is by construction of Algorithm 1, the second last line is by $b_h^\theta(s,a) \lesssim H$, the last follows from (90).

Further, we decompose the second term of Eq.(98) for any $\theta \in \Theta$ by

$$
\begin{aligned}
&\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h} \widehat{d}_h^{\pi}(s,a)\cdot b_h^{\theta}(s,a) \\
&= \sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h} \widehat{d}_h^{\pi}(s,a)\cdot \min\Bigg\{\sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\widehat{N}_h^b(s,a)\vee 1}\Big[\widehat{\mathbb{V}}_h V_{h+1}\Big](s,a)} + \frac{H\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\widehat{N}_h^b(s,a)\vee 1} \\
&+ \frac{\epsilon}{H}\Bigg(1+\sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\widehat{N}_h^b(s,a)\vee 1}}\Bigg), H\Bigg\} \\
&\overset{\text{(i)}}{\leq} \sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h} \widehat{d}_h^{\pi}(s,a)\cdot \Bigg\{\min\Bigg\{\sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\widehat{N}_h^b(s,a)\vee 1}\Big[\widehat{\mathbb{V}}_h V_{h+1}\Big](s,a)}, H\Bigg\} \\
&+ \frac{H\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\widehat{N}_h^b(s,a)\vee 1} + \frac{\epsilon}{H}\Bigg(1+\sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\widehat{N}_h^b(s,a)\vee 1}}\Bigg)\Bigg\} \\
&\overset{\text{(ii)}}{\leq} \sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h} \widehat{d}_h^{\pi}(s,a)\cdot \Bigg\{\sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota\Big[\widehat{\mathbb{V}}_h V_{h+1}\Big](s,a) + H}{\widehat{N}_h^b(s,a)\vee 1 + 1/H}} + \frac{H\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\widehat{N}_h^b(s,a)\vee 1} \\
&+ \frac{\epsilon}{H}\Bigg(1+\sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\widehat{N}_h^b(s,a)\vee 1}}\Bigg)\Bigg\} \\
&\overset{\text{(iii)}}{=} \underbrace{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h} \widehat{d}_h^{\pi}(s,a)\cdot \sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota\Big[\widehat{\mathbb{V}}_h V_{h+1}\Big](s,a) + H}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big] + 1/H}}}_{\text{(II.a)}} \\
&+ \underbrace{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h} \widehat{d}_h^{\pi}(s,a)\cdot \frac{H\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big] + 1/H}}_{\text{(II.b)}} \\
&+ \underbrace{\frac{\epsilon}{H}\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h} \widehat{d}_h^{\pi}(s,a)\cdot \Bigg(1+\sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big] + 1/H}}\Bigg)}_{\text{(II.c)}} \qquad (99)
\end{aligned}
$$

where the (i) is by inequality $\min\{a+b, c\} \leq \min\{a, c\} + b$ $(a, b, c \geq 0)$, (ii) comes from inequality $\min\left\{\frac{x}{y}, \frac{z}{w}\right\} \leq \frac{x+z}{y+w}$ and (iii) is valid since

$$
\widehat{N}_h^b(s,a) = \left[\frac{K}{4}, \mathop{\mathbb{E}}_{\pi\sim\mu^{\mathsf{b}}}[\widehat{d}_h^{\pi}(s,a)] - \frac{K\xi}{8N} - 3\log\frac{HSA}{\delta}\right]_+ \gtrsim K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big] + 1/H
$$

holds for all $(s,a) \in \mathcal{I}_h$ according to definition of $\mathcal{I}$. We then study the three terms separately. For the term (II.a), by the

Cauchy-Schwarz inequality, we have

$$\text{(II.a)} \leq \underbrace{\left\{ \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \widehat{d}_h^{\pi}(s,a) \cdot \left[\log\mathcal{N}(\Theta;\epsilon/H)\iota[\mathbb{V}_h V_{h+1}](s,a) + H\right] \right\}^{1/2}}_{\text{(II.a.1)}}$$
$$\times \underbrace{\left\{ \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\widehat{d}_h^{\pi}(s,a)}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\left[\widehat{d}_h^{\pi'}(s,a)\right] + 1/H} \right\}^{1/2}}_{\text{(II.a.2)}}.$$

Observe that $\|V_{h+1}\|_\infty \leq H$, then the term (II.a.1) can be upper bounded by

$$\begin{aligned}\text{(II.a.1)} &= \sqrt{\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \widehat{d}_h^{\pi}(s,a) \cdot \left[\log\mathcal{N}(\Theta;\epsilon/H)\iota[\mathbb{V}_h V_{h+1}](s,a) + H\right]} \\ &\leq \sqrt{[H^2\log\mathcal{N}(\Theta;\epsilon/H)\iota + H] \cdot \sqrt{\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \widehat{d}_h^{\pi}(s,a)}} \asymp \sqrt{H^3\log\mathcal{N}(\Theta;\epsilon/H)\iota}. \end{aligned} \tag{100}$$

For the term (II.a.2), we have

$$\begin{aligned}\text{(II.a.2)} &= \sqrt{\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\widehat{d}_h^{\pi}(s,a)}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\left[\widehat{d}_h^{\pi'}(s,a)\right] + 1/H}} \\ &= \sqrt{\frac{1}{K} \cdot \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\widehat{d}_h^{\pi}(s,a)}{\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\left[\widehat{d}_h^{\pi'}(s,a)\right] + 1/KH}} \\ &\lesssim \sqrt{\frac{HSA}{K}}, \end{aligned} \tag{101}$$

which the last line comes from Eq.(16). Combining Eq.(100) and (101), we conclude that

$$\text{(II.a)} \lesssim \sqrt{\frac{H^4SA}{K}}. \tag{102}$$

For the term (II.b), by Eq.(16), we have

$$\begin{aligned}\text{(II.b)} &= \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \widehat{d}_h^{\pi}(s,a) \cdot \frac{H\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\left[\widehat{d}_h^{\pi'}(s,a)\right] + 1/H} \\ &= \frac{1}{K} \cdot \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \widehat{d}_h^{\pi}(s,a) \cdot \frac{H\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\left[\widehat{d}_h^{\pi'}(s,a)\right] + 1/KH} \\ &\lesssim \frac{H^2SA\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K}. \end{aligned} \tag{103}$$

For the term (II.c), we have

$$
\begin{aligned}
\text{(II.c)} &= \frac{\epsilon}{H}\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h}\widehat{d}_h^{\pi}(s,a)\cdot\left(1+\sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big]+1/H}}\right)\\
&= \epsilon+\frac{\epsilon}{H}\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h}\sqrt{\widehat{d}_h^{\pi}(s,a)}\cdot\sqrt{\frac{\widehat{d}_h^{\pi}(s,a)\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big]+1/H}}\\
&\le \epsilon+\frac{\epsilon}{H}\sqrt{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h}\widehat{d}_h^{\pi}(s,a)}\cdot\sqrt{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h}\frac{\widehat{d}_h^{\pi}(s,a)\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big]+1/H}}\\
&\le \epsilon(1+\sqrt{\frac{SA\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K}}),
\end{aligned}
\tag{104}
$$

where the second last line is by the Cauchy-Schwarz inequality and the last line is due to Eq.(101).

Then combining Eq.(98), Eq.(102), Eq.(103), and Eq.(104), we obtain the bound for the term (II)

$$
\begin{aligned}
\text{(II)} &\lesssim \text{(II.a)}+\text{(II.b)}+\text{(II.c)}\\
&\lesssim \sqrt{\frac{H^4SA\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K}}+\frac{H^2SA\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K}+\epsilon(1+\sqrt{\frac{SA\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K}})\\
&\lesssim \sqrt{\frac{H^4SA\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K}}+\epsilon,
\end{aligned}
\tag{105}
$$

where the last line is from $\epsilon<1$. Finally, combining Eq.(97) and (102), we get the final bound

$$
\begin{aligned}
D_{\Theta}^{\mathsf{all}}\Big(\mathscr{R}^{\star},\widehat{\mathscr{R}}\Big) = \sup_{\pi,\theta\in\Theta} d^{\pi}\left(r_h^{\theta},\widehat{r}_h^{\theta}\right) &\le \mathrm{I}+\mathrm{II}\\
&\lesssim \frac{\xi H^2SA}{N}+\frac{H^2SA\eta\iota}{K}+\sqrt{\frac{H^4SA\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K}}+\epsilon.
\end{aligned}
$$

Hence, we can guarantee $D_{\Theta}^{\mathsf{all}}\Big(\mathscr{R}^{\star},\widehat{\mathscr{R}}\Big)\le 2\epsilon$, provided that

$$
K\ge\widetilde{\mathcal{O}}\bigg(\frac{H^4SA\log\mathcal{N}(\Theta;\epsilon/H)}{\epsilon^2}+\frac{H^2SA\eta}{\epsilon}\bigg),\qquad KH\ge N\ge\widetilde{\mathcal{O}}\Big(\sqrt{H^9S^7A^7K}\Big).
$$

Here $poly\log(H,S,A,1/\delta)$ are omitted.

Suppose that $\epsilon\le H^{-9}(SA)^{-6}$. We set

$$
N=\widetilde{\mathcal{O}}\big(H^9S^7A^7K\big),\qquad K=\widetilde{\mathcal{O}}\bigg(\frac{H^4SA\log\mathcal{N}(\Theta;\epsilon/H)}{\epsilon^2}+\frac{H^2SA\eta}{\epsilon}\bigg).
\tag{106}
$$

When $\epsilon\le H^{-9}(SA)^{-6}$, we have

$$
\begin{aligned}
KH &\ge \sqrt{K}H\cdot\widetilde{\mathcal{O}}\left(\sqrt{\frac{H^4SA\log\mathcal{N}(\Theta;\epsilon/H)}{\epsilon^2}}\right)\\
&\ge \sqrt{K}H\cdot\widetilde{\mathcal{O}}\Big(H^9S^6A^6\sqrt{H^4SA}\Big)\\
&\ge \sqrt{K}H\cdot\widetilde{\mathcal{O}}\Big(\sqrt{H^9S^7A^7}\Big)\\
&\ge \widetilde{\mathcal{O}}\Big(\sqrt{H^9S^7A^7K}\Big).
\end{aligned}
\tag{107}
$$

Combining Eq.(106) and Eq.(107), we have

$$KH \geq N \geq \widetilde{\mathcal{O}}\Big(\sqrt{H^9 S^7 A^7 K}\Big). \tag{108}$$

Then, the total sample complexity is

$$K + NH \geq \widetilde{\mathcal{O}}\Bigg(\frac{H^4 SA \log \mathcal{N}(\Theta;\epsilon/H)}{\epsilon^2} + \frac{H^2 SA\eta}{\epsilon} + \sqrt{\frac{H^{15} S^8 A \log \mathcal{N}(\Theta;\epsilon/H)}{\epsilon^2}} + \sqrt{\frac{H^{13} S^8 A^8 \eta}{\epsilon}}\Bigg). \tag{109}$$

When $\epsilon \leq H^{-9}(SA)^{-6}$, we have

$$\begin{aligned}
\widetilde{\mathcal{O}}\bigg(\frac{H^4 SA \log \mathcal{N}(\Theta;\epsilon/H)}{\epsilon^2}\bigg) &\geq \widetilde{\mathcal{O}}\bigg(\frac{H^{13} S^7 A^7 \log \mathcal{N}(\Theta;\epsilon/H)}{\epsilon}\bigg) \\
&= \widetilde{\mathcal{O}}\Bigg(\sqrt{\frac{H^{26} S^{14} A^{14} \log \mathcal{N}(\Theta;\epsilon/H)^2}{\epsilon^2}}\Bigg) && (\log \mathcal{N}(\Theta;\epsilon/H) \geq 1) \\
&\geq \sqrt{\frac{H^{15} S^8 A \log \mathcal{N}(\Theta;\epsilon/H)}{\epsilon^2}} && (110)
\end{aligned}$$

and

$$\begin{aligned}
\widetilde{\mathcal{O}}\bigg(\frac{H^2 SA\eta}{\epsilon}\bigg) &\geq \widetilde{\mathcal{O}}\bigg(\frac{H^2 SA\eta}{\epsilon}\bigg) \\
&= \widetilde{\mathcal{O}}\Bigg(\sqrt{\frac{H^4 S^2 A^2 \eta^2}{\epsilon^2}}\Bigg) \\
&\geq \widetilde{\mathcal{O}}\Bigg(\sqrt{\frac{H^{13} S^8 A^8 \eta^2}{\epsilon}}\Bigg) \\
&\geq \widetilde{\mathcal{O}}\Bigg(\sqrt{\frac{H^{13} S^8 A^8 \eta}{\epsilon}}\Bigg), && (111)
\end{aligned}$$

where the last line is due to $\eta \in \{0\} \cup [1,\infty)$. Combining Eq.(109), Eq.(110), and Eq.(111), we obtain that

$$K + NH \geq \widetilde{\mathcal{O}}\bigg(\frac{H^4 SA \log \mathcal{N}(\Theta;\epsilon/H)}{\epsilon^2} + \frac{H^2 SA\eta}{\epsilon}\bigg) \tag{112}$$

holds when $\epsilon \leq H^{-9}(SA)^{-6}$

G. Lower bound in the online setting

G.1. Lower bound of online IRL problems

We focus on the case where $\Theta = \overline{\mathcal{V}} \times \overline{\mathcal{A}}$. In this case $\log \mathcal{N}(\Theta;\epsilon/H) = \widetilde{\mathcal{O}}(S)$, the upper bound of the sample complexity of Algorithm 2 becomes $\widetilde{\mathcal{O}}\big(H^4 S^2 A/\epsilon^2\big)$ (we hide the burn-in term).

Similar to the offline setting, we define (ϵ,δ)-*PAC algorithm for online IRL problems* for all $\epsilon,\delta \in (0,1)$ as follows.

Definition G.1. *Fix a parameter set Θ, we say an online IRL algorithm $\mathfrak{A}$ is a (ϵ,δ)-PAC algorithm for online IRL problems, if for any IRL problem $(\mathcal{M},\pi^{\mathsf{E}})$, with probability $1-\delta$, $\mathfrak{A}$ outputs a reward mapping $\widehat{\mathscr{R}}$ such that*

$$D^{\mathsf{all}}_{\Theta}(\widehat{\mathscr{R}},\mathscr{R}^\star) \leq \epsilon.$$

Theorem G.2 (Lower bound for online IRL problems)**.** *Fix parameter set $\Theta = \overline{\mathcal{V}} \times \overline{\mathcal{A}}$ and let $\mathfrak{A}$ be an (ϵ,δ)-PAC algorithm for online IRL problems, where $\delta \leq 1/3$. Then, there exists an IRL problem $(\mathcal{M},\pi^{\mathsf{E}})$ such that, if $H \geq 4, S \geq 130, A \geq 2$, there exists an absolute constant c_0 such that the expected sample complexity N is lower bounded by*

$$N \geq \frac{c_0 H^3 SA \min\{S, A\}}{\epsilon^2},$$

where $0 < \epsilon \leq (H-2)/1024$*;*

Note that when $S \leq A$, the lower bound scales with $\Omega\big(S^2 A\big)$, matching the $S^2 A$ factor dependence observed in the upper bound (Theorem 11).

G.2. Hard instance construction

Hard Instance Construction Our construction is a modification of the hard instance constructed in the proof of Metelli et al. (2023, Theorem B.3). We construct the hard instance with $2S+1$ states, $A+1$ actions, and $2H+2$ stages for any $H,\ S,\ A > 0$. (This rescaling only affects $S,\ H$ by at most a multiplicative constant and thus does not affect our result). We then define an integer K by

$$K := \min\{S, A\}.$$

Each MDP $\mathcal{M}_{\mathbf{v}}$ is indexed by a vector $\mathbf{w} = \left(w_h^{(i,j,k)}\right)_{h\in[H], i\in[K], j\in[S], k\in[K]} \in \mathbb{R}^{HSKA}$ and is specified as follows:

- State space: $\mathcal{S} = \{s_{\text{start}}, s_{\text{root}}, s_1, \ldots, s_S, \bar{s}_1, \ldots, \bar{s}_S\}$.
- Action space: $\mathcal{A} = \{a_0, a_1, ..., a_A\}$.
- Initial state: s_{start}, that is
$$\mathbb{P}(s_1 = s_{\mathrm{start}}) = 1.$$
- Transitions:
  - At stage 1, s_{start} can only transition to itself or s_i. The transition probabilities are given by
$$\begin{cases} \mathbb{P}_1(s_{\text{start}} \mid s_{\mathrm{start}}, a_0) = 1 & \\ \mathbb{P}_1(s_i \mid s_{\mathrm{start}}, a_i) = 1 & \text{for all } i \in [K], \\ \mathbb{P}_1(s_j \mid s_{\mathrm{start}}, a_k) = \frac{1}{S} & \text{for all } j \in [S],\ k \geq K+1, \end{cases}$$
  - At each stage $h \in \{2, \ldots, H+1\}$, s_{start} can only transition to itself or s_i, s_i can only transition to absorbing state $\bar{s}_j$. The transition probabilities are given by
$$\begin{cases} \mathbb{P}_h(s_{\text{start}} \mid s_{\text{start}}, a_0) = 1, & \\ \mathbb{P}_h(s_i \mid s_{\text{start}}, a_i) = 1 & \text{for all } i \in [K], \\ \mathbb{P}_h(s_j \mid s_{\text{start}}, a_k) = \frac{1}{S} & \text{for all } j \in [S],\ k \geq K+1, \\ \mathbb{P}_h(\bar{s}_j \mid s_i, a_0) = \frac{1}{S} & \text{for all } i \geq K+1,\ j \in [S], \\ \mathbb{P}_h(\bar{s}_j \mid s_i, a_k) = \frac{1+\epsilon' \cdot w_{h-1}^{(i,j,k)}}{S} & \text{for all } i \in [K],\ j \in [S],\ k \in [A], \\ \mathbb{P}_h(\bar{s}_j \mid \bar{s}_j, a_k) = 1 & \text{for all } j \in [S],\ k \geq 0. \end{cases} \tag{113}$$
  - At each stage $h \in \{H+1, \ldots, 2H+2\}$ and s_{start} can only transition to s_i and s_i can only transition to absorbing state $\bar{s}_j$. The transition probabilities are given by
$$\begin{cases} \mathbb{P}_h(s_i \mid s_{\text{start}}, a_0) = \frac{1}{S} & \text{for all } i \in [S], \\ \mathbb{P}_h(s_i \mid s_{\text{start}}, a_i) = 1 & \text{for } i \in [K], \\ \mathbb{P}_h(s_j \mid s_{\text{start}}, a_k) = \frac{1}{S} & \text{for all } j \in [S],\ k \geq K+1, \\ \mathbb{P}_h(\bar{s}_j \mid s_i, a_k) = \frac{1}{S} & \text{for all } i \in [K],\ j \in [S],\ k \geq 0, \\ \mathbb{P}_h(\bar{s}_j \mid \bar{s}_j, a_k) = 1 & \text{for all } i \in [S],\ k \geq 0. \end{cases}$$

- Expert policy: expert policy π^{E} plays action a_0 at every stage $h \in [H]$ and state $s \in \mathcal{S}$. That is

$$\pi^{\mathsf{E}}_h(a_0|s) = 1, \qquad \text{for all } h \in [2H+2],\ s \in \mathcal{S}. \tag{114}$$

In this case, Δ can be 1, which means our lower bound is not derived from a large $\Omega(1/\Delta)$ in our proof. To ensure the definition of $\mathcal{M}_{\mathbf{w}}$ is valid, we enforce the following condition:

$$\sum_{j\in[S]} w_h^{(i,j,k)} = 0,$$

for any $h \in [H]$, $i \in [K]$, $k \in [A]$. We define a vector space $\mathcal{W}$ by

$$\mathcal{W} := \left\{ w = (w_j)_{j\in[S]} \in \{1,-1\}^S : \sum_{j\in[S]} w_j = 0 \right\}.$$

Let $\mathcal{I}$ denote $[H] \times [K] \times [A]$, the Eq.(114) is equivalent to

$$\mathbf{w} \in \mathcal{W}^{\mathcal{I}}.$$

Further, we let $\mathbb{P}^{(\mathbf{w})} = \left\{\mathbb{P}_h^{(\mathbf{w})}\right\}_{h\in[H]}$ to be the transition kernel of MDP\R $\mathcal{M}_{\mathbf{w}}$. In addition, Given $\mathbf{w} \in \mathcal{W}^{\mathcal{I}}, w \in \mathcal{W}$ and index $a \in \mathcal{I}$, we use the notation $\mathbf{w} \overset{a}{\leftarrow} w$ to represent vector obtained by replacing a component of $\mathbf{w}$ with w. For example, let $\mathbf{w} = (w_h^{(i,j,k)})_{h\in[H],i\in[K],j\in[S],k\in[K]}$, $w = (w_j)_{j\in[S]}$, $a = (h_a, i_a, j_a)$ and $\overline{\mathbf{w}} = \mathbf{w} \overset{a}{\leftarrow} w$ and then $\overline{\mathbf{w}}$ can be expressed as follows:

$$\overline{w}_h^{(i,j,k)} = \begin{cases} w_j & (h,i,k) = (h_a, i_a, k_a), \\ w_h^{(i,j,k)} & \text{otherwise.} \end{cases} \tag{115}$$

By Metelli et al. (2023, Lemma E.6), there exists a $\overline{\mathcal{W}} \subseteq \mathcal{W}$ such that

$$\sum_{i\in[n]} (w_i - v_i)^2 \geq \frac{S}{8}, \qquad \forall v, w \in \bar{\mathcal{W}}, \qquad \log\left|\overline{\mathcal{W}}\right| \geq \frac{S}{10}. \tag{116}$$

Notations. To distinguish with different MDP\Rs, we denote $V_h^{\pi}\left(\cdot; r, \mathbb{P}^{(\mathbf{w})}\right)$ be the value function of π in MDP $\mathcal{M}_{\mathbf{w}} \cup r$. Given two rewards r r', we define $d^{\mathsf{all}}(r, r'; \mathbb{P}^{(\mathbf{w})})$ to be the d^{all} metric evaluated in $\mathcal{M}_{\mathbf{w}}$:

$$d^{\mathsf{all}}(r, r'; \mathbb{P}^{(\mathbf{w})}) := \sup_{\pi, h\in[H]} \mathbb{E}_{\mathbb{P}^{(\mathbf{w})},\pi} \left| V_h^{\pi}(s_h; r, \mathbb{P}^{(\mathbf{w})}) - V_h^{\pi}(s_h; r', \mathbb{P}^{(\mathbf{w})}) \right|.$$

Correspondingly, given a parameter set Θ, two reward mappings $\mathscr{R}, \mathscr{R}'$, we can define $D^{\mathsf{all}}_{\Theta}(\mathscr{R}, \mathscr{R}'; \mathbb{P}^{(\mathbf{w})})$ by

$$D^{\mathsf{all}}_{\Theta}(\mathscr{R}, \mathscr{R}'; \mathbb{P}^{(\mathbf{w})}) := \sup_{(V,A)\in\Theta} d^{\mathsf{all}}\Big(\mathscr{R}(V, A), \mathscr{R}'(V, A); \mathbb{P}^{(\mathbf{w})}\Big).$$

In the following, we always assume that $\mathbf{w} \in \bar{\mathcal{W}}$. We then present the following lemma which shows the difference between two MDP\Rs $\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}v}$ and $\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}w}$ for any $\mathbf{w} \in \overline{\mathcal{W}}^{\mathcal{I}}$ and $v \neq w \in \overline{\mathcal{W}}$.

Lemma G.3. *Given any $\mathbf{w} \in \overline{\mathcal{W}}^{\mathcal{I}}$, $w \neq v \in \overline{\mathcal{W}}$, and index $a = (h_a, i_a, k_a) \in \mathcal{I}$, let $\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}w)}$, $\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}v)}$ be the ground truth reward mapping induced by $\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}w}$, $\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}v}$, respectively. Set $\Theta = \overline{\mathcal{V}} \times \overline{\mathcal{A}}$. For any $\epsilon' \in (0, 1/2]$ and any reward mapping $\mathscr{R} : \overline{\mathcal{V}} \times \overline{\mathcal{A}} \to \mathcal{R}^{\mathsf{all}}$, we have*

$$7D^{\mathsf{all}}_{\Theta}\Big(\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}w)}, \mathscr{R}; \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}\Big) + D^{\mathsf{all}}_{\Theta}\Big(\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}v)}, \mathscr{R}; \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}v)}\Big) \geq \frac{H\epsilon'}{16},$$

where ϵ' is specified in Eq.(113).

Proof. **Step 1: Construct the bad parameter $(V^{\mathsf{bad}}, A^{\mathsf{bad}})$.** We construct the bad parameter $(V^{\mathsf{bad}}, A^{\mathsf{bad}}) \in \overline{\mathcal{V}} \times \overline{\mathcal{A}}$ as follows:

- We set $A_h^{\mathsf{bad}}(s, a) = 0$ for all $(h, s, a) \in [2H+2] \times \mathcal{S} \times \mathcal{A}$.
- We set V_h^{bad} by

$$V_h^{\mathsf{bad}}(s) := \begin{cases} \frac{(2H+2-h)\cdot(w_i - v_i)}{2} & \text{if } s = \bar{s}_i, h = h_a + 2, \\ 0 & \text{other.} \end{cases} \tag{117}$$

Directly by the construction of $(V^{\mathsf{bad}}, A^{\mathsf{bad}})$, we obtain that

$$\sum_{i \in [S]} (w_i - v_i) \cdot V_{h_a+2}^{\mathsf{bad}}(\bar{s}_i) = \sum_{i \in [S]} \frac{(2H - h_a)(w_i - v_i)^2}{2} \geq \frac{H\,(w_i - v_i)^2}{2} \geq \frac{HS}{16}, \tag{118}$$

where the last inequality is due to Eq.(116). We then denote $\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}w)}\big(V^{\mathsf{bad}}, A^{\mathsf{bad}}\big)$, $\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}v)}\big(V^{\mathsf{bad}}, A^{\mathsf{bad}}\big)$ as r_w^{bad}, r_v^{bad}, respectively.

Since $A^{\mathsf{bad}} \equiv \mathbf{0}$, any policy $\pi \in \Pi^{\star}_{\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}w} \cup r_w^{\mathsf{bad}}}, \Pi^{\star}_{\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}v} \cup r_v^{\mathsf{bad}}}$. More explicitly, any policy is optimal in $\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}w} \cup r_w^{\mathsf{bad}}$ and $\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}v} \cup r_v^{\mathsf{bad}}$.

Step 2: Construct test policies $\pi^{\mathsf{test},(1)}$, $\pi^{\mathsf{test},(2)}$. Let $r = \mathscr{R}\big(V^{\mathrm{bad}}, A^{\mathrm{bad}}\big)$. Let $\pi^{\mathsf{g}} \in \Pi^{\det}$ be a optimal policy of $\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}w} \cup r$. By Lemma 2, there exist a pair $(V, A) \in \overline{\mathcal{V}} \times \overline{\mathcal{A}}$ such that

$$r_h(s, a) = -A_h(s, a) \cdot \mathbf{1}\left\{a \notin \operatorname{supp}\left(\pi_h^{\mathsf{g}}(\cdot \mid s)\right)\right\} + V_h(s) - \left[\mathbb{P}_h^{(\mathbf{w}\overset{a}{\leftarrow}w)} V_{h+1}\right](s, a), \tag{119}$$

We then construct test policy $\pi^{\mathsf{test},(1)}$ by

$$\begin{cases} \pi_h^{\mathsf{test},(1)}(a_0 \mid s_{\mathsf{start}}) = 1 & h \leq h_a - 1 \\ \pi_h^{\mathsf{test},(1)}(a_{i_a} \mid s_{\mathsf{start}}) = 1 & h = h_a \\ \pi_h^{\mathsf{test},(1)}(a_{k_a} \mid s_{i_a}) = 1 & h = h_a + 1 \\ \pi_h^{\mathsf{test},(1)} \quad = \pi_h^{\mathsf{g}} & h \geq h_a + 2 \end{cases}$$

which implies that at stage $h \leq h_a - 1$, $\pi^{\mathsf{test},(1)}$ always plays a_0, at stage h_a, $\pi^{\mathsf{test},(1)}$ plays a_{i_a}, then transition to s_{i_a}, at stage $h_a + 1$, $\pi^{\mathsf{test},(1)}$ plays a_{k_a}, then at stage $h \geq h_a + 2$, $\pi^{\mathsf{test},(1)}$ is equal to the greedy policy π^{g}. By construction, we can conclude that

$$d_{h_a+1}^{\pi^{\mathsf{test},(1)}}(s_{i_a}; \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}) = 1, \qquad V_{h_a+2}^{\pi^{\mathsf{test},(1)}}(\cdot|r, \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}) = V_{h_a+2}(\cdot)., \tag{120}$$

the second equality is due to $\pi_h^{\mathsf{test},(1)} = \pi_h^{\mathsf{g}}$ for any $h \geq h_a + 2$.

Further, we have

$$\begin{aligned} V_{h_a+1}^{\pi^{\mathsf{test},(1)}}(s_{i_a}; r, \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}) &= r_{h_a+1}(s_{i_a}, a_{k_a}) + \left[\mathbb{P}_{h_a+1}^{(\mathbf{w}\overset{a}{\leftarrow}w)} V_{h_a+2}^{\pi^{\mathsf{test},(1)}}(\cdot|r, \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)})\right](s_{i_a}, a_{k_a}) \\ &= -A_{h_a+1}(s_{i_a}, a_{k_a}) \cdot \mathbf{1}\left\{a_{k_a} \notin \operatorname{supp}\left(\pi_{h_a+1}^{\mathsf{g}}(\cdot \mid s_{i_a})\right)\right\} \\ &+ V_{h_a+1}(s) - \left[\mathbb{P}_{h_a+1}^{(\mathbf{w}\overset{a}{\leftarrow}w)} V_{h_a+2}\right](s_{i_a}, a_{k_a}) + \left[\mathbb{P}_{h_a+1}^{(\mathbf{w}\overset{a}{\leftarrow}w)} V_{h_a+2}\right] \\ &= V_{h_a+1}(s_{i_a}) - \mathsf{gap}, \end{aligned} \tag{121}$$

where the first line is by the Bellman equation, the second line is due to Eq.(119) and Eq.(120). Here gap is the advantage function at $(h_a + 1, s_{i_a}, a_{k_a})$, i.e, $\mathsf{gap} := A_{h_a+1}(s_{i_a}, a_{k_a}) \cdot \mathbf{1}\left\{a_{k_a} \in \operatorname{supp}\left(\pi_{h_a+1}^{\mathsf{g}}(\cdot \mid s_{i_a})\right)\right\}$. Then by definition of

$D^{\mathsf{all}}_{\Theta}(\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}w)},\mathscr{R};\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)})$, we can obtain that

$$\begin{aligned}
&D^{\mathsf{all}}_{\Theta}\Big(\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}w)},\mathscr{R};\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}\Big)\\
&\geq d^{\mathsf{all}}\Big(\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}w)}\big(V^{\mathsf{bad}},A^{\mathsf{bad}}\big),\mathscr{R}\big(V^{\mathsf{bad}},A^{\mathsf{bad}}\big);\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}\Big)\\
&=d^{\mathsf{all}}\Big(r^{\mathsf{bad}}_w,r;\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}\Big)\\
&\geq \mathbb{E}_{\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)},\pi^{\mathsf{test},(1)}}\Big|V^{\pi^{\mathsf{test},(1)}}_{h_a+1}(s;r^{\mathsf{bad}}_w,\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)})-V^{\pi^{\mathsf{test},(1)}}_{h_a+1}(s;r,\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)})\Big|\\
&=\Big|V^{\pi^{\mathsf{test},(1)}}_{h_a+1}(s_{i_a};r^{\mathsf{bad}}_w,\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)})-V^{\pi^{\mathsf{test},(1)}}_{h_a+1}(s_{i_a};r,\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)})\Big|\\
&=\big|V^{\mathsf{bad}}_{h_a+1}(s_{i_a})-V_{h_a+1}(s_{i_a})+\mathsf{gap}\big|, \qquad (122)
\end{aligned}$$

where the second last line is due to Eq.(120) and the last line is by Eq.(121) and $\pi^{\mathsf{test},(1)}\in\Pi^{\star}_{\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}w}\cup r^{\mathsf{bad}}_w}\Rightarrow V^{\pi^{\mathsf{test},(1)}}_{h_a+1}(s_{i_a};r^{\mathsf{bad}}_w,\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)})=V^{\mathsf{bad}}_{h_a+1}(s_{i_a})$.

Next, we construct another test policy $\pi^{\mathsf{test},(2)}$ as follows:

$$\begin{cases}
\pi^{\mathsf{test},(2)}_h(a_0\mid s_{\text{start}})=1 & h\leq h_a-1\\
\pi^{\mathsf{test},(2)}_h(a_{i_a}\mid s_{\text{start}})=1 & h=h_a\\
\pi^{\mathsf{test},(2)}_h=\pi^{\mathsf{g}}_h & h\geq h_a+1.
\end{cases}$$

The difference between $\pi^{\mathsf{test},(2)}$ and $\pi^{\mathsf{test},(1)}$ is that at stage h_a $\pi^{\mathsf{test},(2)}$ play the $\pi^{\mathsf{g}}_{h_a+1}(s_{i_a})$ instead of a_{k_a}. Similar to Eq.(120), we have

$$d^{\pi^{\mathsf{test},(2)}}_{h_a+1}(s_{i_a};\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)})=1,\qquad V^{\pi^{\mathsf{test},(2)}}_{h_a+1}(s_{i_a};r,\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)})=V_{h_a+1}(s_{i_a}) \qquad (123)$$

where the seconed equality is valid since $\pi^{\mathsf{test},(2)}_h=\pi^{\mathsf{g}}_h$ for any $h\geq h_a+1$.

Similar to Eq.(122), we have

$$\begin{aligned}
D^{\mathsf{all}}_{\Theta}\Big(\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}w)},\mathscr{R};\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}\Big)&\geq d^{\mathsf{all}}\Big(r^{\mathsf{bad}}_w,r;\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}\Big)\\
&\geq \mathbb{E}_{\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)},\pi^{\mathsf{test},(2)}}\Big|V^{\pi^{\mathsf{test},(2)}}_{h_a+1}(s;r^{\mathsf{bad}}_w,\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)})-V^{\pi^{\mathsf{test},(2)}}_{h_a+1}(s;r,\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)})\Big|\\
&=\Big|V^{\pi^{\mathsf{test},(2)}}_{h_a+1}(s_{i_a};r^{\mathsf{bad}}_w,\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)})-V^{\pi^{\mathsf{test},(2)}}_{h_a+1}(s_{i_a};r,\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)})\Big|\\
&=\big|V^{\mathsf{bad}}_{h_a+1}(s_{i_a})-V_{h_a+1}(s_{i_a})\big|, \qquad (124)
\end{aligned}$$

where the second last is due to Eq.(123), the last line follows from $\pi^{\mathsf{test},(2)}\in\Pi^{\star}_{\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}w}\cup r^{\mathsf{bad}}_w}$: $V^{\pi^{\mathsf{test},(2)}}_{h_a+1}(s_{i_a};r^{\mathsf{bad}}_w,\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)})=V^{\mathsf{bad}}_{h_a+1}(s_{i_a})$. Combing Eq.(122) and Eq.(124), we have

$$\begin{aligned}
2D^{\mathsf{all}}_{\Theta}\Big(\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}w)},\mathscr{R};\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}\Big)&\geq\big|V^{\mathsf{bad}}_{h_a+1}(s_{i_a})-V_{h_a+1}(s_{i_a})\big|+\big|V^{\mathsf{bad}}_{h_a}(s_{i_a})-V_{h_a+1}(s_{i_a})+\mathsf{gap}\big|\\
&\geq\mathsf{gap}, \qquad (125)
\end{aligned}$$

where the second line comes from the triangle inequality.

Step 3: lower bound $D^{\mathsf{all}}_{\Theta}\Big(\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}v)},\mathscr{R};\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}v)}\Big)$. We still use the test policy $\pi^{\mathsf{test},(1)}$ in $\mathcal{M}_{(\mathbf{w}\overset{a}{\leftarrow}v)}$. Since $\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}v)}_h=\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}_h$ for any $h\geq h_a+2$, we have

$$V^{\pi^{\mathsf{test},(1)}}_{h_a+2}(\bar{s}_i|r,\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}v)})=V^{\pi^{\mathsf{test},(1)}}_{h_a+2}(\bar{s}_i|r,\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)})=V_{h_a+2}(\bar{s}_i),\qquad\text{for all } i\in[S], \qquad (126)$$

where the second equality comes from Eq.(120).

By the definition of $D_\Theta^{\mathsf{all}}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}, \mathscr{R}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right)$, we have

$$
\begin{aligned}
& D_\Theta^{\mathsf{all}}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}, \mathscr{R}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right) \\
& \geq d^{\mathsf{all}}\left(r_v^{\mathsf{bad}}, r; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right) \\
& \geq \mathbb{E}_{\mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}, \pi^{\mathsf{test},(1)}} \left| V_{h_a+1}^{\pi^{\mathsf{test},(1)}}\left(s; r_v^{\mathsf{bad}}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right) - V_{h_a+1}^{\pi^{\mathsf{test},(1)}}\left(s; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right)\right| \\
& = \left| V_{h_a+1}^{\pi^{\mathsf{test},(1)}}\left(s_{i_a}; r_v^{\mathsf{bad}}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right) - V_{h_a+1}^{\pi^{\mathsf{test},(1)}}\left(s_{i_a}; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right)\right| \qquad \text{(by construction of policy } \pi^{\mathsf{test},(1)}.\text{)} \\
& = \left| V_{h_a+1}^{\mathsf{bad}}(s_{i_a}) - V_{h_a+1}^{\pi^{\mathsf{test},(1)}}\left(s_{i_a}; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right)\right| \\
& \overset{\text{(i)}}{=} \left| V_{h_a+1}^{\mathsf{bad}}(s_{i_a}) - r_{h_a+1}(s_{i_a}, a_{k_a}) - \left[\mathbb{P}_{h_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)} V_{h_a+2}\right](s_{i_a}, a_{k_a})\right| \\
& = \left| V_{h_a+1}^{\mathsf{bad}}(s_{i_a}) - r_{h_a+1}(s_{i_a}, a_{k_a}) - \left[\mathbb{P}_{h_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)} V_{h_a+2}\right](s_{i_a}, a_{k_a}) - \left[\left(\mathbb{P}_{h_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)} - \mathbb{P}_{h_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}\right) V_{h_a+2}\right](s_{i_a}, a_{k_a})\right| \\
& \geq \left| \left[\left(\mathbb{P}_{h_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)} - \mathbb{P}_{h_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}\right) V_{h_a+2}\right](s_{i_a}, a_{k_a})\right| - \left| V_{h_a+1}^{\mathsf{bad}}(s_{i_a}) - r_{h_a+1}(s_{i_a}, a_{k_a}) - \left[\mathbb{P}_{h_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)} V_{h_a+2}\right](s_{i_a}, a_{k_a})\right| \\
& \qquad \text{(by triangle inequality)} \\
& \overset{\text{(ii)}}{=} \left| \left[\left(\mathbb{P}_{h_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)} - \mathbb{P}_{h_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}\right) V_{h_a+2}\right](s_{i_a}, a_{k_a})\right| - \left| V_{h_a+1}^{\mathsf{bad}}(s_{i_a}) - V_{h_a+1}(s_{i_a}) + \mathsf{gap}\right| \\
& \geq \left| \left[\left(\mathbb{P}_{h_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)} - \mathbb{P}_{h_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}\right) V_{h_a+2}\right](s_{i_a}, a_{k_a})\right| - \left| V_{h_a+1}^{\mathsf{bad}}(s_{i_a}) - V_{h_a+1}(s_{i_a})\right| - \mathsf{gap} \\
& \overset{\text{(iii)}}{\geq} \left| \left[\left(\mathbb{P}_{s_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)} - \mathbb{P}_{h_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}\right) V_{h_a+2}\right](s_{i_a}, a_{k_a})\right| - 3D_\Theta^{\mathsf{all}}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}, \mathscr{R}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}\right), \qquad (127)
\end{aligned}
$$

where (i) is by the Bellman equation, (ii) is valid since

$$
\begin{aligned}
& r_{h_a+1}(s_{i_a}, a_{k_a}) + \left[\mathbb{P}_{h_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)} V_{h_a+2}\right](s_{i_a}, a_{k_a}) \\
& \quad = -A_{h_a+1}(s_{i_a}, a_{k_a}) \cdot \mathbf{1}\left\{a_{k_a} \in \operatorname{supp}\left(\pi_{h_a+1}^{g}(\cdot|s_{i_a})\right)\right\} + V_{h_a+1}(s_{i_a}) \\
& \quad - \left[\mathbb{P}_{h_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)} V_{h_a+2}\right](s_{i_a}, a_{k_a}) + \left[\mathbb{P}_{h_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)} V_{h_a+2}\right](s_{i_a}, a_{k_a}) \qquad \text{(by Eq.(119))} \\
& \quad = -\mathsf{gap} + V_{h_a+1}(s_{i_a})
\end{aligned}
$$

and (iii) is due to Eq.(124) and Eq.(125). We next analyse $\left| \left[\left(\mathbb{P}_{h_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)} - \mathbb{P}_{h_a+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}\right) V_{h_a+2}\right](s_{i_a}, a_{k_a})\right|$. We move back to $\pi^{\mathsf{test},(1)}$. By the construction of $\pi^{\mathsf{test},(1)}$ and the transition probabilities of $\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}w}$, we have

$$
d_{h_a+2}^{\pi^{\mathsf{test},(1)}}\left(\bar{s}_i; \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}\right) = \frac{1+\epsilon' w_i}{S}, \qquad V_{h_a+2}^{\pi^{\mathsf{test},(1)}}\left(\bar{s}_i; r, \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}\right) = V_{h_a+2}(\bar{s}_i), \qquad \forall i \in [S]. \qquad (128)
$$

By definition of $D_{\Theta}^{\mathsf{all}}(\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}w)}, \mathscr{R}; \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)})$, we have

$$
\begin{aligned}
&D_{\Theta}^{\mathsf{all}}\left(\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}w)}, \mathscr{R}; \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}\right) \\
&\geq \mathbb{E}_{\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}, \pi^{\mathsf{test},(1)}} \left| V_{h_a+2}^{\pi^{\mathsf{test},(1)}}(s; r_w^{\mathsf{bad}}, \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}) - V_{h_a+2}^{\pi^{\mathsf{test},(1)}}(s; r, \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}) \right| \\
&\geq \sum_{i\in[S]} d_{h_a+2}^{\pi^{\mathsf{test},(1)}}(\bar{s}_i; \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}) \cdot \left| V_{h_a+2}^{\pi^{\mathsf{test},(1)}}(\bar{s}_i; r_w^{\mathsf{bad}}, \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}) - V_{h_a+2}^{\pi^{\mathsf{test},(1)}}(\bar{s}_i; r,\ \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}) \right| \\
&= \sum_{i\in[S]} \frac{1+\epsilon' w_i}{S} \cdot \left| V_{h_a+2}^{\mathsf{bad}}(\bar{s}_i) - V_{h_a+2}(\bar{s}_i) \right| \\
&\geq \sum_{i\in[S]} \frac{1}{2S} \cdot \left| V_{h_a+2}^{\mathsf{bad}}(\bar{s}_i) - V_{h_a+2}(\bar{s}_i) \right|,
\end{aligned} \tag{129}
$$

where the last second is by Eq.(128) and the last line comes from $\epsilon' \in (0, 1/2]$. Applying Eq.(129), we obtain that

$$
\begin{aligned}
&\left| \left[\left(\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}v)} - \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)} \right) V_{h_a+2} \right] (s_{i_a}, a_{k_a}) \right| \\
&= \left| \frac{\epsilon'}{S} \cdot \sum_{i\in[S]} V_{h_a+2}(\bar{s}_i) \cdot (w_i - v_i) \right| \\
&\geq \left| \frac{\epsilon'}{S} \cdot \sum_{i\in[S]} V_{h_a+2}^{\mathsf{bad}}(\bar{s}_i) \cdot (w_i - v_i) \right| - \frac{\epsilon'}{S} \cdot \sum_{i\in[S]} \left| V_{h_a+2}^{\mathsf{bad}}(\bar{s}_i) - V_{h_a+2}(\bar{s}_i) \right| \cdot |(w_i - v_i)| && \text{(by triangle inequality)} \\
&\geq \left| \frac{\epsilon'}{S} \cdot \sum_{i\in[S]} V_{h_a+2}^{\mathsf{bad}}(\bar{s}_i) \cdot (w_i - v_i) \right| - \frac{2\epsilon'}{S} \cdot \sum_{i\in[S]} \left| V_{h_a+2}^{\mathsf{bad}}(\bar{s}_i) - V_{h_a+2}(\bar{s}_i) \right| \\
&\geq \frac{H\epsilon'}{16} - 2D_{\Theta}^{\mathsf{all}}\left(\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}w)}, \mathscr{R}; \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}\right),
\end{aligned} \tag{130}
$$

where the second line is by the triangle inequality and the last line comes from Eq.(118)and Eq.(129). Combining Eq.(127) and Eq.(130), we complete the proof.

□

G.3. Proof of Theorem G.2

Proof of Theorem G.2. Our method is similar to the one used for the proof of Metelli et al. (2023, Theorem B.3). For any $\epsilon \in (0, 1/2]$, $\delta \in (0, 1)$, we consider an online algorithm $\mathfrak{A}$ such that for any IRL problem $(\mathcal{M}, \pi^{\mathsf{E}})$, we have

$$
\mathop{\mathbb{P}}_{(\mathcal{M},\pi^{\mathsf{E}}),\mathfrak{A}} \left(D_{\Theta}^{\mathsf{all}}\left(\mathscr{R}^{\star}, \widehat{\mathscr{R}}\right) \leq \epsilon \right) \geq 1 - \delta, \tag{131}
$$

where $\mathop{\mathbb{P}}_{(\mathcal{M},\pi^{\mathsf{E}}),\mathfrak{A}}$ denotes the probability measure induced by executing the algorithm $\mathfrak{A}$ in the IRL problem $(\mathcal{M}, \pi^{\mathsf{E}})$, $\mathscr{R}^{\star}$ is the ground truth reward mapping and $\widehat{\mathscr{R}}$ is the estimated reward mapping outputted by executing $\mathfrak{A}$ in $(\mathcal{M}, \pi^{\mathsf{E}})$. We define the the identification function for any $(a, \mathbf{w}) \in \mathcal{I} \times \overline{\mathcal{W}}^{\mathcal{I}}$ by

$$
\mathbf{\Phi}_{a,\mathbf{w}} := \mathop{\arg\min}_{v\in\overline{\mathcal{W}}} D_{\Theta}^{\mathsf{all}}\left(\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}v)}, \widehat{\mathscr{R}}; \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}v)}\right),
$$

where $\mathscr{R}^{(\mathbf{w})}$ is the ground truth reward mapping induced by $(\mathcal{M}_{\mathbf{w}}, \pi^{\mathsf{E}})$. Let $v^{\star} = \mathbf{\Phi}_{a,\mathbf{w}}$. For any $v \neq v^{\star} \in \overline{\mathcal{W}}$, by definition of $v^{\star}$, we have

$$
D_{\Theta}^{\mathsf{all}}\left(\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}v^{\star})}, \widehat{\mathscr{R}}; \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}v^{\star})}\right) \leq D_{\Theta}^{\mathsf{all}}\left(\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}v)}, \widehat{\mathscr{R}}; \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}v)}\right).
$$

By applying Lemma G.3, we obtain that

$$\frac{H\epsilon'}{16} \leq D^{\mathsf{all}}_{\Theta}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}, \widehat{\mathscr{R}}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right) + 7D^{\mathsf{all}}_{\Theta}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}v^{\star}\right)}, \widehat{\mathscr{R}}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v^{\star}\right)}\right) \leq 8D^{\mathsf{all}}_{\Theta}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}, \widehat{\mathscr{R}}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right).$$

Next, we set $\epsilon' = \frac{256\epsilon}{H}$ which implies that

$$\frac{H\epsilon'}{16} \geq 16\epsilon. \tag{132}$$

Here, to employ Lemma G.3, we need $\epsilon' \in (0, 1/2]$ which is equivalent to $0 < \epsilon \leq H/512$. Then, it holds that

$$D^{\mathsf{all}}_{\Theta}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}, \widehat{\mathscr{R}}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right) \geq 2\epsilon > \epsilon,$$

which implies that

$$\{v \neq \mathbf{\Phi}_{a,\mathbf{w}}\} \subseteq \left\{D^{\mathsf{all}}_{\Theta}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}, \widehat{\mathscr{R}}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right) > \epsilon\right\}. \tag{133}$$

By Eq.(133), we have the following lower bound for the probability

$$\begin{aligned}
\delta &\geq \sup_{v\in\overline{\mathcal{W}}} \underset{(\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}v},\pi^{\mathsf{E}}),\mathfrak{A}}{\mathbb{P}}\left(D^{\mathsf{all}}_{\Theta}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}, \widehat{\mathscr{R}}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right) > \epsilon\right) \\
&\geq \sup_{v\in\overline{\mathcal{W}}} \underset{(\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}v},\pi^{\mathsf{E}}),\mathfrak{A}}{\mathbb{P}}(v \neq \mathbf{\Phi}_{a,\mathbf{w}}) \\
&\geq \frac{1}{|\overline{\mathcal{W}}|}\sum_{v\in\overline{\mathcal{W}}} \underset{(\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}v},\pi^{\mathsf{E}}),\mathfrak{A}}{\mathbb{P}}(v \neq \mathbf{\Phi}_{a,\mathbf{w}}),
\end{aligned} \tag{134}$$

By applying Theorem B.3 with $\mathbb{P}_0 = \underset{(\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}0},\pi^{\mathsf{E}}),\mathfrak{A}}{\mathbb{P}}$, $\mathbb{P}_w = \underset{(\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}w},\pi^{\mathsf{E}}),\mathfrak{A}}{\mathbb{P}}$, we have

$$\frac{1}{|\overline{\mathcal{W}}|}\sum_{(\mathcal{M}_{\mathbf{w}}\overset{a}{\leftarrow}v,\pi^{\mathsf{E}}),\mathfrak{A}} (v \neq \mathbf{\Phi}_{a,\mathbf{w}}) \geq 1 - \frac{1}{\log|\overline{\mathcal{W}}|}\left(\frac{1}{|\overline{\mathcal{W}}|}\sum_{v\in\overline{\mathcal{W}}} D_{\mathrm{KL}}(\underset{(\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}v},\pi^{\mathsf{E}}),\mathfrak{A}}{\mathbb{P}}, \underset{(\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}0},\pi^{\mathsf{E}}),\mathfrak{A}}{\mathbb{P}}) - \log 2\right). \tag{135}$$

Our next step is to bound the KL divergence. Using the same scheme in the proof Metelli et al. (2021, Theorem B.3), we can compute the KL-divergence as follows:

$$\begin{aligned}
&D_{\mathrm{KL}}(\underset{(\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}v},\pi^{\mathsf{E}}),\mathfrak{A}}{\mathbb{P}}, \underset{(\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}0},\pi^{\mathsf{E}}),\mathfrak{A}}{\mathbb{P}}) \\
&= \mathbb{E}_{(\mathcal{M}_{\mathbf{w}}\overset{a}{\leftarrow}v,\pi^{\mathsf{E}}),\mathfrak{A}}\left[\sum_{t=1}^{N} D_{\mathrm{KL}}\left(\mathbb{P}_{h_t}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}(\cdot\,|\,s_t,a_t), \mathbb{P}_{h_t}^{\left(\mathbf{w}\overset{a}{\leftarrow}0\right)}(\cdot\,|\,s_t,a_t)\right)\right] \\
&\leq \mathbb{E}_{(\mathcal{M}_{\mathbf{w}}\overset{a}{\leftarrow}v,\pi^{\mathsf{E}}),\mathfrak{A}}[N_{h_a}(s_{i_a},a_{k_a})] D_{\mathrm{KL}}\left(\mathbb{P}_{h_a}^{\left(\left(\mathbf{w}\overset{a}{\leftarrow}v\right)\right)}(\cdot\,|\,s_{i_a},a_{k_a}), \mathbb{P}_{h_a}^{\left(\mathbf{w}\overset{a}{\leftarrow}0\right)}(\cdot\,|\,s_{i_a},a_{k_a})\right) \\
&\leq 2(\epsilon')^2\mathbb{E}_{(\mathcal{M}_{\mathbf{w}}\overset{a}{\leftarrow}v,\pi^{\mathsf{E}}),\mathfrak{A}}[N_{h_a}(s_{i_a},a_{k_a})],
\end{aligned} \tag{136}$$

where $N_h(s,a) := \sum_{t=1}^{N}\mathbf{1}\{(h_t,s_t,a_t) = (h,s,a)\}$ for any given $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$ and the last inequality comes from Metelli et al. (2021, Lemma E.4). Combining Eq.(134) and Eq.(135), we have

$$\delta \geq 1 - \frac{1}{\log(|\overline{\mathcal{W}}|)}\left(\frac{1}{|\overline{\mathcal{W}}|}\sum_{v\in\overline{\mathcal{W}}} 2(\epsilon')^2\mathbb{E}_{(\mathcal{M}_{\mathbf{w}}\overset{a}{\leftarrow}v,\pi^{\mathsf{E}}),\mathfrak{A}}[N_{h_a}(s_{i_a},a_{k_a})] - \log 2\right)$$

for any $\mathbf{w}$. It also holds for any $a \in \mathcal{I}$ that

$$\frac{1}{|\overline{\mathcal{W}}|}\sum_{v\in\overline{\mathcal{W}}} \mathbb{E}_{(\mathcal{M}_{\mathbf{w}}\overset{a}{\leftarrow}v,\pi^{\mathsf{E}}),\mathfrak{A}}[N_{h_a}(s_{i_a},a_{k_a})] \geq \frac{(1-\delta)\log|\overline{\mathcal{W}}| - \log 2}{2(\epsilon')^2}. \tag{137}$$

By summing Eq.(137) over all $\mathbf{w}$, we obtain that

$$\begin{aligned}
&\sum_{a\in\mathcal{I}}\frac{1}{|\overline{\mathcal{W}}^{\mathcal{I}}|}\sum_{\mathbf{w}\in\overline{\mathcal{W}}^{\mathcal{I}}}\frac{1}{|\overline{\mathcal{W}}|}\sum_{v\in\overline{\mathcal{W}}}\mathbb{E}_{(\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}v},\pi^{\mathsf{E}}),\mathfrak{A}}[N_{h_a}(s_{i_a},a_{k_a})]\\
&=\frac{1}{|\overline{\mathcal{W}}^{\mathcal{I}}|}\sum_{\mathbf{w}\in\overline{\mathcal{W}}^{\mathcal{I}}}\sum_{a\in\mathcal{I}}\mathbb{E}_{(\mathcal{M}_{\mathbf{w}},\pi^{\mathsf{E}}),\mathfrak{A}}[N_{h_a}(s_{i_a},a_{k_a})]\\
&\geq HKA\frac{(1-\delta)\log|\overline{\mathcal{W}}|-\log 2}{2(\epsilon')^2}.
\end{aligned}\tag{138}$$

Hence, there exists a $\mathbf{w}^{\mathrm{bad}}\in\overline{\mathcal{W}}^{\mathcal{I}}$ such that

$$\begin{aligned}
\mathbb{E}_{(\mathcal{M}_{\mathbf{w}^{\mathrm{bad}}},\pi^{\mathsf{E}}),\mathfrak{A}}[N]&\geq\sum_{a\in\mathcal{I}}\mathbb{E}_{(\mathcal{M}_{\mathbf{w}^{\mathrm{bad}}},\pi^{\mathsf{E}}),\mathfrak{A}}\left[N^t_{h_a}(s_{i_a},a_{k_a})\right]\geq HKA\frac{(1-\delta)\log|\overline{\mathcal{W}}|-\log 2}{2(\epsilon')^2}\\
&=H^3KA\frac{(1-\delta)\log|\overline{\mathcal{W}}|-\log 2}{131072\epsilon^2},
\end{aligned}\tag{139}$$

where the last line is by $\epsilon'=\frac{\epsilon}{256H}$. By taking $\delta=1/3$, we obtain that

$$\begin{aligned}
\mathbb{E}_{(\mathcal{M}_{\mathbf{w}^{\mathrm{bad}}},\pi^{\mathsf{E}}),\mathfrak{A}}[N]&\geq H^3KA\frac{(1-\delta)\log|\overline{\mathcal{W}}|-\log 2}{131072\epsilon^2}=H^3KA\frac{2\log|\overline{\mathcal{W}}|-3\log 2}{393216\epsilon^2}\\
&=\Omega\left(\frac{H^3SKA}{\epsilon^2}\right)=\Omega\left(\frac{H^3SA\min\{S,A\}}{\epsilon^2}\right),
\end{aligned}\tag{140}$$

where the last line follows from Eq.(132) and $\log|\overline{\mathcal{W}}|\geq\frac{S}{10}$. □

H. Lower bound in the offline setting

H.1. Lower bound of offline IRL problems

We direct our attention towards the lower bound analysis of the offline IRL problems, particularly in scenarios where $\Theta=\overline{\mathcal{V}}\times\overline{\mathcal{A}}$. In this case $\log\mathcal{N}(\Theta;\epsilon/H)$ is upper-bounded by $\widetilde{\mathcal{O}}(S)$, and the corresponding upper bound of the sample complexity becomes $\widetilde{\mathcal{O}}\left(\frac{C^\star H^4S^2A}{\epsilon^2}\right)$.

Following Metelli et al. (2023) we define the (ϵ,δ)-PAC algorithm for offline IRL problems for all $\epsilon,\delta\in(0,1)$.

Definition H.1 ((ϵ,δ)-PAC algorithm for offline IRL problems). *We say an offline IRL algorithm $\mathfrak{A}$ is an (ϵ,δ)-PAC algorithm for offline IRL problems if for any offline IRL problem $(\mathcal{M},\pi^{\mathsf{E}},\pi^{\mathsf{b}},\pi^{\mathsf{eval}})$ and any parameter set Θ, with probability $1-\delta$, $\mathfrak{A}$ outputs a reward mapping $\widehat{\mathscr{R}}$ such that*

$$D_{\Theta}^{\pi^{\mathsf{eval}}}(\widehat{\mathscr{R}},\mathscr{R}^\star)\leq\epsilon.$$

Theorem H.2 (Lower bound for offline IRL problems). *Fix $\Theta=\overline{\mathcal{V}}\times\overline{\mathcal{A}}$ and let $\mathfrak{A}$ be an (ϵ,δ)-PAC algorithm for offline IRL problems, where $\delta\leq1/3$. Then, there exists an offline IRL problem $(\mathcal{M},\pi^{\mathsf{E}},\pi^{\mathsf{b}},\pi^{\mathsf{eval}})$ such that, if $H,S\geq4$, $A\geq2$, $C^\star\geq2$, there exists an absolute constant c_0 such that the sample complexity N is lower bounded by*

$$N\geq\frac{c_0H^2SC^\star\min\{S,A\}}{\epsilon^2}.$$

where $0<\epsilon\leq(H-2)/1024$.

The hard instance construction and the proof of Theorem H.2 can be found to Section H.2 and Section H.3, respectively. Our proof involves a modification of the challenging instance constructed in Metelli et al. (2023). Specifically, when $S\leq A$, the lower bound scales with $\Omega(C^\star S^2)$, matching the $C^\star S^2$ factor dependence observed in the upper bound (Theorem 8).

H.2. Hard instance construction

We consider the MDP\R $\mathcal{M}_{\mathbf{w}}$ indexed by vector $\mathbf{w} \in \overline{\mathcal{W}}^{\mathcal{I}}$, defined in Section G. We assume $C^\star \geq 2$. Fix $i^\star \in [K]$, we construct the behavior policy π^{b} as follows:

$$\begin{cases} \pi^{\mathsf{b}}_h(a_0|s_{\text{start}}) = 1 \qquad \text{for all } i \in [K] \text{ and } h \in [H-1], \\ \pi^{\mathsf{b}}_H(a_i|s_{\text{start}}) = \frac{1}{K} \qquad \text{for all } i \in [K], \\ \pi^{\mathsf{b}}_{H+1}(a_0|s_i) = 1 \qquad \text{for all } i \neq i^\star, \\ \pi^{\mathsf{b}}_{H+1}(a_0|s_{i^\star}) = 1 - \frac{1}{C^\star}, \qquad \pi^{\mathsf{b}}_{H+1}(a_1|s_{i^\star}) = \frac{1}{C^\star}, \\ \pi^{\mathsf{b}}_h(a_0|\bar{s}_i) = 1 \qquad \text{for all } i \in [S] \text{ and } h \geq H+2. \end{cases} \tag{141}$$

And evaluation policy π^{eval} is defined by

$$\begin{cases} \pi^{\mathsf{eval}}_h(a_0|s_{\text{start}}) = 1 \qquad \text{for all } h \in [H-1], \\ \pi^{\mathsf{eval}}_H(a_{i^\star}|s_{\text{start}}) = 1, \\ \pi^{\mathsf{eval}}_{H+1}(a_0|s_i) = 1 \qquad \text{for all } i \neq i^\star, \\ \pi^{\mathsf{eval}}_{H+1}(a_1|s_{i^\star}) = 1, \\ \pi^{\mathsf{eval}}_h(a_0|\bar{s}_i) = 1 \qquad \text{for all } i \in [S] \text{ and } h \geq H+2. \end{cases} \tag{142}$$

For all $\mathbf{w} \in \overline{\mathcal{W}}^{\mathcal{I}}$, we can show that π^{eval} has $C^\star$-concentrability in $\mathcal{M}_{\mathbf{w}}$.

Lemma H.3. *Suppose that $\epsilon' \in (0, 1/2]$. For any $\mathbf{w} \in \overline{\mathcal{W}}^{\mathcal{I}}$, it holds that*

$$\sum_{(h,s,a)\in[2H+2]\times\mathcal{S}\times\mathcal{A}} \frac{d^{\pi^{\mathsf{eval}}}(s,a)}{d^{\pi^{\mathsf{b}}}(s,a)} \leq 3C^\star(H+2)S.$$

Proof. By the construction of behavior policy π^{b}, we have

$$\operatorname{supp}\left(d^{\pi^{\mathsf{eval}}}_h(\cdot,\cdot)\right) \subseteq \{(s_{\text{start}}, a_0), (s_{\text{start}}, a_{k^\star}), (s_{i^\star}, a_1), (\bar{s}_1, a_0), \dots, (\bar{s}_S, a_0)\}.$$

Since $\pi^{\mathsf{b}}_h = \pi^{\mathsf{eval}}_h$ for all $h \in [H-1]$, then

$$d^{\pi^{\mathsf{b}}}_h(s_{\text{start}}, a_0) = d^{\pi^{\mathsf{eval}}}_h(s_{\text{start}}, a_0) = 1 \tag{143}$$

for all $h \in [H-1]$.

At stage $h = H$, we have

$$d^{\pi^{\mathsf{b}}}_H(s_{\text{start}}, a_{i^\star}) = \frac{1}{K}, \qquad d^{\pi^{\mathsf{eval}}}_H(s_{\text{start}}, a_{i^\star}) = 1. \tag{144}$$

At stage $h = H+1$, we have

$$d^{\pi^{\mathsf{b}}}_{H+1}(s_{i^\star}, a_1) = \frac{1}{C^\star K}, \qquad d^{\pi^{\mathsf{eval}}}_{H+1}(s_{i^\star}, a_1) = 1. \tag{145}$$

At stage $h \in \{H+2, \dots, 2H+2\}$, by direct computation, we obtain that

$$d^{\pi^{\mathsf{b}}}_h(\bar{s}_j, a_0) = \frac{C^\star K - 1}{C^\star SK} + \frac{1 + \epsilon' w_H^{(i^\star,j,1)}}{C^\star SK}, \qquad d^{\pi^{\mathsf{eval}}}_h(\bar{s}_j, a_1) = \frac{1 + \epsilon' w_H^{(i^\star,j,1)}}{S}, \tag{146}$$

for all $j \in [S]$. Since $0 < \epsilon \leq 1/2$ and $C^\star \geq 1$, we have

$$\begin{aligned} d^{\pi^{\mathsf{b}}}_h(\bar{s}_j, a_0) &= \frac{C^\star K - 1}{C^\star SK} + \frac{1 + \epsilon' w_H^{(i^\star,j,1)}}{C^\star SK} \\ &\geq \frac{C^\star K - 1}{C^\star SK} + \frac{1}{2C^\star SK} = \frac{1}{S}(1 - \frac{1}{2C^\star K}) \geq \frac{1}{2S} \end{aligned} \tag{147}$$

and

$$d_h^{\pi^{\mathsf{eval}}}(s_{i^\star}, a_1) = \frac{1+\epsilon' w_{H+1}^{(i^\star,j,1)}}{S} \leq \frac{3}{2S}, \tag{148}$$

for all $h \geq H+2$. By Eq.(147) and (147), we obtain that

$$\frac{d_h^{\pi^{\mathsf{eval}}}(\bar{s}_j, a_0)}{d_h^{\pi^{\mathsf{b}}}(\bar{s}_j, a_0)} \leq 3, \tag{149}$$

for all $h \geq H+2$.

Combining Eq.(143), Eq.(144) and Eq.(145), we have

$$\sum_{h=1}^{2H+2} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{d^{\pi^{\mathsf{eval}}}(s,a)}{d^{\pi^{\mathsf{b}}}(s,a)} = \sum_{h\in[H-1]} \frac{d_h^{\pi^{\mathsf{eval}}}(s_{\mathrm{start}}, a_0)}{d_h^{\pi^{\mathsf{b}}}(s_{\mathrm{start}}, a_0)} + \frac{d_H^{\pi^{\mathsf{eval}}}(s_{\mathrm{start}}, a_{i^\star})}{d_H^{\pi^{\mathsf{eval}}}(s_{\mathrm{start}}, a_{i\star})} \tag{150}$$

$$+ \frac{d_{H+1}^{\pi^{\mathsf{eval}}}(s_{i^\star}, a_1)}{d_{H+1}^{\pi^{\mathsf{b}}}(s_{i^\star}, a_1)} + \sum_{h\geq H+2} \sum_{i\in[S]} \frac{d_h^{\pi^{\mathsf{eval}}}(\bar{s}_i, a_0)}{d_h^{\pi^{\mathsf{b}}}(\bar{s}_i, a_0)} \tag{151}$$

$$= H - 1 + K + C^\star K + \sum_{h\geq H+2} \sum_{i\in[S]} \frac{d_h^{\pi^{\mathsf{eval}}}(\bar{s}_i, a_0)}{d_h^{\pi^{\mathsf{b}}}(\bar{s}_i, a_0)} \tag{152}$$

$$\leq H - 1 + K + C^\star K + 3(H+1)S \leq C^\star(2H+2)(2S+1), \tag{153}$$

where the last second inequality is by Eq.(149) and the last inequality is by $C^\star \geq 2$. This completes the proof. □

Lemma H.3 demonstrate that π^{b} and π^{eval} satisfies $C^\star$-concentrability (Assumption B) in any $\mathcal{M}_{\mathbf{w}}$.

Notations. To distinguish with different MDP\Rs, we still use $V_h^\pi\left(\cdot; r, \mathbb{P}^{(\mathbf{w})}\right)$ to denote the value function of π in MDP $\mathcal{M}_{\mathbf{w}} \cup r$. Given two rewards r r' and $\mathbf{w} \in \overline{\mathcal{W}}^{\mathcal{I}}$, we define $d^{\pi^{\mathsf{eval}}}(r, r'; \mathbb{P}^{(\mathbf{w})})$ by:

$$d^{\pi^{\mathsf{eval}}}(r, r'; \mathbb{P}^{(\mathbf{w})}) := \sup_{\pi, h\in[H]} \mathbb{E}_{\mathbb{P}^{(\mathbf{w})}} \left| V_h^{\pi^{\mathsf{eval}}}(s_h; r, \mathbb{P}^{(\mathbf{w})}) - V_h^{\pi^{\mathsf{eval}}}(s_h; r', \mathbb{P}^{(\mathbf{w})}) \right|.$$

Correspondingly, given a parameter set Θ, two reward mappings $\mathscr{R}, \mathscr{R}'$, we define $D_\Theta^{\pi^{\mathsf{eval}}}(\mathscr{R}, \mathscr{R}'; \mathbb{P}^{(\mathbf{w})})$ by

$$D_\Theta^{\pi^{\mathsf{eval}}}(\mathscr{R}, \mathscr{R}'; \mathbb{P}^{(\mathbf{w})}) := \sup_{(V,A)\in\Theta} d^{\pi^{\mathsf{eval}}}\Big(\mathscr{R}(V, A), \mathscr{R}'(V, A); \mathbb{P}^{(\mathbf{w})}\Big).$$

In this section, we only consider the case that $\Theta = \overline{\mathcal{V}} \times \overline{\mathcal{A}}$.

Lemma H.4. *Given any $\mathbf{w} \in \overline{\mathcal{W}}^{\mathcal{I}}$, $w \neq v \in \overline{\mathcal{W}}$, and $i^\star \in [K]$. Let $\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}w)}$, $\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}v)}$ be the ground truth reward mappings induced by $\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}w}$, $\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}v}$ where $a = (i^\star, H+1, 1) \in \mathcal{I}$. Set $\Theta = \overline{\mathcal{V}} \times \overline{\mathcal{A}}$. For any rewarding mapping $\mathscr{R}$ and $\epsilon' \in (0, 1/2]$, we have*

$$7D_\Theta^{\pi^{\mathsf{eval}}}\left(\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}w)}, \mathscr{R}; \mathbb{P}^{\mathbf{w}\overset{a}{\leftarrow}w}\right) + D_\Theta^{\pi^{\mathsf{eval}}}\left(\mathscr{R}^{(\mathbf{w}\overset{a}{\leftarrow}v)}, \mathscr{R}; \mathbb{P}^{\mathbf{w}\overset{a}{\leftarrow}v}\right) \geq \frac{H\epsilon'}{16}.$$

Proof. We consider similar construction of bad parameter V^{bad}, A^{bad} in the Proof of Lemma G.3. To summarize, $\left(V^{\mathsf{bad}}, A^{\mathsf{bad}}\right)$ is given by

- We set $A_h^{\mathsf{bad}}(s,a) = 0$ for all $(h, s, a) \in [2H+2] \times \mathcal{S} \times \mathcal{A}$.
- We set V_h^{bad} by

$$V_h^{\mathsf{bad}}(s) := \begin{cases} \frac{(2H+2-h)\cdot(w_i - v_i)}{2} & \text{if } s = \bar{s}_i, h = H+2, \\ 0 & \text{otherwise.} \end{cases} \tag{154}$$

Similarly, we define r_w^{bad}, r_v^{bad} and r by

$$r_w^{\mathsf{bad}} := \mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}\big(V^{\mathsf{bad}}, A^{\mathsf{bad}}\big), \qquad r_w^{\mathsf{bad}} := \mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\big(V^{\mathsf{bad}}, A^{\mathsf{bad}}\big), \qquad r := \mathscr{R}\big(V^{\mathsf{bad}}, A^{\mathsf{bad}}\big).$$

By definition of $\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}$, $\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}$, we have

$$\begin{aligned}
\left|r_{w,H+1}^{\mathsf{bad}}(s_{i^\star}, a_1) - r_{v,H+1}^{\mathsf{bad}}(s_{i^\star}, a_1)\right| &= \left|\left[\left(\mathbb{P}_{H+2}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)} - \mathbb{P}_{H+2}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right)V_{H+1}^{\mathsf{bad}}\right](s_{i^\star}, a_1)\right| \\
&= \epsilon' \cdot \left|\sum_{i\in[S]} \frac{(w_i - v_i)V_{H+2}^{\mathsf{bad}}}{S}\right| \\
&= \frac{H\epsilon'}{2S} \cdot \sum_{i\in[S]} (w_i - v_i)^2 \geq \frac{H\epsilon'}{16},
\end{aligned} \tag{155}$$

where the last inequality follows from Eq.(116). By definition of $D_\Theta^{\pi^{\mathsf{eval}}}$, we have

$$\begin{aligned}
D_\Theta^{\pi^{\mathsf{eval}}}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}, \mathscr{R}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}\right) &\geq d^{\pi^{\mathsf{eval}}}\left(r_w^{\mathsf{bad}}, r; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}\right) \\
&\geq \mathbb{E}_{\mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}, \pi^{\mathsf{eval}}}\left|V_{H+2}^{\pi^{\mathsf{eval}}}(s; r_w^{\mathsf{bad}}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}) - V_{H+2}^{\pi^{\mathsf{eval}}}(s; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)})\right| \\
&= \sum_{i\in[S]} \frac{1 + \epsilon' \cdot w_i}{S}\left|V_{H+2}^{\pi^{\mathsf{eval}}}(\bar{s}_i; r_w^{\mathsf{bad}}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}) - V_{H+2}^{\pi^{\mathsf{eval}}}(\bar{s}_i; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)})\right| \\
&\geq \sum_{i\in[S]} \frac{1}{2S}\left|V_{H+2}^{\pi^{\mathsf{eval}}}(s; r_w^{\mathsf{bad}}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}) - V_{H+2}^{\pi^{\mathsf{eval}}}(s; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)})\right|,
\end{aligned} \tag{156}$$

where the last line is due to $\epsilon' \in (0, 1/2]$. By construction of π^{eval}, in MDP\R $\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}v}$, the visiting probability $d_{H+1}^{\pi^{\mathsf{eval}}}$ is given by

$$d_{H+1}^{\pi^{\mathsf{eval}}}\left(s_{i^\star}, a_1; \mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}w)}\right) = 1.$$

For $D_\Theta^{\pi^{\mathsf{eval}}}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}, \mathscr{R}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right)$, we also have

$$\begin{aligned}
&D_\Theta^{\pi^{\mathsf{eval}}}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}, \mathscr{R}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right) \geq d^{\pi^{\mathsf{eval}}}\left(r_v^{\mathsf{bad}}, r; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right) \\
&\geq \mathbb{E}_{\mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}, \pi^{\mathsf{eval}}}\left|V_{H+1}^{\pi^{\mathsf{eval}}}(s; r_v^{\mathsf{bad}}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}) - V_{H+1}^{\pi^{\mathsf{eval}}}(s; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)})\right| \\
&= \left|V_{H+1}^{\pi^{\mathsf{eval}}}(s_{i^\star}; r_v^{\mathsf{bad}}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}) - V_{H+1}^{\pi^{\mathsf{eval}}}(s_{i^\star}; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)})\right| \\
&= \Bigg|r_{v,H+1}^{\mathsf{bad}}(s_{i^\star}, a_1) - r_{H+1}(s_{i^\star}, a_1) \\
&\quad - \sum_{i\in[S]} \mathbb{P}_{H+1}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}(\bar{s}_i|s_{i^\star}, a_1) \cdot \left(V_{H+2}^{\pi^{\mathsf{eval}}}(\bar{s}_i; r_v^{\mathsf{bad}}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}) - V_{H+2}^{\pi^{\mathsf{eval}}}(\bar{s}_i; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)})\right)\Bigg| \\
&\geq \left|r_{v,H+1}^{\mathsf{bad}}(s_{i^\star}, a_1) - r_{H+1}(s_{i^\star}, a_1)\right| \\
&\quad - \sum_{i\in[S]} \frac{1 + \epsilon' \cdot v_i}{S} \cdot \left|V_{H+2}^{\pi^{\mathsf{eval}}}(\bar{s}_i; r_v^{\mathsf{bad}}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}) - V_{H+2}^{\pi^{\mathsf{eval}}}(\bar{s}_i; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)})\right|,
\end{aligned} \tag{157}$$

where the second last line is by the bellman equation and the last line is due to the triangle inequality. Since $\mathbb{P}_h^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)} = \mathbb{P}_h^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}$ and $r_{w,h}^{\mathsf{bad}} = r_{v,h}^{\mathsf{bad}}$ for all $h \geq H + 2$, we have

$$V_{H+2}^{\pi^{\mathsf{eval}}}(\bar{s}_i; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}) = V_{H+2}^{\pi^{\mathsf{eval}}}(\bar{s}_i; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}), \qquad V_{H+2}^{\pi^{\mathsf{eval}}}(\bar{s}_i; r_v^{\mathsf{bad}}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}) = V_{H+2}^{\pi^{\mathsf{eval}}}(\bar{s}_i; r_w^{\mathsf{bad}}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}). \tag{158}$$

Apply Eq.(158) to Eq.(157), we have

$$\begin{aligned}
&D^{\pi^{\mathsf{eval}}}_{\Theta}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}, \mathscr{R}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right) \\
&\geq \left|r^{\mathsf{bad}}_{v,H+1}(s_{i^\star}, a_1) - r_{H+1}(s_{i^\star}, a_1)\right| \\
&- \sum_{i\in[S]} \frac{1+\epsilon'\cdot v_i}{S} \cdot \left|V^{\pi^{\mathsf{eval}}}_{H+2}(\bar{s}_i; r^{\mathsf{bad}}_{v}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}) - V^{\pi^{\mathsf{eval}}}_{H+2}(\bar{s}_i; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)})\right| \\
&= \left|r^{\mathsf{bad}}_{v,H+1}(s_{i^\star}, a_1) - r_{H+1}(s_{i^\star}, a_1)\right| \\
&- \sum_{i\in[S]} \frac{1+\epsilon'\cdot v_i}{S} \cdot \left|V^{\pi^{\mathsf{eval}}}_{H+2}(\bar{s}_i; r^{\mathsf{bad}}_{w}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}) - V^{\pi^{\mathsf{eval}}}_{H+2}(\bar{s}_i; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)})\right| \\
&\geq \left|r^{\mathsf{bad}}_{v,H+1}(s_{i^\star}, a_1) - r_{H+1}(s_{i^\star}, a_1)\right| \\
&- \sum_{i\in[S]} \frac{3}{2S} \cdot \left|V^{\pi^{\mathsf{eval}}}_{H+2}(\bar{s}_i; r^{\mathsf{bad}}_{w}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}) - V^{\pi^{\mathsf{eval}}}_{H+2}(\bar{s}_i; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)})\right| \\
&\geq \left|r^{\mathsf{bad}}_{v,H+1}(s_{i^\star}, a_1) - r_{H+1}(s_{i^\star}, a_1)\right| - 3D^{\pi^{\mathsf{eval}}}_{\Theta}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}, \mathscr{R}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}\right),
\end{aligned} \tag{159}$$

where the last second inequality comes from $\epsilon' \in (0, 1/2]$ and the last inequality comes from Eq.(156).

We next bound $\left|r^{\mathsf{bad}}_{w,H+1}(s_{i^\star}, a_1) - r_{H+1}(s_{i^\star}, a_1)\right|$ by $D^{\pi^{\mathsf{eval}}}_{\Theta}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}, \mathscr{R}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}\right)$.

$$\begin{aligned}
&D^{\pi^{\mathsf{eval}}}_{\Theta}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}, \mathscr{R}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}\right) \geq d^{\pi^{\mathsf{eval}}}\left(r^{\mathsf{bad}}_{w}, r; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}\right) \\
&\geq \mathbb{E}_{\mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}, \pi^{\mathsf{eval}}}\left|V^{\pi^{\mathsf{eval}}}_{H+1}(s; r^{\mathsf{bad}}_{v}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}) - V^{\pi^{\mathsf{eval}}}_{H+1}(s; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)})\right| \\
&= \left|V^{\pi^{\mathsf{eval}}}_{H+1}(s_{i^\star}; r^{\mathsf{bad}}_{w}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}) - V^{\pi^{\mathsf{eval}}}_{H+1}(s_{i^\star}; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)})\right| \\
&= \Bigg| r^{\mathsf{bad}}_{w,H+1}(s_{i^\star}, a_1) - r_{H+1}(s_{i^\star}, a_1) \\
&- \sum_{i\in[S]} \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}_{H+1}(\bar{s}_i|s_{i^\star}, a_1) \cdot \left(V^{\pi^{\mathsf{eval}}}_{H+2}(\bar{s}_i; r^{\mathsf{bad}}_{w}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}) - V^{\pi^{\mathsf{eval}}}_{H+2}(\bar{s}_i; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)})\right)\Bigg| \\
&\geq \left|r^{\mathsf{bad}}_{w,H+1}(s_{i^\star}, a_1) - r_{H+1}(s_{i^\star}, a_1)\right| \\
&- \sum_{i\in[S]} \frac{1+\epsilon'\cdot w_i}{S} \cdot \left|V^{\pi^{\mathsf{eval}}}_{H+2}(\bar{s}_i; r^{\mathsf{bad}}_{w}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}) - V^{\pi^{\mathsf{eval}}}_{H+2}(\bar{s}_i; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)})\right| \\
&\geq \left|r^{\mathsf{bad}}_{w,H+1}(s_{i^\star}, a_1) - r_{H+1}(s_{i^\star}, a_1)\right| \\
&- \sum_{i\in[S]} \frac{3}{2S} \cdot \left|V^{\pi^{\mathsf{eval}}}_{H+2}(\bar{s}_i; r^{\mathsf{bad}}_{w}, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}) - V^{\pi^{\mathsf{eval}}}_{H+2}(\bar{s}_i; r, \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)})\right| \\
&\geq \left|r^{\mathsf{bad}}_{w,H+1}(s_{i^\star}, a_1) - r_{H+1}(s_{i^\star}, a_1)\right| - 3D^{\pi^{\mathsf{eval}}}_{\Theta}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}, \mathscr{R}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}\right),
\end{aligned} \tag{160}$$

where the last second inequality comes from $\epsilon' \in (0, 1/2]$ and the last inequality is by Eq.(156). Eq.(160) is equivalent to

$$4D^{\pi^{\mathsf{eval}}}_{\Theta}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}, \mathscr{R}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}\right) \geq \left|r^{\mathsf{bad}}_{w,H+1}(s_{i^\star}, a_1) - r_{H+1}(s_{i^\star}, a_1)\right|. \tag{161}$$

Combining Eq.(159) and Eq.(161), we conclude that

$$\begin{aligned}
&7D^{\pi^{\mathsf{eval}}}_{\Theta}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}, \mathscr{R}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}w\right)}\right) + D^{\pi^{\mathsf{eval}}}_{\Theta}\left(\mathscr{R}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}, \mathscr{R}; \mathbb{P}^{\left(\mathbf{w}\overset{a}{\leftarrow}v\right)}\right) \\
&\geq \left|r^{\mathsf{bad}}_{w,H+1}(s_{i^\star}, a_1) - r_{H+1}(s_{i^\star}, a_1)\right| + \left|r^{\mathsf{bad}}_{v,H+1}(s_{i^\star}, a_1) - r_{H+1}(s_{i^\star}, a_1)\right| \\
&\geq \left|r^{\mathsf{bad}}_{v,H+1}(s_{i^\star}, a_1) - r^{\mathsf{bad}}_{w,H+1}(s_{i^\star}, a_1)\right| \geq \frac{H\epsilon'}{16},
\end{aligned} \tag{162}$$

where the last inequality comes from Eq.(155). This completes the proof.

□

H.3. Proof for Theorem H.2

Our proof is similar to the proof of Theorem G.2 in Section G.

Proof of Theorem H.2. For any $\epsilon \in (0, 1/2]$, $\delta \in (0, 1)$, We consider an offline IRL algorithm $\mathfrak{A}$ such that for any IRL problem $(\mathcal{M}, \pi^{\mathsf{E}})$, we have

$$\mathop{\mathbb{P}}_{(\mathcal{M},\pi^{\mathsf{E}}),\mathfrak{A}}\left(D_{\Theta}^{\pi^{\mathsf{eval}}}\left(\mathscr{R}^{\star}, \widehat{\mathscr{R}}\right) \leq \epsilon\right) \geq 1-\delta, \tag{163}$$

where $\mathop{\mathbb{P}}_{(\mathcal{M},\pi^{\mathsf{E}}),\mathfrak{A}}$ denotes the probability measure induced by executing the algorithm $\mathfrak{A}$ in the IRL problem $(\mathcal{M}, \pi^{\mathsf{E}})$, $\mathscr{R}^{\star}$ is the ground truth reward mapping and $\widehat{\mathscr{R}}$ is the estimated reward mapping outputted by executing $\mathfrak{A}$ in $(\mathcal{M}, \pi^{\mathsf{E}})$. Fix $i^{\star} \in [S]$, We define the the identification function for any $\mathbf{w} \in \overline{\mathcal{W}}$ by

$$\mathbf{\Phi}_{\mathbf{w}} := \mathop{\arg\min}_{v \in \overline{\mathcal{W}}} D_{\Theta}^{\pi^{\mathsf{eval}}}\left(\mathscr{R}^{\left(\mathbf{w} \overset{a}{\leftarrow} v\right)}, \widehat{\mathscr{R}}; \mathbb{P}^{\left(\mathbf{w} \overset{a}{\leftarrow} v\right)}\right),$$

where $a = (i^{\star}, H+1, 1)$, $\mathscr{R}^{\left(\mathbf{w} \overset{a}{\leftarrow} v\right)}$ is the ground truth reward mapping induced by $(\mathcal{M}_{\mathbf{w} \overset{a}{\leftarrow} v}, \pi^{\mathsf{E}})$. Let $v^{\star} = \mathbf{\Phi}_{a,\mathbf{w}}$. For any $v \neq v^{\star} \in \overline{\mathcal{W}}$, by definition of $v^{\star}$, we have

$$D_{\Theta}^{\pi^{\mathsf{eval}}}\left(\mathscr{R}^{\left(\mathbf{w} \overset{a}{\leftarrow} v^{\star}\right)}, \widehat{\mathscr{R}}; \mathbb{P}^{\left(\mathbf{w} \overset{a}{\leftarrow} v^{\star}\right)}\right) \leq D_{\Theta}^{\pi^{\mathsf{eval}}}\left(\mathscr{R}^{\left(\mathbf{w} \overset{a}{\leftarrow} v\right)}, \widehat{\mathscr{R}}; \mathbb{P}^{\left(\mathbf{w} \overset{a}{\leftarrow} v\right)}\right).$$

By applying Lemma G.3, we obtain that

$$\frac{H\epsilon'}{16} \leq D_{\Theta}^{\pi^{\mathsf{eval}}}\left(\mathscr{R}^{\left(\mathbf{w} \overset{a}{\leftarrow} v\right)}, \widehat{\mathscr{R}}; \mathbb{P}^{\left(\mathbf{w} \overset{a}{\leftarrow} v\right)}\right) + 7D_{\Theta}^{\pi^{\mathsf{eval}}}\left(\mathscr{R}^{\left(\mathbf{w} \overset{a}{\leftarrow} v^{\star}\right)}, \widehat{\mathscr{R}}; \mathbb{P}^{\left(\mathbf{w} \overset{a}{\leftarrow} v^{\star}\right)}\right) \leq 8D_{\Theta}^{\mathsf{all}}\left(\mathscr{R}^{\left(\mathbf{w} \overset{a}{\leftarrow} v\right)}, \widehat{\mathscr{R}}; \mathbb{P}^{\left(\mathbf{w} \overset{a}{\leftarrow} v\right)}\right).$$

Next, we set $\epsilon' = \frac{256\epsilon}{H}$ which implies that

$$\frac{H\epsilon'}{16} \geq 16\epsilon. \tag{164}$$

Here, to employ Lemma H.4, we need $\epsilon' \in (0, 1/2]$ which is equivalent to $0 < \epsilon \leq H/512$. Then, it holds that

$$D_{\Theta}^{\pi^{\mathsf{eval}}}\left(\mathscr{R}^{\left(\mathbf{w} \overset{a}{\leftarrow} v\right)}, \widehat{\mathscr{R}}; \mathbb{P}^{\left(\mathbf{w} \overset{a}{\leftarrow} v\right)}\right) \geq 2\epsilon > \epsilon,$$

which implies that

$$\{v \neq \mathbf{\Phi}_{\mathbf{w}}\} \subseteq \left\{D_{\Theta}^{\pi^{\mathsf{eval}}}\left(\mathscr{R}^{\left(\mathbf{w} \overset{a}{\leftarrow} v\right)}, \widehat{\mathscr{R}}; \mathbb{P}^{\left(\mathbf{w} \overset{a}{\leftarrow} v\right)}\right) \geq \epsilon\right\}.$$

By Eq.(163), we have the following lower bound for the probability

$$\begin{aligned}
\delta &\geq \sup_{v \in \overline{\mathcal{W}}} \mathop{\mathbb{P}}_{(\mathcal{M}_{\mathbf{w} \overset{a}{\leftarrow} v},\pi^{\mathsf{E}}),\mathfrak{A}}\left(D_{\Theta}^{\pi^{\mathsf{eval}}}\left(\mathscr{R}^{\left(\mathbf{w} \overset{a}{\leftarrow} v\right)}, \widehat{\mathscr{R}}; \mathbb{P}^{\left(\mathbf{w} \overset{a}{\leftarrow} v\right)}\right) \geq \epsilon\right) \\
&\geq \sup_{v \in \overline{\mathcal{W}}} \mathop{\mathbb{P}}_{(\mathcal{M}_{\mathbf{w} \overset{a}{\leftarrow} v},\pi^{\mathsf{E}}),\mathfrak{A}}(v \neq \mathbf{\Phi}_{\mathbf{w}}) \\
&\geq \frac{1}{|\overline{\mathcal{W}}|} \sum_{v \in \overline{\mathcal{W}}} \mathop{\mathbb{P}}_{(\mathcal{M}_{\mathbf{w} \overset{a}{\leftarrow} v},\pi^{\mathsf{E}}),\mathfrak{A}}(v \neq \mathbf{\Phi}_{\mathbf{w}}).
\end{aligned} \tag{165}$$

By applying Theorem B.3 with $\mathbb{P}_0 = \mathop{\mathbb{P}}_{(\mathcal{M}_{\mathbf{w} \overset{a}{\leftarrow} 0},\pi^{\mathsf{E}}),\mathfrak{A}}$, $\mathbb{P}_w = \mathop{\mathbb{P}}_{(\mathcal{M}_{\mathbf{w} \overset{a}{\leftarrow} w},\pi^{\mathsf{E}}),\mathfrak{A}}$, we have

$$\frac{1}{|\overline{\mathcal{W}}|} \sum_{v \in \overline{\mathcal{W}}} \mathop{\mathbb{P}}_{(\mathcal{M}_{\mathbf{w} \overset{a}{\leftarrow} w},\pi^{\mathsf{E}}),\mathfrak{A}}(v \neq \mathbf{\Phi}_{\mathbf{w}}) \geq 1 - \frac{1}{\log|\overline{\mathcal{W}}|}\left(\frac{1}{|\overline{\mathcal{W}}|} \sum_{v \in \overline{\mathcal{W}}} D_{\mathrm{KL}}\left(\mathop{\mathbb{P}}_{(\mathcal{M}_{\mathbf{w} \overset{a}{\leftarrow} v},\pi^{\mathsf{E}}),\mathfrak{A}}, \mathop{\mathbb{P}}_{(\mathcal{M}_{\mathbf{w} \overset{a}{\leftarrow} 0},\pi^{\mathsf{E}}),\mathfrak{A}}\right) - \log 2\right). \tag{166}$$

Our next step is to bound the KL divergence. Using the same scheme in the proof Metelli et al. (2021, Theorem B.3), we can compute the KL-divergence as follows:

$$\begin{aligned}
&D_{\mathrm{KL}}(\underset{(\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}v},\pi^{\mathsf{E}}),\mathfrak{A}}{\mathbb{P}},\underset{(\mathcal{M}_{\mathbf{w}\overset{a}{\leftarrow}0},\pi^{\mathsf{E}}),\mathfrak{A}}{\mathbb{P}})\\
&=\mathbb{E}_{(\mathcal{M}_{\mathbf{w}}\overset{a}{\leftarrow}v,\pi^{\mathsf{E}}),\mathfrak{A}}\left[\sum_{t=1}^{N}D_{\mathrm{KL}}\left(\mathbb{P}^{(\mathbf{w}\overset{a}{\leftarrow}v)}h_t(\cdot\,|\,s_t,a_t),\mathbb{P}_{h_t}^{(\mathbf{w}\overset{a}{\leftarrow}0)}(\cdot\,|\,s_t,a_t)\right)\right]\\
&\le\mathbb{E}_{(\mathcal{M}_{\mathbf{w}}\overset{a}{\leftarrow}v,\pi^{\mathsf{E}}),\mathfrak{A}}[N_{h_a}(s_{i_a},a_{k_a})]D_{\mathrm{KL}}\left(\mathbb{P}_{H+1}^{(\mathbf{w}\overset{a}{\leftarrow}v)}(\cdot\,|\,s_{i^\star},a_1),\mathbb{P}_{H+1}^{(\mathbf{w}\overset{a}{\leftarrow}0)}(\cdot\,|\,s_{i^\star},a_1)\right)\\
&\le 2(\epsilon')^2\mathbb{E}_{(\mathcal{M}_{\mathbf{w}}\overset{a}{\leftarrow}v,\pi^{\mathsf{E}}),\mathfrak{A}}[N_{H+1}(s_{i^\star},a_1)],
\end{aligned}\tag{167}$$

where $N_h(s,a):=\sum_{t=1}^{N}\mathbf{1}\{(h_t,s_t,a_t)=(h,s,a)\}$ for any given $(h,s,a)\in[2H+2]\times\mathcal{S}\times\mathcal{A}$ and the last inequality comes from Metelli et al. (2021, Lemma E.4). Combining Eq.(165) and (166), we have

$$\delta\ge 1-\frac{1}{\log(|\overline{\mathcal{W}}|)}\left(\frac{1}{|\overline{\mathcal{W}}|}\sum_{v\in\overline{\mathcal{W}}}2(\epsilon')^2\mathbb{E}_{(\mathcal{M}_{\mathbf{w}}\overset{a}{\leftarrow}v,\pi^{\mathsf{E}}),\mathfrak{A}}[N_{h_a}(s_{i^\star},a_1)]-\log 2\right)$$

for any $\mathbf{w}$. It also holds that

$$\frac{1}{|\overline{\mathcal{W}}|}\sum_{v\in\overline{\mathcal{W}}}\mathbb{E}_{(\mathcal{M}_{\mathbf{w}}\overset{a}{\leftarrow}v,\pi^{\mathsf{E}}),\mathfrak{A}}[N_{H+1}(s_{i^\star},a_1)]\ge\frac{(1-\delta)\log|\overline{\mathcal{W}}|-\log 2}{2(\epsilon')^2}.\tag{168}$$

Hence, there exists a $\mathbf{w}^{\text{hard}}\in\overline{\mathcal{W}}$ such that

$$\mathbb{E}_{(\mathcal{M}_{\mathbf{w}^{\text{hard}}},\pi^{\mathsf{E}}),\mathfrak{A}}[N_{H+1}(s_{i^\star},a_1)]\ge\frac{(1-\delta)\log|\overline{\mathcal{W}}|-\log 2}{2(\epsilon')^2}.\tag{169}$$

By taking $\delta=1/3$, we have

$$\mathbb{E}_{(\mathcal{M}_{\mathbf{w}^{\text{hard}}},\pi^{\mathsf{E}}),\mathfrak{A}}[N_{H+1}(s_{i^\star},a_1)]\ge\frac{(1-\delta)\log|\overline{\mathcal{W}}|-\log 2}{2(\epsilon')^2}=\frac{2\log|\overline{\mathcal{W}}|-3\log 2}{6(\epsilon')^2}=\Omega\left(\frac{H^2S}{\epsilon^2}\right),\tag{170}$$

where the last equality follows from $\epsilon'=\frac{256\epsilon}{H}$ and $\log\left|\overline{\mathcal{W}}\right|\ge\frac{S}{10}$. By construction of π^{b}, it holds that $N_{H+1}(s_{i^\star},a_1)\sim\mathrm{Bin}\big(K,\frac{1}{C^\star K}\big)$, which implies that

$$\mathbb{E}_{(\mathcal{M}_{\mathbf{w}^{\text{hard}}},\pi^{\mathsf{E}}),\mathfrak{A}}[N]\ge C^\star K\cdot\Omega\left(\frac{H^2S}{\epsilon}\right)=\Omega\left(\frac{C^\star H^2SK}{\epsilon^2}\right)=\Omega\left(\frac{C^\star H^2S\min\{S,A\}}{\epsilon^2}\right).$$

□

I. Transfer learning

In this section, we explore the application of IRL in the context of transfer learning. Specifically, we apply the rewards learned by Algorithm 1 and Algorithm 2 to do RL in a *different* environment.

To distinguish different environments, given a transition dynamics $\mathbb{P}$ and policy π, we introduce the following notations: $\left\{d_h^{\mathbb{P},\pi}\right\}_{h\in[H]}$ represents the visitation probability induced by $\mathbb{P}$ and π, $d^{\mathbb{P},\pi}$ signifies the metric d^π evaluated on $\mathbb{P}$, and correspondingly $D_\Theta^{\mathbb{P},\pi}$ denotes the metric D_Θ^π evaluated on $\mathbb{P}$.

I.1. Transfer learning between IRL problems

We introduce the transfer learning setting outlined in Metelli et al. (2021), where they consider two IRL problems: $(\mathcal{M},\pi^{\mathsf{E}})$ (the source IRL problem), $(\mathcal{M}',(\pi')^{\mathsf{E}})$ (the target IRL problem). Here, $\mathcal{M}$, $\mathcal{M}'$ share the same state space and action space, but different dynamics. Suppose that we can learn the source IRL problem and obtain a solution r. However, r is not necessarily a solution for $(\mathcal{M}',(\pi')^{E})$, hence, in order to facilitate the transfer learning, we enforce the following assumption.

Assumption C. *If r represents a solution to the source IRL problem $(\mathcal{M}, \pi^{\mathsf{E}})$, it also stands as a solution to the target IRL problem $(\mathcal{M}', (\pi')^{E})$.*

Assumption C is also supposed in Metelli et al. (2021). We remark that in numerous practical scenarios, Assumption C may not be precisely met, but could be approximated: when the two IRL problems are very close[5] to each other, the solutions to the two IRL problems exhibit a high degree of similarity..

I.2. Transfer learning between two MDP\Rs

In this section, we consider a more general setting, where we focus solely on a source IRL problem and a target MDP\R.

We consider two MDP\Rs $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P})$ (source MDP\R), $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}')$ (the target MDP\R), which share the same state space and action space, but different dynamics, and an expert policy π^{E}. Let $\mathscr{R}^\star$ be the ground truth reward mapping of the IRL problem $(\mathcal{M}, \pi^{\mathsf{E}})$ and $\widehat{\mathscr{R}}$ be the estimated reward mapping learned from $(\mathcal{M}, \pi^{\mathsf{E}})$. In this setting, we evaluate $\widehat{\mathscr{R}}$ in $\mathcal{M}'$.

As we see in Section 1, Inverse reinforcement learning (IRL) and behavioral cloning (BC) are highly related. As mentioned in (Metelli et al., 2021), transfer learning makes IRL more powerful than BC, and a lot of literature has used IRL to do transfer learning (Syed & Schapire, 2007; Metelli et al., 2021; Abbeel & Ng, 2004; Fu et al., 2017; Levine et al., 2011).

Inspired by the single policy concentrability of policies, we propose the following transferability assumption.

Definition I.1 (Weak transferability)**.** *Given transitions $(\mathbb{P}, \mathbb{P}')$, and policies (π, π'), we say $(\mathbb{P}', \pi')$ is C^{wtran}-weakly transferable from $(\mathbb{P}, \pi)$ if it holds that*

$$\sup_{s,a} \frac{d_h^{\mathbb{P}',\pi'}(s,a)}{d_h^{\mathbb{P},\pi}(s,a)} \le C^{\mathsf{wtran}}.$$

Definition I.2 (Transferability)**.** *Given source and target transitions $\mathbb{P}$, $\mathbb{P}'$, and target policy π', we say π' is C^{tran}-transferable from $\mathbb{P}$ to $\mathbb{P}'$ if it holds that*

$$\inf_{\pi} \sup_{s,a} \frac{d_h^{\mathbb{P}',\pi'}(s,a)}{d_h^{\mathbb{P},\pi}(s,a)} \le C^{\mathsf{tran}}.$$

We remark that given a policy π and a dynamics $(\mathbb{P}, \mathbb{P}')$, transferability measures how hard one can learn the states π' frequently goes to in $\mathbb{P}$ in a different environment $\mathbb{P}'$ while given a policy pair (π, π') and a dynamics pair $(\mathbb{P}, \mathbb{P}')$, weak-transferability measures how hard one learn the states π frequently visits in $\mathbb{P}$ via policy π' in $\mathbb{P}'$. Without transferability, we can't obtain information on the policy of interest in the target MDP, which makes transfer learning hard to perform.

I.3. Theoretical guarantee

We then present the main theorems in this section.

Theorem I.3 (Transfer learning in the offline setting)**.** *Suppose $(\mathbb{P}', \pi^{\mathsf{eval}})$ is C^{wtran}-weakly transferable from $(\mathbb{P}, \pi^{\mathsf{b}})$ (Definition I.1). In addition, we assume π^{E} is well-posed (Definition A) when we receive feedback in option 1. Then for both options, with probability at least $1 - \delta$,* RLP *(Algorithm 1) outputs a reward mapping $\widehat{\mathscr{R}}$ such that*

$$D_{\Theta}^{\mathbb{P}',\pi^{\mathsf{eval}}}\Big(\mathscr{R}^\star, \widehat{\mathscr{R}}\Big) \le \epsilon, \Big[\widehat{\mathscr{R}}(V, A)\Big]_h(s,a) \le [\mathscr{R}^\star(V, A)]_h(s,a)$$

for all $(V, A) \in \Theta$ and $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$, as long as the number of episodes

$$K \ge \widetilde{\mathcal{O}}\bigg(\frac{H^4 S C^{\mathsf{wtran}} A \log \mathcal{N}}{\epsilon^2} + \frac{H^2 S C^{\mathsf{wtran}} A \eta}{\epsilon}\bigg).$$

Above, $\log \mathcal{N} := \log \mathcal{N}(\Theta; \epsilon/H)$, $\eta := \Delta^{-1} \mathbf{1}\{\text{option } 1\}$*, and $\widetilde{\mathcal{O}}(\cdot)$ hides* $\mathrm{polylog}(H, S, A, 1/\delta)$ *factors.*

Theorem I.4 (Transfer learning in the online setting)**.** *Suppose π^{E} is well-posed (Definition 11) when we receive feedback in option 1. Let $\mathscr{R}^\star$ be the ground truth reward mapping of IRL problem $(\mathcal{M}, \pi^{\mathsf{E}})$. Then for the online setting, for sufficiently*

[5] Here, we say $(\mathcal{M}, \pi^{\mathsf{E}})$ and $(\mathcal{M}', (\pi')^{E})$ are very close if the transitions of the two IRL problems are close under certain metric.

small $\epsilon \le H^{-9}(SA)^{-6}$, with probability at least $1-\delta$, RLE *(Algorithm 2) with $N = \widetilde{\mathcal{O}}(\sqrt{H^9S^7A^7K})$ outputs a reward mapping $\widehat{\mathscr{R}}$ such that*

$$\sup_{\pi^{\mathsf{eval}} \text{is } C^{\mathsf{tran}}\text{-transferable from } \mathbb{P} \text{ to } \mathbb{P}'} D_{\Theta}^{\mathbb{P}',\pi^{\mathsf{eval}}}(\mathscr{R}^\star, \widehat{\mathscr{R}}) \le \epsilon, \qquad \Big[\widehat{\mathscr{R}}(V,A)\Big]_h(s,a) \le [\mathscr{R}^\star(V,A)]_h(s,a)$$

for all $(V,A)\in\Theta$ and $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$, as long as the total the number of episodes

$$K + NH \ge \widetilde{\mathcal{O}}\Bigg(\frac{HSC^{\mathsf{tran}}A\big(C^{\mathsf{tran}} + H^3\log\mathcal{N}\big)}{\epsilon^2} + \frac{H^2SC^{\mathsf{tran}}A\eta}{\epsilon}\Bigg).$$

Application: Performing RL algorithms in different environments With Theorem I.3 and Theorem I.4 in place, as a concrete application, we consider utilizing rewards learned by IRL algorithms to execute RL algorithms in a different environment ($\mathcal{M}'$). The following two corollaries provide guarantees for the performance of learned rewards in executing RL algorithms in the offline and the online setting, respectively. Both of these corollaries are direct consequences of Proposition 6.

Corollary I.5 (Performing RL algorithms with learned rewards in the offline setting)**.** *Fix $\theta = (V,A)\in\Theta$, let $r^\theta := \mathscr{R}^\star(V,A)$ and $\widehat{r}^\theta := \widehat{\mathscr{R}}(V,A)$, where $\widehat{\mathscr{R}}$ are recovered reward mapping outputted by Algorithm 1. Suppose that there exists a policy π such that π is $\bar{\epsilon}$-optimal in MDP $\mathcal{M}'\cup r^\theta$ and $(\mathbb{P}',\pi)$ is C^{wtran}-weakly transferable from $(\mathbb{P},\pi^{\mathsf{b}})$ (Definition I.1). Let $\widehat{\pi}$ be an ϵ'-optimal policy in $\mathcal{M}'\cup\widehat{r}^\theta$ (learned by some RL algorithms with $\widehat{r}^\theta$). Under the same assumption of Theorem I.3 , for both options, we have $V_1^\star(s_1;\mathcal{M}'\cup r^\theta) - V_1^{\widehat{\pi}}(s_1;\mathcal{M}'\cup r^\theta) \le \epsilon+\epsilon'+2\bar{\epsilon}$, as long as the number of episodes*

$$K \ge \widetilde{\mathcal{O}}\Bigg(\frac{H^4SC^{\mathsf{wtran}}A\log\mathcal{N}}{\epsilon^2} + \frac{H^2SC^{\mathsf{wtran}}A\eta}{\epsilon}\Bigg).$$

Above, $\log\mathcal{N} := \log\mathcal{N}(\Theta;\epsilon/H)$, $\eta := \Delta^{-1}\mathbf{1}\{\text{option } 1\}$*, and $\widetilde{\mathcal{O}}(\cdot)$ hides* $\mathrm{polylog}(H,S,A,1/\delta)$ *factors.*

Corollary I.6 (Performing RL algorithms with learned rewards in the online setting)**.** *Fix $\theta = (V,A)\in\Theta$, let $r^\theta := \mathscr{R}^\star(V,A)$ and $\widehat{r}^\theta := \widehat{\mathscr{R}}(V,A)$, where $\widehat{\mathscr{R}}$ are recovered reward mapping outputted by Algorithm 2 with $N = \widetilde{\mathcal{O}}(\sqrt{H^9S^7A^7K})$.Suppose that there exists a policy π such that π is $\bar{\epsilon}$-optimal in MDP $\mathcal{M}'\cup r^\theta$ and π is C^{tran}-transferable from $\mathbb{P}$ to $\mathbb{P}'$ (Definition I.2). Let $\widehat{\pi}$ be an ϵ'-optimal policy in $\mathcal{M}'\cup\widehat{r}^\theta$ (learned by some RL algorithms with $\widehat{r}^\theta$), then for the online setting, for sufficiently small $\epsilon \le H^{-9}(SA)^{-6}$, we have $V_1^\star(s_1;\mathcal{M}'\cup r^\theta) - V_1^{\widehat{\pi}}(s_1;\mathcal{M}'\cup r^\theta) \le \epsilon+\epsilon'+2\bar{\epsilon}$, as long as the number of episodes*

$$K + NH \ge \widetilde{\mathcal{O}}\Bigg(\frac{HSC^{\mathsf{tran}}A\big(C^{\mathsf{tran}} + H^3\log\mathcal{N}\big)}{\epsilon^2} + \frac{H^2SC^{\mathsf{tran}}A\eta}{\epsilon}\Bigg).$$

Application: learning IRL problems by transfer learning We return to the topic of transfer learning between IRL problems. We note that our findings related to transfer learning between MDP\Rs can also be employed in the context of transfer learning between IRL problems. As the illustrated in Theorem I.3 and Theorem I.4, we can efficiently learn a $\widehat{\mathscr{R}}$ such that the distance $D_{\Theta}^{\pi^{\mathsf{eval}}}(\widehat{\mathscr{R}},\mathscr{R}^\star) \le 2\epsilon$, where $\mathscr{R}^\star$ is the ground truth reward mapping of $(\mathcal{M},\pi^{\mathsf{E}})$. By Assumption C, the rewards induced by $\mathscr{R}^\star$ are solutions of $\Big(\mathcal{M}',(\pi')^{\mathsf{E}}\Big)$, hence the rewards induced by $\widehat{\mathscr{R}}$ also approximate the solutions of $\Big(\mathcal{M}',(\pi')^{\mathsf{E}}\Big)$.

I.4. Proof of Theorem I.3

Note that under the same assumptions in Theorem I.3, the concentration event $\mathcal{E}$ defined in Lemma E.1 still holds with $1-\delta$. By the week-transferablity of $(\pi^{\mathsf{eval}},\pi^{\mathsf{b}})$, we have

$$\begin{aligned}\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{d_h^{\mathbb{P}',\pi^{\mathsf{eval}}}(s,a)}{d_h^{\mathbb{P},\pi^{\mathsf{b}}}(s,a)} &\le C^{\mathsf{wtran}}\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathbf{1}\left\{d_h^{\mathbb{P}',\pi^{\mathsf{eval}}}(s,a)\neq 0\right\}\\ &\le C^{\mathsf{wtran}}\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathbf{1}\left\{a\in\pi_h^{\mathsf{eval}}(\cdot|s)\right\} \le C^{\mathsf{wtran}}HSA. \qquad (171)\end{aligned}$$

For any $\theta = (V, A) \in \Theta$, define $r^\theta = \mathscr{R}^\star(V, A)$, and $\widehat{r}^\theta = \widehat{\mathscr{R}}(V, A)$. With Eq.(171) at hand, we can repeat the proof of Lemma E.2, thereby obtaining that

$$d^{\mathbb{P}', \pi^{\mathsf{eval}}}\left(r^\theta, \widehat{r}^\theta\right) \lesssim \frac{C^{\mathsf{wtran}} H^2 S A \eta \iota}{K} + \underbrace{\sum_{h \in [H]} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_h^{\mathbb{P}', \pi^{\mathsf{eval}}}(s, a) b_h^\theta(s, a)}_{\text{(I)}}, \tag{172}$$

holds on the event $\mathcal{E}$. where η, $b_h^\theta(s, a)$ are specified in Lemma E.1.

Furthermore, similar to Eq.(66), and through the application of the triangle inequality, we can decompose $\sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_h^{\mathbb{P}', \pi^{\mathsf{eval}}}(s, a) b_h^\theta(s, a)$ as follows:

$$\begin{aligned}
\text{(I)} &= \sum_{h \in [H]} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_h^{\mathbb{P}', \pi^{\mathsf{eval}}}(s, a) b_h^\theta(s, a) \\
&\lesssim \sum_{h \in [H]} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s, a) \cdot \left\{ \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H) \iota}{N_h^b(s, a) \vee 1} \left[\widehat{\mathbb{V}}_h V_{h+1}\right](s, a)} + \frac{H \log \mathcal{N}(\Theta; \epsilon/H) \iota}{N_h^b(s, a) \vee 1} \right\} \\
&\quad + \sum_{h \in [H]} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_h^{\pi^{\mathsf{eval}}}(s, a) \cdot \frac{\epsilon}{H} \left(1 + \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H) \iota}{N_h^b(s, a) \vee 1}}\right) \\
&\leq \underbrace{\sum_{h \in [H]} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_h^{\mathbb{P}', \pi^{\mathsf{eval}}}(s, a) \cdot \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H) \iota}{N_h^b(s, a) \vee 1} [\mathbb{V}_h V_{h+1}](s, a)}}_{\text{(I.a)}} \\
&\quad + \underbrace{\sum_{h \in [H]} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_h^{\mathbb{P}', \pi^{\mathsf{eval}}}(s, a) \cdot \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H) \iota}{N_h^b(s, a) \vee 1} \left[\left(\widehat{\mathbb{V}}_h - \mathbb{V}_h\right) V_{h+1}\right](s, a)}}_{\text{(I.b)}} \\
&\quad + \underbrace{\sum_{h \in [H]} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_h^{\mathbb{P}', \pi^{\mathsf{eval}}}(s, a) \cdot \frac{H \log \mathcal{N}(\Theta; \epsilon/H) \iota}{N_h^b(s, a) \vee 1}}_{\text{(I.c)}} \\
&\quad + \underbrace{\epsilon \cdot \sum_{h \in [H]} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_h^{\mathbb{P}', \pi^{\mathsf{eval}}}(s, a) \cdot \left(\frac{1}{H} + \sqrt{\frac{\log \mathcal{N}(\Theta; \epsilon/H) \iota}{H^2 \cdot N_h^b(s, a) \vee 1}}\right)}_{\text{(I.d)}}.
\end{aligned} \tag{173}$$

Thanks to Eq.(171), we can employ a similar argument as in the proof of Eq.(67), Eq.(72), Eq.(73), and Eq.(74), which allows us to deduce that

$$\begin{aligned}
\text{(I.a)} &\lesssim \sqrt{\frac{C^{\mathsf{wtran}} H^4 S A n^{\mathsf{E}} \log \mathcal{N}(\Theta; \epsilon/H) \iota}{K}}, \\
\text{(I.b)} &\lesssim \sqrt{\frac{C^{\mathsf{wtran}} H^2 S A \log \mathcal{N}(\Theta; \epsilon/H)}{K}} + \frac{C^{\mathsf{wtran}} H^3 S A \log \mathcal{N}(\Theta; \epsilon/H) \iota^{5/2}}{K}, \\
\text{(I.c)} &\lesssim \frac{C^{\mathsf{wtran}} H^2 S A \log \mathcal{N}(\Theta; \epsilon/H) \iota}{K}, \epsilon \cdot (1 + \sqrt{\frac{C^{\mathsf{wtran}} S A \log \mathcal{N}(\Theta; \epsilon/H) \iota}{K}}),
\end{aligned} \tag{174}$$

Combining Eq.(172), Eq.(173) and Eq.(174), we conclude that

$$\begin{aligned}
D_\Theta^{\mathbb{P}', \pi^{\mathsf{eval}}}\left(\mathscr{R}^\star, \widehat{\mathscr{R}}\right) = \sup_{\theta \in \Theta} d^{\mathbb{P}', \pi^{\mathsf{eval}}}\left(r^\theta, \widehat{r}^\theta\right) &\lesssim \frac{C^{\mathsf{wtran}} H^2 S A \eta \iota}{K} + \sqrt{\frac{C^{\mathsf{wtran}} H^4 S A n^{\mathsf{E}} \log \mathcal{N}(\Theta; \epsilon/H) \iota}{K}} \\
&\quad + \frac{C^{\mathsf{wtran}} H^3 S A n^{\mathsf{E}} \log \mathcal{N}(\Theta; \epsilon/H) \iota^{5/2}}{K} + \epsilon.
\end{aligned} \tag{175}$$

The right-hand-side is upper bounded by 2ϵ as long as

$$K \geq \widetilde{\mathcal{O}}\bigg(\frac{C^{\mathsf{wtran}} H^4 SA \log \mathcal{N}(\Theta; \epsilon/H)}{\epsilon^2} + \frac{C^{\mathsf{wtran}} H^2 SA\eta}{\epsilon}\bigg).$$

Here $poly \log \left(H, S, A, 1/\delta\right)$ are omitted.

I.5. Proof of Theorem I.4

Under the assumptions in Theorem I.4, the concentration event $\mathcal{E}$ defined in Lemma F.2 still holds with $1-\delta$. Fix π such that π satisfies C^{tran}-concentrability from $\mathbb{P}$ to $\mathbb{P}'$. We define

$$\bar{\mathcal{I}}_h := \bigg\{(s,a) \in \mathcal{S}\times\mathcal{A} \,|\, \widehat{d}_h^{\mathbb{P}',\pi}(s,a) \geq \frac{\xi}{N} + e_h^{\pi}(s,a)\bigg\},$$

for all $h \in [H]$. Similar to Eq.(96), we have the following decomposition:

$$\begin{aligned} d^{\mathbb{P}',\pi}\big(r_h^{\theta}, \widehat{r}_h^{\theta}\big) &\leq \sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} d_h^{\mathbb{P}',\pi}(s,a)\cdot\big|r_h^{\theta}(s,a) - \widehat{r}_h^{\theta}(s,a)\big| \\ &\leq \underbrace{\sum_{h\in[H]}\sum_{(s,a)\notin\mathcal{I}_h\cup\bar{\mathcal{I}}_h} d_h^{\mathbb{P}',\pi}(s,a)\cdot\big|r_h^{\theta}(s,a) - \widehat{r}_h^{\theta}(s,a)\big|}_{\text{(I)}} + \underbrace{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h\cup\bar{\mathcal{I}}_h} d_h^{\mathbb{P}',\pi}(s,a)\cdot\big|r^{\theta}(s,a) - \widehat{r}_h^{\theta}(s,a)\big|}_{\text{(II)}}, \end{aligned} \tag{176}$$

where set $\mathcal{I}_h$ is defined in Eq.(93).

We further decompose the term (I) as follows:

$$\text{(I)} \leq \underbrace{\sum_{h\in[H]}\sum_{(s,a)\notin\mathcal{I}_h} d_h^{\mathbb{P}',\pi}(s,a)\cdot\big|r_h^{\theta}(s,a) - \widehat{r}_h^{\theta}(s,a)\big|}_{\text{(I.a)}} + \underbrace{\sum_{h\in[H]}\sum_{(s,a)\notin\bar{\mathcal{I}}_h} d_h^{\mathbb{P}',\pi}(s,a)\cdot\big|r_h^{\theta}(s,a) - \widehat{r}_h^{\theta}(s,a)\big|}_{\text{(I.b)}}. \tag{177}$$

By the definition of transferability, there exists a policy π' such that

$$d_h^{\mathbb{P}',\pi}(s,a) \leq 2C^{\mathsf{tran}} d_h^{\mathbb{P},\pi'}(s,a),$$

for any $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$. For the term (I.a), we have

$$\text{(I.a)} = \sum_{h\in[H]}\sum_{(s,a)\notin\mathcal{I}_h\cup\bar{\mathcal{I}}_h} d_h^{\mathbb{P}',\pi}(s,a)\cdot\big|r_h^{\theta}(s,a) - \widehat{r}_h^{\theta}(s,a)\big| \leq 2C^{\mathsf{tran}} \sum_{(s,a)\notin\mathcal{I}_h\cup\bar{\mathcal{I}}_h} d_h^{\mathbb{P},\pi'}(s,a)\cdot\big|r_h^{\theta}(s,a) - \widehat{r}_h^{\theta}(s,a)\big| \tag{178}$$

Similar to Eq.(97), on the event $\mathcal{E}$, we have

$$\begin{aligned} \sum_{h\in[H]}\sum_{(s,a)\notin\mathcal{I}_h\cup\bar{\mathcal{I}}_h} d_h^{\mathbb{P},\pi'}(s,a)\cdot\big|r_h^{\theta}(s,a) - \widehat{r}_h^{\theta}(s,a)\big| &\leq \sum_{h\in[H]}\sum_{(s,a)\notin\mathcal{I}_h} d_h^{\mathbb{P},\pi'}(s,a)\cdot\big|r_h^{\theta}(s,a) - \widehat{r}_h^{\theta}(s,a)\big| \\ &\lesssim \frac{\xi H^2 SA}{N} + \frac{H^2 SA\eta}{K} + \sqrt{\frac{HSA}{K}}, \end{aligned}$$

which allows us to bound the term (I.a) as follows:

$$\text{(I.a)} \lesssim \frac{C^{\mathsf{tran}}\xi H^2 SA}{N} + \frac{C^{\mathsf{tran}} H^2 SA\eta}{K} + C^{\mathsf{tran}}\sqrt{\frac{HSA}{K}}. \tag{179}$$

For the term (I.b), on the event $\mathcal{E}$, we have

$$
\begin{aligned}
\text{(I.b)} &= \sum_{h\in[H]} \sum_{(s,a)\notin\bar{\mathcal{I}}_h} d_h^{\mathbb{P}',\pi}(s,a) \cdot \left| r^\theta(s,a) - \widehat{r}^\theta(s,a) \right| \\
&= \sum_{h\in[H]} \sum_{(s,a)\notin\bar{\mathcal{I}}_h} d_h^{\mathbb{P}',\pi}(s,a) \cdot \Bigg| - A_h(s,a)\Big(\mathbf{1}\left\{a \in \text{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\} - \mathbf{1}\left\{a \in \text{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\}\Big) \\
&\quad - \left[\left(\mathbb{P}_h - \widehat{\mathbb{P}}_h\right)V_{h+1}\right](s,a)(s,a) - b_h^\theta(s,a)\Bigg| \\
&\leq \sum_{h\in[H]} \sum_{(s,a)\notin\bar{\mathcal{I}}_h} d_h^{\mathbb{P}',\pi}(s,a) \cdot \Bigg\{ \left| A_h(s,a) \cdot \Big(\mathbf{1}\left\{a \in \text{supp}\left(\widehat{\pi}_h^{\mathsf{E}}(\cdot|s)\right)\right\} - \cdot\mathbf{1}\left\{a \in \text{supp}\left(\pi_h^{\mathsf{E}}(\cdot|s)\right)\right\}\Big) \right| \\
&\quad + \left| \left[(\mathbb{P}_h - \widehat{\mathbb{P}}_h)V_{h+1}\right](s,a) \right| + b_h^\theta(s,a) \Bigg\} \qquad \text{(by triangle inequality)} \\
&\overset{\text{(i)}}{\lesssim} H \cdot \sum_{h\in[H]} \sum_{(s,a)\notin\bar{\mathcal{I}}_h} d_h^{\mathbb{P}',\pi}(s,a) \\
&\overset{\text{(ii)}}{\leq} 2C^{\mathsf{tran}} H \cdot \sum_{h\in[H]} \sum_{(s,a)\notin\bar{\mathcal{I}}_h} \left(\widehat{d}_h^{\mathbb{P},\pi'}(s,a) + e_h^{\pi'}(s,a) + \frac{\xi}{N} \right) \\
&\overset{\text{(iii)}}{\lesssim} C^{\mathsf{tran}} H \cdot \sum_{h\in[H]} \sum_{(s,a)\notin\bar{\mathcal{I}}_h} \left(e_h^{\pi'}(s,a) + \frac{\xi}{N} \right) \\
&\lesssim \frac{C^{\mathsf{tran}} \xi H^2 SA}{N} + C^{\mathsf{tran}} \sqrt{\frac{HSA}{K}},
\end{aligned} \tag{180}
$$

where (i) is by $\|A_h\|_\infty, \|V_{h+1}\|_\infty, b_h^\theta(s,a) \leq H$, (ii) comes from Eq.(90) and the concentration event $\mathcal{E}(ii)$, and (iii) follows from the definition of $\bar{\mathcal{I}}_h$.

Combining Eq.(179) and Eq.(180), we can conclude that

$$
\text{(I)} \lesssim \frac{C^{\mathsf{tran}} \xi H^2 SA}{N} + \frac{C^{\mathsf{tran}} H^2 SA \eta}{K} + C^{\mathsf{tran}} \sqrt{\frac{HSA}{K}}. \tag{181}
$$

For the term (II), following a similar approach as in Eq.(98), we have

$$
\begin{aligned}
\text{(II)} &= \sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h\cup\bar{\mathcal{I}}_h} d_h^{\mathbb{P}',\pi}(s,a)\cdot b_h^{\theta}(s,a)\\
&= \sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h\cup\bar{\mathcal{I}}_h} d_h^{\mathbb{P}',\pi}(s,a)\cdot \min\Bigg\{\sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\widehat{N}_h^b(s,a)\vee 1}\Big[\widehat{\mathbb{V}}_h V_{h+1}\Big](s,a)}+\frac{H\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\widehat{N}_h^b(s,a)\vee 1}\\
&\quad+\frac{\epsilon}{H}\left(1+\sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\widehat{N}_h^b(s,a)\vee 1}}\right),H\Bigg\}\\
&\overset{\text{(i)}}{\leq} \sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h\cup\bar{\mathcal{I}}_h} d_h^{\mathbb{P}',\pi}(s,a)\cdot\Bigg\{\min\Bigg\{\sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\widehat{N}_h^b(s,a)\vee 1}\Big[\widehat{\mathbb{V}}_h V_{h+1}\Big](s,a)},H\Bigg\}+\frac{H\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\widehat{N}_h^b(s,a)\vee 1} \qquad (182)\\
&\quad+\frac{\epsilon}{H}\left(1+\sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\widehat{N}_h^b(s,a)\vee 1}}\right)\Bigg\}\\
&\overset{\text{(ii)}}{\leq} \sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h\cup\bar{\mathcal{I}}_h} d_h^{\mathbb{P}',\pi}(s,a)\cdot\Bigg\{\sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota\Big[\widehat{\mathbb{V}}_h V_{h+1}\Big](s,a)+H}{\widehat{N}_h^b(s,a)\vee 1+1/H}}+\frac{H\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\widehat{N}_h^b(s,a)\vee 1}\\
&\quad+\frac{\epsilon}{H}\left(1+\sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\widehat{N}_h^b(s,a)\vee 1}}\right)\Bigg\}\\
&\overset{\text{(iii)}}{=} \underbrace{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h\cup\bar{\mathcal{I}}_h} d_h^{\mathbb{P}',\pi}(s,a)\cdot\sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota\Big[\widehat{\mathbb{V}}_h V_{h+1}\Big](s,a)+H}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big]+1/H}}}_{\text{(II.a)}}\\
&\quad+\underbrace{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h\cup\bar{\mathcal{I}}_h} d_h^{\mathbb{P}',\pi}(s,a)\cdot\frac{H\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big]+1/H}}_{\text{(II.b)}}\\
&\quad+\underbrace{\frac{\epsilon}{H}\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h\cup\bar{\mathcal{I}}_h} d_h^{\mathbb{P}',\pi}(s,a)\cdot\left(1+\sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big]+1/H}}\right)}_{\text{(II.c)}}. \qquad (183)
\end{aligned}
$$

For the term (II.a), by the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\text{(II.a)} &\leq \sqrt{C^{\mathsf{tran}}} \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{I}_h\cup\bar{\mathcal{I}}_h} \sqrt{d_h^{\mathbb{P}',\pi}(s,a)\cdot\left(\log\mathcal{N}(\Theta;\epsilon/H)\iota[\mathbb{V}_h V_{h+1}](s,a)+H\right)} \\
&\quad \cdot\sqrt{\frac{d_h^{\mathbb{P},\pi'}(s,a)}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big]+1/H}} \\
&\lesssim \sqrt{C^{\mathsf{tran}}} \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{I}_h\cup\bar{\mathcal{I}}_h} \sqrt{d_h^{\mathbb{P}',\pi}(s,a)\cdot\left(\log\mathcal{N}(\Theta;\epsilon/H)\iota[\mathbb{V}_h V_{h+1}](s,a)+H\right)} \\
&\quad \cdot\sqrt{\frac{2\widehat{d}_h^{\mathbb{P},\pi'}(s,a)+2e_h^{\pi'}(s,a)+\frac{\xi}{2N}}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big]+1/H}} \\
&\lesssim \sqrt{C^{\mathsf{tran}}} \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{I}_h\cup\bar{\mathcal{I}}_h} \sqrt{d_h^{\mathbb{P}',\pi}(s,a)\cdot\left(\log\mathcal{N}(\Theta;\epsilon/H)\iota[\mathbb{V}_h V_{h+1}](s,a)+H\right)} \\
&\quad \cdot\sqrt{\frac{\widehat{d}_h^{\mathbb{P},\pi'}(s,a)}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big]+1/H}} \\
&\leq \sqrt{C^{\mathsf{tran}}} \underbrace{\left\{\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\mathbb{P}',\pi}(s,a)\cdot\left(\log\mathcal{N}(\Theta;\epsilon/H)\iota[\mathbb{V}_h V_{h+1}](s,a)+H\right)\right\}^{1/2}}_{\text{(II.a.1)}} \\
&\quad \times \underbrace{\left\{\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\widehat{d}_h^{\mathbb{P},\pi'}(s,a)}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big]+1/H}\right\}^{1/2}}_{\text{(II.a.2)}}.
\end{aligned}$$

Following similar approaches as in Eq.(100) and Eq.(101), we have

$$\text{(II.a.1)} \lesssim \sqrt{H^3\log\mathcal{N}(\Theta;\epsilon/H)\iota}, \qquad \text{(II.a.2)} \lesssim \sqrt{\frac{HSA}{K}}, \tag{184}$$

which implies that

$$\text{(II.a)} \lesssim \sqrt{\frac{C^{\mathsf{tran}}H^4SA\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K}}. \tag{185}$$

For the term (II.b), by Eq.(16), we have

$$\begin{aligned}
\text{(II.b)} &= \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\mathbb{P}',\pi}(s,a)\cdot\frac{H\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big]+1/H} \\
&= \frac{C^{\mathsf{tran}}}{K}\cdot\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\mathbb{P},\pi'}(s,a)\cdot\frac{H\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big]+1/KH} \\
&\lesssim \frac{C^{\mathsf{tran}}}{K}\cdot\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \widehat{d}_h^{\mathbb{P},\pi'}(s,a)\cdot\frac{H\log\mathcal{N}(\Theta;\epsilon/H)\iota}{\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big]+1/KH} \\
&\lesssim \frac{C^{\mathsf{tran}}H^2SA\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K}.
\end{aligned} \tag{186}$$

For the term (II.c), we have

$$
\begin{aligned}
\text{(II.c)} &= \frac{\epsilon}{H}\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h\cup\bar{\mathcal{I}}_h} d_h^{\mathbb{P},\pi}(s,a)\cdot\left(1+\sqrt{\frac{\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big]+1/H}}\right)\\
&= \epsilon+\frac{\sqrt{C^{\mathsf{tran}}}\epsilon}{H}\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h\cup\bar{\mathcal{I}}_h}\sqrt{d_h^{\mathbb{P}',\pi}(s,a)}\cdot\sqrt{\frac{d_h^{\mathbb{P},\pi'}(s,a)\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big]+1/H}}\\
&\lesssim \epsilon+\frac{\sqrt{C^{\mathsf{tran}}}\epsilon}{H}\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h\cup\bar{\mathcal{I}}_h}\sqrt{d_h^{\mathbb{P}',\pi}(s,a)}\cdot\sqrt{\frac{\widehat{d}_h^{\mathbb{P},\pi'}(s,a)\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big]+1/H}}\\
&\le \epsilon+\frac{\epsilon}{H}\sqrt{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h\cup\bar{\mathcal{I}}_h}d_h^{\mathbb{P}',\pi}(s,a)}\cdot\sqrt{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{I}_h}\frac{\widehat{d}_h^{\mathbb{P},\pi'}(s,a)\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K\mathbb{E}_{\pi'\sim\mu^{\mathsf{b}}}\Big[\widehat{d}_h^{\pi'}(s,a)\Big]+1/H}}\\
&\le \epsilon\left(1+\sqrt{\frac{C^{\mathsf{tran}}SA\log\mathcal{N}(\Theta;\epsilon/H)\iota}{HK}}\right),
\end{aligned}
\tag{187}
$$

where the second last line is by the Cauchy-Schwarz inequality and the last line is by Eq.(16).

Then combining Eq.(185) Eq.(186), and Eq.(187), we obtain the bound for the term (II)

$$
\begin{aligned}
\text{(II)} &\lesssim \text{(II.a)}+\text{(II.b)}+\text{(II.c)}\\
&\lesssim \sqrt{\frac{C^{\mathsf{tran}}H^4SA\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K}}+\frac{C^{\mathsf{tran}}H^2SA\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K}+\epsilon(1+\sqrt{\frac{C^{\mathsf{tran}}\log\mathcal{N}(\Theta;\epsilon/H)\iota}{HK}})\\
&\lesssim \sqrt{\frac{C^{\mathsf{tran}}H^4SA\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K}}+\epsilon,
\end{aligned}
\tag{188}
$$

where the last line is from $\epsilon<1$.

Finally, combining Eq.(181) and Eq.(188), we get the final bound

$$
\begin{aligned}
D_{\Theta}^{\mathsf{all}}\Big(\mathscr{R}^{\star},\widehat{\mathscr{R}}\Big) &= \sup_{\pi,\theta\in\Theta} d^{\pi}\big(r_h^{\theta},\widehat{r}_h^{\theta}\big)\le\text{(I)}+\text{(II)}\\
&\lesssim \frac{C^{\mathsf{tran}}\xi H^2SA}{N}+\sqrt{\frac{C^{\mathsf{tran}}H^4SA\log\mathcal{N}(\Theta;\epsilon/H)\iota}{K}}+\frac{C^{\mathsf{tran}}H^2SA\eta}{K}+C^{\mathsf{tran}}\sqrt{\frac{HSA}{K}}+\epsilon
\end{aligned}
$$

Hence, we can guarantee $D_{\Theta}^{\mathsf{all}}\Big(\mathscr{R}^{\star},\widehat{\mathscr{R}}\Big)\le 2\epsilon$, provided that

$$
\begin{aligned}
&KH\ge N\ge\widetilde{\mathcal{O}}\Big(\sqrt{H^9S^7A^7K}\Big),\\
&K\ge\widetilde{\mathcal{O}}\left(\frac{C^{\mathsf{tran}}HSA\big(C^{\mathsf{tran}}+H^3\log\mathcal{N}(\Theta;\epsilon/H)\big)}{\epsilon^2}+\frac{C^{\mathsf{tran}}H^2SA\eta}{\epsilon}\right)
\end{aligned}
\tag{189}
$$

Here $poly\log(H,S,A,1/\delta)$ are omitted. Similar to the proof of Theorem 11, suppose $\epsilon\le H^{-9}(SA)^{-6}$, set $N=\widetilde{\mathcal{O}}$, when

$$
K\ge\widetilde{\mathcal{O}}\left(\frac{C^{\mathsf{tran}}HSA\big(C^{\mathsf{tran}}+H^3\log\mathcal{N}(\Theta;\epsilon/H)\big)}{\epsilon^2}+\frac{C^{\mathsf{tran}}H^2SA\eta}{\epsilon}\right),
\tag{190}
$$

Eq.(189) holds. And at this time, the total sample complexity is

$$
K+NH\ge\widetilde{\mathcal{O}}\left(\frac{C^{\mathsf{tran}}HSA\big(C^{\mathsf{tran}}+H^3\log\mathcal{N}(\Theta;\epsilon/H)\big)}{\epsilon^2}+\frac{C^{\mathsf{tran}}H^2SA\eta}{\epsilon}\right).
\tag{191}
$$