Offline Inverse RL: New Solution Concepts and Provably Efficient Algorithms

Filippo Lazzati [1] **Mirco Mutti** [2] **Alberto Maria Metelli** [1]

Abstract

Inverse reinforcement learning (IRL) aims to recover the reward function of an *expert* agent from demonstrations of behavior. It is well-known that the IRL problem is fundamentally ill-posed, i.e., many reward functions can explain the demonstrations. For this reason, IRL has been recently reframed in terms of estimating the *feasible reward set* (Metelli et al., 2021), thus, postponing the selection of a single reward. However, so far, the available formulations and algorithmic solutions have been proposed and analyzed mainly for the *online* setting, where the learner can interact with the environment and query the expert at will. This is clearly unrealistic in most practical applications, where the availability of an *offline* dataset is a much more common scenario. In this paper, we introduce a novel notion of feasible reward set capturing the opportunities and limitations of the offline setting and we analyze the complexity of its estimation. This requires the introduction of an original learning framework that copes with the intrinsic difficulty of the setting, for which the data coverage is not under control. Then, we propose two computationally and statistically efficient algorithms, `IRLO` and `PIRLO`, for addressing the problem. In particular, the latter adopts a specific form of *pessimism* to enforce the novel, desirable property of *inclusion monotonicity* of the delivered feasible set. With this work, we aim to provide a panorama of the challenges of the offline IRL problem and how they can be fruitfully addressed.

[1]Politecnico di Milano, Milan, Italy [2]Technion, Haifa, Israel. Correspondence to: Filippo Lazzati <filippo.lazzati@polimi.it>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Inverse reinforcement learning (IRL), also called inverse optimal control, consists of recovering a reward function from expert's demonstrations (Russell, 1998). Specifically, the reward is required to be *compatible* with the expert's behavior, i.e., it shall make the expert's policy optimal. As pointed out in Arora & Doshi (2018), IRL allows mitigating the challenging task of the manual specification of the reward function, thanks to the presence of demonstrations, and provides an effective method for *imitation learning* (Osa et al., 2018). In opposition to mere *behavioral cloning*, IRL allows focusing on the expert *intent* (instead of *behavior*), and, for this reason, it has the potential to reveal the underlying objectives that drive the expert's choices. In this sense, IRL enables *interpretability*, improving the interaction with the expert by explaining and predicting its behavior, and *transferability*, as the reward (more than a policy) can be employed under environment shifts (Adams et al., 2022).

One of the main concerns of IRL is that the problem is inherently *ill-posed* or *ambiguous* (Ng & Russell, 2000), i.e., there exists a variety of reward functions compatible with expert's demonstrations. In the literature, many criteria for the selection of a single reward among the compatible ones were proposed (e.g., Ng & Russell, 2000; Ratliff et al., 2006; Ziebart et al., 2008; Boularias et al., 2011). Nevertheless, the ambiguity issue has limited the theoretical understanding of the IRL problem for a long time.

Recently, IRL has been reframed by Metelli et al. (2021) into the problem of computing the *set* of all rewards compatible with expert's demonstrations, named *feasible reward set* (or just *feasible set*). By postponing the choice of a specific reward within the feasible set, this formulation has opened the doors to a new perspective that has enabled a deeper theoretical understanding of the IRL problem. The majority of previous works on the reconstruction of the feasible set have focused mostly on the *online* setting (e.g., Metelli et al., 2021; Lindner et al., 2022; Zhao et al., 2023; Metelli et al., 2023), in which the learner is allowed to actively interact with the environment and with the expert to collect samples.

Although these works succeeded in obtaining sample efficient algorithms and represent a fundamental step ahead in

the understanding of the challenges of the IRL problem (e.g., providing sample complexity lower bounds), the underlying basic assumption that the learner is allowed to govern the exploration and query the expert wherever is far from being realistic. Indeed, the most common IRL applications are naturally framed in an *offline* scenario, in which the learner is given in advance a dataset of trajectories of the expert (and, possibly, an additional dataset collected with a *behavioral* policy, e.g., Boularias et al. 2011). Typically, no further interaction with the environment and with the expert is allowed (Likmeta et al., 2021). The offline setting has been widely studied in (forward) *reinforcement learning* (RL, Sutton & Barto, 2018), and a surge of works have analyzed the problem from theoretical and practical perspectives (e.g., Munos, 2007; Levine et al., 2020; Buckman et al., 2020; Yu et al., 2020; Jin et al., 2021). In this context, a powerful technique is represented by *pessimism*, which discourages the learner from assigning credit to options that have not been sufficiently explored in the available dataset, allowing for sample efficiency guarantees (Buckman et al., 2020).

The IRL offline setting has been investigated for the problem of recovering the feasible set in the recent preprint (Zhao et al., 2023). The authors consider the same feasible set definition employed for the online case, which enforces the optimality of the expert's policy *in every state* (Metelli et al., 2021; Lindner et al., 2022). However, in the offline setting, this learning target is unrealistic unless the dataset covers the full space. This implies that the produced rewards can be safely used in forward RL when the behavioral policy covers the whole reachable portion of the state-action space *only*. For this reason, Zhao et al. (2023) apply a form of *pessimism* which allows delivering rewards that make the expert's policy ϵ-optimal even in the presence of partial covering of the behavioral policy but only when the latter is sufficiently close to the expert's. These demanding requirements, however, collide with the intuition that, regardless of the sampling policy, if we observe the expert's actions, we can deliver *at least one* reward, making the expert optimal.[1]

Desired Properties In this paper, we seek to develop novel appropriate *solution concepts* for the feasible reward set and new effective actionable *algorithms* for recovering them in the offline IRL setting. Specifically, we aim at fulfilling the following three *key properties*:

- (i) (*Sample Efficiency*) We should output, with high probability, an estimated feasible set using a number of samples polynomial w.r.t. the desired accuracy, error probability, and relevant sizes of the problem.
- (ii) (*Computational Efficiency*) We should be able to check the *membership* of a candidate reward in the feasible set in polynomial time w.r.t. the relevant sizes of the problem.
- (iii) (*Inclusion Monotonicity*) We should output one estimated feasible set that *includes* and one that *is included* in the true feasible set with high probability.

[1]For instance, simply assign 0 when playing the expert actions and -1 otherwise.

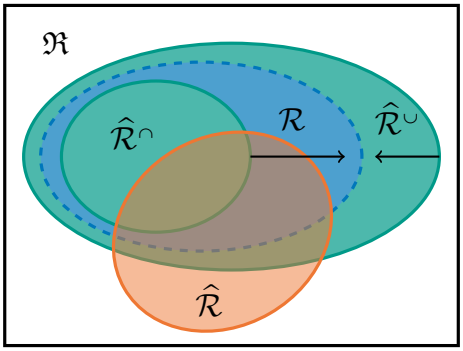

Figure 1. $\mathfrak{R}$ = set of all rewards, $\mathcal{R}$ = true feasible set, $\widehat{\mathcal{R}}^{\cap}$ and $\widehat{\mathcal{R}}^{\cup}$ = examples of *inclusion monotonic* estimated feasible set (i.e., $\widehat{\mathcal{R}}^{\cap} \subseteq \mathcal{R} \subseteq \widehat{\mathcal{R}}^{\cup}$), $\widehat{\mathcal{R}}$ = example of *inclusion non-monotonic* estimated feasible set (i.e., $\widehat{\mathcal{R}} \not\subseteq \mathcal{R}$ and $\mathcal{R} \not\subseteq \widehat{\mathcal{R}}$).

While properties (i) and (ii) are commonly requested, (iii) deserves some comments. *Inclusion monotonicity*, intuitively, guarantees that we produce a set that *does not exclude* any reward function that can be feasible and a set that *includes only* reward functions that are surely feasible, given the current samples (Figure 1). This, remarkably, allows delivering (with high probability) reward functions that make the expert's policy optimal (not just ϵ-optimal) regardless of the accuracy with which the feasible set is recovered.

Contributions The contributions of this paper are summarized as follows:

- We propose a novel definition of *feasible set* that takes into account the intrinsic challenges of the *offline* setting (i.e., partial covering). Moreover, we introduce appropriate *solution concepts*, which are *learnable* based on the coverage of the given dataset (Section 3).
- We adapt the *probably approximately correct* (PAC) framework from Metelli et al. (2023) to our offline setting by proposing novel *semimetrics* which, differently from previous works, allow us to naturally deal with *unbounded* rewards (Section 4).
- We present a novel algorithm, named **IRLO** (Inverse Reinforcement Learning for Offline data), for solving offline IRL. We show that it satisfies the requirements of (i) sample and (ii) computational efficiency (Section 5).
- After having formally defined the notion of *inclusion monotonicity*, we propose a *pessimism*-based algorithm, named **PIRLO** (Pessimistic Inverse Reinforcement Learning for Offline data), that achieves (iii) inclusion monotonicity preserving sample and computational efficiency, at the price of a larger sample complexity (Section 6).
- We discuss a specific application of our algorithm **PIRLO**

for *reward sanity check* (Section 7).
- We present a negative result for *offline* IRL when only data from a deterministic expert are available (Section 8).

Additional related works are reported in Appendix A. The proofs of all the results are reported in the Appendix B-J.

2. Preliminaries

Notation Given a finite set $\mathcal{X}$, we denote by $|\mathcal{X}|$ its cardinality and by $\Delta^{\mathcal{X}} := \{q \in [0,1]^{|\mathcal{X}|} \mid \sum_{x\in\mathcal{X}} q(x) = 1\}$ the simplex on $\mathcal{X}$. Given two sets $\mathcal{X}$ and $\mathcal{Y}$, we denote the set of conditional distributions as $\Delta^{\mathcal{X}}_{\mathcal{Y}} := \{q : \mathcal{Y} \to \Delta^{\mathcal{X}}\}$. Given $N \in \mathbb{N}$, we denote $[\![N]\!] := \{1, \dots, N\}$. Given an equivalence relation $\equiv \subseteq \mathcal{X} \times \mathcal{X}$, and an item $x \in \mathcal{X}$, we denote by $[x]_{\equiv}$ the equivalence class of x.

Markov Decision Processes (MDPs) without Reward A finite-horizon *Markov decision process* (MDP, Puterman, 1994) without reward is defined as $\mathcal{M} := \langle \mathcal{S}, \mathcal{A}, \mu_0, p, H \rangle$, where $\mathcal{S}$ is the finite state space ($S := |\mathcal{S}|$), $\mathcal{A}$ is the finite action space ($A := |\mathcal{A}|$), $\mu_0 \in \Delta^{\mathcal{S}}$ is the initial-state distribution, $p = \{p_h\}_{h\in[\![H]\!]}$ where $p_h \in \Delta^{\mathcal{S}}_{\mathcal{S}\times\mathcal{A}}$ for every $h \in [\![H]\!]$ is the transition model, and $H \in \mathbb{N}$ is the horizon. A policy is defined as $\pi = \{\pi_h\}_{h\in[\![H]\!]}$ where $\pi_h \in \Delta^{\mathcal{A}}_{\mathcal{S}}$ for every $h \in [\![H]\!]$. $\mathbb{P}_{p,\pi}$ denotes the trajectory distribution induced by π and $\mathbb{E}_{p,\pi}$ the expectation w.r.t. $\mathbb{P}_{p,\pi}$ (we omit μ_0 in the notation). The state-action visitation distribution induced by p and π is defined as $\rho^{p,\pi}_h(s,a) := \mathbb{P}_{p,\pi}(s_h = s, a_h = a)$ and the state visitation distribution as $\rho^{p,\pi}_h(s) := \sum_{a\in\mathcal{A}} \rho^{p,\pi}_h(s,a)$, so that $\sum_{s\in\mathcal{S}} \rho^{p,\pi}_h(s) = 1$ for every $h \in [\![H]\!]$.

Additional Definitions The sets of transition models, policies, and rewards are denoted as $\mathcal{P} := \Delta^{\mathcal{S}}_{\mathcal{S}\times\mathcal{A}\times[\![H]\!]}$, $\Pi := \Delta^{\mathcal{A}}_{\mathcal{S}\times[\![H]\!]}$, and $\mathfrak{R} := \{r : \mathcal{S}\times\mathcal{A}\times[\![H]\!] \to \mathbb{R}\}$, respectively.[2] For every $h \in [\![H]\!]$, we define the set of states and state-action pairs reachable by π at stage $h \in [\![H]\!]$ as $\mathcal{S}^{p,\pi}_h := \{s \in \mathcal{S} \mid \rho^{p,\pi}_h(s) > 0\}$ and $\mathcal{Z}^{p,\pi}_h := \{(s,a) \in \mathcal{S}\times\mathcal{A} \mid \rho^{p,\pi}_h(s,a) > 0\}$, respectively. Moreover, we define $\mathcal{S}^{p,\pi} := \{(s,h) : h \in [\![H]\!], s \in \mathcal{S}^{p,\pi}_h\}$ and $\mathcal{Z}^{p,\pi} := \{(s,a,h) : h \in [\![H]\!], (s,a) \in \mathcal{Z}^{p,\pi}_h\}$, with cardinality $S^{p,\pi} \leqslant SH$ and $Z^{p,\pi} \leqslant SAH$, respectively. We refer to these sets as the "support" of $\rho^{p,\pi}$. We denote the cardinality of the largest set $\mathcal{S}^{p,\pi}_h$ varying $h \in [\![H]\!]$, as $S^{p,\pi}_{\max} := \max_{h\in[\![H]\!]} |\mathcal{S}^{p,\pi}_h| \leqslant S$. Finally, we denote the minimum of the state-action distribution on set $\mathcal{Y} \subseteq \mathcal{S}\times\mathcal{A}\times[\![H]\!]$ as $\rho^{\pi,\mathcal{Y}}_{\min} := \min_{(s,a,h)\in\mathcal{Y}} \rho^{p,\pi}_h(s,a)$.

Value Functions and Optimality The *Q-function* of policy π with transition model p and reward function r is defined as $Q^{\pi}_h(s,a;p,r) := \mathbb{E}_{p,\pi}[\sum_{t=h}^{H} r_t(s_t,a_t) | s_h = s, a_h = a]$ and the optimal Q-function as $Q^{*}_h(s,a;p,r) := \max_{\pi\in\Pi} Q^{\pi}_h(s,a;p,r)$. The *utility* (i.e., expected return) of policy π under the initial-state distribution μ_0 is given by $J(\pi;\mu_0,p,r) := \mathbb{E}_{s\sim\mu_0, a\sim\pi(\cdot|s)}[Q^{\pi}_1(s,a;p,r)]$ and the optimal utility by $J^*(\mu_0,p,r) := \max_{\pi\in\Pi} J(\pi;\mu_0,p,r)$. An *optimal policy* π^* is a policy that maximizes the utility $\pi^* \in \arg\max_{\pi\in\Pi} J(\pi;\mu_0,p,r)$. The existence of a deterministic optimal policy is guaranteed (Puterman, 1994).

Equivalence Relations We introduce two *equivalence* relations: $\equiv_{\overline{\mathcal{S}}}$ (over policies) and $\equiv_{\overline{\mathcal{Z}}}$ (over transition models), defined for arbitrary $\overline{\mathcal{S}} \subseteq \mathcal{S}\times[\![H]\!]$ and $\overline{\mathcal{Z}} \subseteq \mathcal{S}\times\mathcal{A}\times[\![H]\!]$. Specifically, let $\pi, \pi' \in \Pi$ be two policies, we have:

$$\pi \equiv_{\overline{\mathcal{S}}} \pi' \quad \text{iff} \quad \forall (s,h) \in \overline{\mathcal{S}} : \pi_h(\cdot|s) = \pi'_h(\cdot|s). \tag{1}$$

Similarly, let $p, p' \in \mathcal{P}$, be two transition models, we have:

$$p \equiv_{\overline{\mathcal{Z}}} p' \quad \text{iff} \quad \forall (s,a,h) \in \overline{\mathcal{Z}} : p_h(\cdot|s,a) = p_h(\cdot|s,a). \tag{2}$$

We will often use $\overline{\mathcal{S}} = \mathcal{S}^{p,\pi}$ and $\overline{\mathcal{Z}} = \mathcal{Z}^{p,\pi}$ for some $p \in \mathcal{P}$ and $\pi \in \Pi$. Intuitively, the equivalence relation $\equiv_{\mathcal{S}^{p,\pi}}$ (resp. $\equiv_{\mathcal{Z}^{p,\pi}}$) group policies (resp. transition models) indistinguishable given the support $\mathcal{S}^{p,\pi}$ (resp. $\mathcal{Z}^{p,\pi}$) of $\rho^{p,\pi}$.

Offline Setting We assume the availability of two datasets $\mathcal{D}^b = \{\langle s^{b,i}_1, a^{b,i}_1, \dots, s^{b,i}_{H-1}, a^{b,i}_{H-1}, s^{b,i}_H \rangle\}_{i\in[\![\tau^b]\!]}$ and $\mathcal{D}^E = \{\langle s^{E,i}_1, a^{E,i}_1, \dots, s^{E,i}_{H-1}, a^{E,i}_{H-1}, s^{E,i}_H \rangle\}_{i\in[\![\tau^E]\!]}$ of τ^b and τ^E independent trajectories collected by playing a *behavioral policy* π^b and the *expert's policy* π^E, respectively. Furthermore, we enforce the following assumption.

Assumption 2.1 (Expert's covering)**.** *The behavioral policy π^b plays with non-zero probability the actions prescribed by the expert's policy π^E in its support $\mathcal{S}^{p,\pi^E}$:*

$$\forall (s,h) \in \mathcal{S}^{p,\pi^E} : \qquad \pi^b_h(\pi^E_h(s)|s) > 0.$$

Assumption 2.1 holds when $\pi^b = \pi^E$ and generalizes that setting when the behavioral policy π^b is "more explorative", possibly playing actions other than expert's ones.[3] It should be remarked that Assumption 2.1 is useful but not strictly necessary. As we will explain later on, it is possible to avoid it by using all samples $\mathcal{D}^E \cup \mathcal{D}^b$ to compute the various estimates that will be needed. Even though this seems reasonable, from a practical viewpoint, it complicates the theoretical analysis of the algorithms. Thus, we will enforce Assumption 2.1 in the following for simplicity.

3. Solution Concepts for Offline IRL

In this section, we introduce a novel definition of *feasible reward set*, discuss its learnability properties, and propose suitable solution concepts to be targeted for the *offline* IRL.

[2] We remark that we consider *real-valued* rewards without requiring boundedness.

[3] We elaborate on the limits of learning with just a dataset collected with the expert's policy π^E in Section 8. Moreover, we discuss how we can use a single dataset collected with π^b, at the price of a slightly larger sample complexity in Appendix D.1.

A New Definition of Feasible Set Let us start by recalling the original definition of *feasible set* presented in the literature and discussing its limitations for offline IRL.

Definition 3.1 ("Old" Feasible Set $\overline{\mathcal{R}}_{p,\pi^E}$, Metelli et al. 2021)**.** *Let $\mathcal{M}$ be an MDP without reward and let π^E be the deterministic expert's policy. The* "old" feasible set $\overline{\mathcal{R}}_{p,\pi^E}$ *of rewards compatible with π^E in $\mathcal{M}$ is defined as:*[4]

$$\overline{\mathcal{R}}_{p,\pi^E} := \{r \in \mathfrak{R} \,|\, \forall (s,h) \in \mathcal{S} \times [\![H]\!], \forall a \in \mathcal{A}: Q_h^{\pi^E}(s, \pi_h^E(s); p, r) \geqslant Q_h^{\pi^E}(s,a;p,r)\}. \quad (3)$$

In words, $\overline{\mathcal{R}}_{p,\pi^E}$ contains all the reward functions that make the expert's policy optimal *in every* state-stage pair $(s,h) \in \mathcal{S} \times [\![H]\!]$. However, forcing the optimality of π^E in states that are never reached from the initial-state distribution μ_0 is unnecessary (and even impossible) if our ultimate goal is to use the learned reward function r to train a policy π^* that achieves the maximum utility, i.e., $\pi^* \in \arg\max_{\pi \in \Pi} J(\pi; \mu_0, p, r)$. This suggests an alternative definition of feasible set.

Definition 3.2 (Feasible Set $\mathcal{R}_{p,\pi^E}$)**.** *Let $\mathcal{M}$ be an MDP without reward and let π^E be the deterministic expert's policy. The* feasible set $\mathcal{R}_{p,\pi^E}$ *of rewards compatible with π^E in $\mathcal{M}$ is defined as:*

$$\mathcal{R}_{p,\pi^E} := \{r \in \mathfrak{R} \,|\, J(\pi^E; \mu_0, p, r) = J^*(\mu_0, p, r)\}.$$

In words, $\mathcal{R}_{p,\pi^E}$ contains all the reward functions that make the expert's policy π^E a utility maximizer. Clearly, since Definition 3.1 enforces optimality *uniformly* over $\mathcal{S} \times [\![H]\!]$, we have the inclusion $\overline{\mathcal{R}}_{p,\pi^E} \subseteq \mathcal{R}_{p,\pi^E}$, where the equality holds when $\mathcal{S}^{p,\pi^E} = \mathcal{S} \times [\![H]\!]$, i.e., $[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}} = \{\pi^E\}$. The following result formalizes the intuition that for $\mathcal{R}_{p,\pi^E}$, differently from $\overline{\mathcal{R}}_{p,\pi^E}$, the expert's policy π^E has to be optimal (as in Equation 3) in a subset of $\mathcal{S} \times [\![H]\!]$ only.

Theorem 3.1. *In the setting of Definition 3.2, the feasible reward set $\mathcal{R}_{p,\pi^E}$ satisfies:*

$$\mathcal{R}_{p,\pi^E} = \{r \in \mathfrak{R} \,|\, \forall \overline{\pi} \in [\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}, \forall (s,h) \in \mathcal{S}^{p,\pi^E}, \forall a \in \mathcal{A}: Q_h^{\overline{\pi}}(s, \pi_h^E(s); p, r) \geqslant Q_h^{\overline{\pi}}(s,a;p,r)\}. \quad (4)$$

Theorem 3.1 shows that the optimal action induced by a reward $r \in \mathcal{R}_{p,\pi^E}$ outside $\mathcal{S}^{p,\pi^E}$, i.e., outside the support of ρ^{p,π^E} induced by the expert's policy π^E, is not relevant. The optimality condition of Equation (4) is requested for all the policies $\overline{\pi}$ that play the expert's action within its support. Intuitively, those policies cover the same portion of state space as π^E, i.e., $\mathcal{S}^{p,\overline{\pi}} = \mathcal{S}^{p,\pi^E}$ and, since they all prescribe the same action in there,[5] they all achieve the same utility, i.e., $J(\overline{\pi}; \mu_0, p, r) = J(\pi^E; \mu_0, p, r) = J^*(\mu_0, p, r)$. Thus, if we train an RL agent with a reward function $\widehat{r} \in \mathcal{R}_{p,\pi^E} \setminus \overline{\mathcal{R}}_{p,\pi^E}$, among the optimal policies we obtain a policy $\widehat{\pi} \in [\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}$, i.e., a policy that plays optimal (expert) actions inside $\mathcal{S}^{p,\pi^E}$. Clearly, $\widehat{\pi}$ will prescribe different actions than π^E outside $\mathcal{S}^{p,\pi^E}$, but this is irrelevant since those states will never be reached by $\widehat{\pi}$. This has important consequences from the offline IRL perspective. Indeed, we can recover this new notion $\mathcal{R}_{p,\pi^E}$ (Definition 3.2) without the knowledge of π^E in the states outside $\mathcal{S}^{p,\pi^E}$. Instead, to learn the old notion $\overline{\mathcal{R}}_{p,\pi^E}$ (Definition 3.1), we would need to enforce that the policy used to collect samples (either π^E or π^b) covers the full space $\mathcal{S} \times [\![H]\!]$.[6]

Solution Concepts and Learnability To compute the feasible set $\mathcal{R}_{p,\pi^E}$, we need to learn the expert's policy $\pi_h^E(s)$ in every $(s,h) \in \mathcal{S}^{p,\pi^E}$ and the transition model $p_h(\cdot|s,a)$ in every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]$, so that we are able to compare the Q-functions. In the *online* setting (e.g., Metelli et al., 2021), this is a reasonable requirement because the learner can explore the environment and, thus, collect samples over the whole $\mathcal{S} \times \mathcal{A} \times [\![H]\!]$ space.[7] However, in our *offline* setting, even in the limit of infinite samples, triples $(s,a,h) \notin \mathcal{Z}^{p,\pi^b}$, i.e., outside the support of ρ^{p,π^b} are never sampled. Thus, we can identify the transition model p up to its equivalence class $[p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}$ only. Intuitively, this means that, unless $\mathcal{Z}^{p,\pi^b} = \mathcal{S} \times \mathcal{A} \times [\![H]\!]$, i.e., π^b covers the entire space, since $\mathcal{R}_{p,\pi^E}$ depends on the value of the transition model in the whole $\mathcal{S} \times \mathcal{A} \times [\![H]\!]$, the problem of estimating the feasible set $\mathcal{R}_{p,\pi^E}$ offline is not *learnable*.[6] Thus, instead of learning $\mathcal{R}_{p,\pi^E}$ directly, we propose to target as solution concepts (i) the *largest learnable* set of rewards *contained* into $\mathcal{R}_{p,\pi^E}$, and (ii) the *smallest learnable* set of rewards that *contains* $\mathcal{R}_{p,\pi^E}$, defined as follows.

Definition 3.3 (Sub- and Super-Feasible Sets)**.** *Let $\mathcal{M}$ be an MDP without reward and let π^E be the deterministic expert's policy. We define the* sub-feasible set $\mathcal{R}^{\cap}_{p,\pi^E}$ *and the* super-feasible set $\mathcal{R}^{\cup}_{p,\pi^E}$ *as:*

$$\mathcal{R}^{\cap}_{p,\pi^E} := \bigcap_{p' \in [p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}} \mathcal{R}_{p',\pi^E}, \quad \mathcal{R}^{\cup}_{p,\pi^E} := \bigcup_{p' \in [p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}} \mathcal{R}_{p',\pi^E}.$$

[4] Actually, Metelli et al. (2021) consider rewards bounded in $[0,1]$, while we consider all real-valued rewards in $\mathfrak{R}$.

[5] It is worth noting that, since $(s,h) \in \mathcal{S}^{p,\pi^E}$, the following identity hold: $Q_h^{\overline{\pi}}(s, \pi_h^E(s); p, r) = Q_h^{\pi^E}(s, \pi_h^E(s); p, r)$.

[6] A formal definition of *learnability* and the proofs that $\overline{\mathcal{R}}_{p,\pi^E}$ and $\mathcal{R}_{p,\pi^E}$ are not learnable under partial cover (i.e., $\mathcal{S}^{p,\pi^E} \neq \mathcal{S} \times [\![H]\!]$ and $\mathcal{Z}^{p,\pi^b} \neq \mathcal{S} \times \mathcal{A} \times [\![H]\!]$) are reported in Appendix C.

[7] This is true for the *generative* model case. In a *forward* model, in which we are allowed to interact through trajectories, we just need to learn the transition model in all state-action pairs (s,a,h) reachable from μ_0 with *any policy*, i.e., $(s,a,h) \in \bigcup_{\pi \in \Pi} \mathcal{Z}^{p,\pi}$.

Since $p \in [p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}$, we "squeeze" the feasible set $\mathcal{R}_{p,\pi^E}$ between these two learnable solution, i.e., $\mathcal{R}^{\cap}_{p,\pi^E} \subseteq \mathcal{R}_{p,\pi^E} \subseteq \mathcal{R}^{\cup}_{p,\pi^E}$. A more explicit representation is given as follows:

$$\mathcal{R}^{\cap}_{p,\pi^E} = \{r \in \mathfrak{R} | \forall p' \in [p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}, \forall \overline{\pi} \in [\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}, \\ \forall (s,h) \in \mathcal{S}^{p,\pi^E}, \forall a \in \mathcal{A} : Q^{\overline{\pi}}_h(s, \pi^E_h(s); p', r) \geqslant Q^{\overline{\pi}}_h(s,a;p',r)\},$$

$$\mathcal{R}^{\cup}_{p,\pi^E} = \{r \in \mathfrak{R} | \exists p' \in [p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}, \forall \overline{\pi} \in [\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}, \\ \forall (s,h) \in \mathcal{S}^{p,\pi^E}, \forall a \in \mathcal{A} : Q^{\overline{\pi}}_h(s, \pi^E_h(s); p', r) \geqslant Q^{\overline{\pi}}_h(s,a;p',r)\}.$$

Intuitively, to be robust against the missing knowledge of the transition model outside $\mathcal{Z}^{p,\pi^b}$, we have to account for all the possible $p' \in [p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}$ and retain the rewards compatible with *all* of them (for the sub-feasible set $\mathcal{R}^{\cap}_{p,\pi^E}$) and with *at least one* of them (for super-feasible set $\mathcal{R}^{\cup}_{p,\pi^E}$), as apparent from the quantifiers. Moreover, when $\mathcal{Z}^{p,\pi^b} = \mathcal{S} \times \mathcal{A} \times [\![H]\!]$, i.e., $[p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}} = \{p\}$, we have the equality: $\mathcal{R}^{\cap}_{p,\pi^E} = \mathcal{R}_{p,\pi^E} = \mathcal{R}^{\cup}_{p,\pi^E}$. We now show that the $\mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E}$ are indeed the *tightest learnable* subset and superset of $\mathcal{R}_{p,\pi^E}$ (formal statement and proof in Appendix B).

Theorem 3.2. (**Informal**) *Let $\mathcal{M}$ be an MDP without reward, let π^E and π^b be the deterministic expert's policy and the behavioral policy, respectively. Then, $\mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E}$ are the* tightest subset and superset of $\mathcal{R}_{p,\pi^E}$ learnable *from data collected in $\mathcal{M}$ by executing π^b and π^E.*

4. PAC Framework

We now propose a PAC framework for learning $\mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E}$ from datasets $\mathcal{D}^E$ and $\mathcal{D}^b$, collected with π^E and π^b. We first present the functions to evaluate the dissimilarity between feasible sets and then define the PAC requirement.

Dissimilarity Functions Being $\mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E}$ sets of rewards, we need (i) a function to assess the dissimilarity between items (i.e., reward functions), and (ii) a way of converting it into a dissimilarity function between sets (i.e., the sub- and super-feasible sets) (Metelli et al., 2021). For (i), we propose the following two semimetrics.

Definition 4.1 (Semimetrics d and d_∞ between rewards)**.** *Let $\mathcal{M}$ be an MDP without reward and let π^E be the expert's policy. Let π^b be the behavioral policy and let $\{\mathcal{Z}^{p,\pi^b}_h\}_h$ be its support. Given two reward functions $r, \widehat{r} \in \mathfrak{R}$, we define $d : \mathfrak{R} \times \mathfrak{R} \to \mathbb{R}$ and $d_\infty : \mathfrak{R} \times \mathfrak{R} \to \mathbb{R}$ as:*

$$d(r,\widehat{r}) := \frac{1}{M(r,\widehat{r})} \sum_{h \in [\![H]\!]} \Bigg(\mathop{\mathbb{E}}_{(s,a) \sim \rho^{p,\pi^b}_h} \left| r_h(s,a) - \widehat{r}_h(s,a) \right| \\ + \max_{(s,a) \notin \mathcal{Z}^{p,\pi^b}_h} \left| r_h(s,a) - \widehat{r}_h(s,a) \right| \Bigg),$$

$$d_\infty(r,\widehat{r}) := \frac{1}{M(r,\widehat{r})} \sum_{h \in [\![H]\!]} \|r_h - \widehat{r}_h\|_\infty,$$

where $M(r,\widehat{r}) := \max\{\|r\|_\infty, \|\widehat{r}\|_\infty\}$. Moreover, we conventionally set both d and d_∞ to 0 when $M(r,\widehat{r}) = 0$.

First, d_∞ corresponds to the ℓ_∞-norm between reward functions, while d combines the ℓ_1-norm between rewards in $\mathcal{Z}^{p,\pi^b}$ weighted by the visitation distribution of the behavioral policy ρ^{p,π^b} and the ℓ_∞-norm outside $\mathcal{Z}^{p,\pi^b}$. The intuition is that, inside $\mathcal{Z}^{p,\pi^b}$, we weigh the error based on the number of samples, which are collected by π^b. Instead, outside $\mathcal{Z}^{p,\pi^b}$, we can afford the ℓ_∞-norm because we adopt as solution concepts $\mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E}$ that intrinsically manage the lack of samples so that we can confidently achieve zero error in that region. Second, it is easy to verify that both d and d_∞ are *semimetrics*.[8] Third, the two semimetrics are related by the following double inequality, where $\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min} > 0$ by definition:

Proposition 4.1. *For any $r, r' \in \mathfrak{R}$, it holds that:*

$$d(r,r') \leqslant 2 d_\infty(r,r') \leqslant \frac{2}{\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{min}} d(r,r').$$

Moreover, the normalization term $1/M(r,\widehat{r})$ enforces that $d(r,r')$ and $d_\infty(r,r')$ lie in $[0, 2H]$ for every $r, r' \in \mathfrak{R}$. Differently from previous works (e.g., Metelli et al., 2021; Lindner et al., 2022), this term allows to deal with *(unbounded) real-valued rewards* more naturally and effectively, at the price of accepting a relaxed triangular inequality. We stress that we have chosen distances d and d_∞ since they enforce non-zero weight to the absolute difference between rewards at all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]$. This property allows us to control the distance between the optimal value function and the value function of the policy $\hat{\pi}^*$, i.e., the optimal policy under the recovered reward $\widehat{r}$. This can be obtained with an analogous reasoning as that contained in Section 4.3 of Metelli et al. (2023). More specifically, let

$$d^G_{V^*}(r,\widehat{r}) \\ := \frac{1}{M(r,\widehat{r})} \sup_{\hat{\pi}^* \in \Pi^*(\widehat{r})} \max_{(s,h) \in \mathcal{S} \times [\![H]\!]} |V^*_h(s;r) - V^{\hat{\pi}^*}_h(s;r)|,$$

be the adaptation of the dissimilarity index defined in Metelli et al. (2023), which measures the distance between the optimal value function $V^*(\cdot;r)$ (under a ground-truth reward r) and the value function $V^{\hat{\pi}^*}(\cdot;r)$ (under the same ground-truth reward r) of the policy $\hat{\pi}^*$ that is learned using the recovered reward $\widehat{r}$. Then, it can be shown that:

Proposition 4.2. *For any $r, r' \in \mathfrak{R}$, it holds that:*

$$d^G_{V^*}(r,\widehat{r}) \leqslant 2 d_\infty(r,\widehat{r}) \leqslant \frac{2 d(r,\widehat{r})}{\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min}}.$$

[8] A *semimetric* fulfills all the properties of a *metric* except for the triangular inequality. We show in Appendix I that our semimetrics fulfill a "relaxed" form of triangular inequality.

Clearly, a small value of d entails a small value of d^G_{V*}. Thus, controlling distances d and d_∞, by enforcing non-zero weight to the absolute difference between rewards at all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]$, we can control the distance between value functions. Finally, as mentioned above, notice that we can get rid of Assumption 2.1 by replacing the expectation w.r.t. ρ^{p,π^b} in the definition of d with some mixture between π^E and π^b. However, for the sake of simplicity, we continue with the current definition.

Next, to obtain a dissimilarity function between reward sets (ii), we make use of the Hausdorff distance.

Definition 4.2 (Hausdorff distance, Rockafellar & Wets 1998)**.** *Let $\mathcal{R}, \widehat{\mathcal{R}} \subseteq \mathfrak{R}$ be two sets of reward functions, and let $c \in \{d, d_\infty\}$. The* Hausdorff distance *between $\mathcal{R}$ and $\widehat{\mathcal{R}}$ with inner distance $c : \mathfrak{R} \times \mathfrak{R} \to \mathbb{R}$ is defined as:*

$$\mathcal{H}_c(\mathcal{R}, \widehat{\mathcal{R}}) := \max \left\{ \sup_{r \in \mathcal{R}} \inf_{\hat{r} \in \widehat{\mathcal{R}}} c(r, \hat{r}), \sup_{\hat{r} \in \widehat{\mathcal{R}}} \inf_{r \in \mathcal{R}} c(r, \hat{r}) \right\}. \quad (5)$$

Moreover, we abbreviate $\mathcal{H}_{d_\infty}$ with $\mathcal{H}_\infty$.

Since the feasible sets are closed (see Appendix I), using d or d_∞, the Hausdorff distance is a semimetric and satisfies a relaxed triangle inequality as well. Thus, $\mathcal{H}_c(\mathcal{R}, \widehat{\mathcal{R}}) = 0$ if and only if the two sets coincide, i.e., $\mathcal{R} = \widehat{\mathcal{R}}$.

(ϵ, δ)-PAC Requirement We now formally define the *sample efficiency* requirement. To distinguish between the two semimetrics d and d_∞, we denote by c-IRL the problem of estimating $\mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E}$ under $\mathcal{H}_c$, where $c \in \{d, d_\infty\}$.

Definition 4.3 ((ϵ, δ)-PAC Algorithm)**.** *Let $\epsilon \in [0, 2H]$ and $\delta \in (0,1)$. An algorithm $\mathfrak{A}$ outputting the estimated sub- and super-feasible sets $\widehat{\mathcal{R}}^{\cap}$ and $\widehat{\mathcal{R}}^{\cup}$ is* (ϵ, δ)-PAC *for c-IRL if:*

$$\mathbb{P}_{(p,\pi^E,\pi^b)} \Big(\big\{ \mathcal{H}_c(\mathcal{R}^{\cap}_{p,\pi^E}, \widehat{\mathcal{R}}^{\cap}) \leqslant \epsilon \big\} \cap \big\{ \mathcal{H}_c(\mathcal{R}^{\cup}_{p,\pi^E}, \widehat{\mathcal{R}}^{\cup}) \leqslant \epsilon \big\} \Big) \geqslant 1 - \delta,$$

where $\mathbb{P}_{(p,\pi^E,\pi^b)}$ denotes the probability measure induced by π^E and π^b in $\mathcal{M}$. The sample complexity *is the number of trajectories τ^E and τ^b in $\mathcal{D}^E$ and $\mathcal{D}^b$, respectively.*

5. Inverse Reinforcement Learning for Offline data (IRLO)

Our goal is to devise an algorithm that is (i) statistically efficient, (ii) computationally efficient, and that provides (iii) guarantees about the inclusion monotonicity property. As a warm-up, in this section, we present IRLO (Inverse Reinforcement Learning for Offline data), fulfilling (i) and (ii), but not (iii).

Algorithm The pseudo-code of IRLO is reported in Algorithm 1 (IRLO box). It receives two datasets $\mathcal{D}^E$ and $\mathcal{D}^b$ of trajectories collected by policies π^E and π^b, respectively,

Algorithm 1 IRLO and PIRLO.

Input : Datasets $\mathcal{D}^E = \{\langle s_h^{E,i}, a_h^{E,i} \rangle_h\}_i$, $\mathcal{D}^b = \{\langle s_h^{b,i}, a_h^{b,i} \rangle_h\}_i$
Output : Estimated sub- and super-feasible sets $\widehat{\mathcal{R}}^{\cap}, \widehat{\mathcal{R}}^{\cup}$

1 Estimate the expert's support:
$\widehat{\mathcal{S}}^{p,\pi^E} \leftarrow \{(s,h) \in \mathcal{S} \times [\![H]\!] \mid \exists i \in [\![\tau^E]\!] : s_h^{E,i} = s\}$
2 Estimate the expert's policy:
for $(s,h) \in \widehat{\mathcal{S}}^{p,\pi^E}$ **do**
3 $\quad \widehat{\pi}^E_h(s) \leftarrow a_h^{E,i}$ for some $i \in [\![\tau^E]\!]$ s.t. $s_h^i = s$
4 **end**
5 Estimate the state-action behavioral policy support:
$\widehat{\mathcal{Z}}^{p,\pi^b} \leftarrow \{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!] \mid \exists i \in [\![\tau^b]\!] : (s_h^{b,i}, a_h^{b,i}) = (s,a)\}$
6 Compute the counts for every $(s,a,h) \in \widehat{\mathcal{Z}}^{p,\pi^b}$ and $s' \in \mathcal{S}$:
$N_h^b(s,a,s') \leftarrow \sum_{i \in [\![\tau^b]\!]} \mathbb{1}\{(s_h^{b,i}, a_h^{b,i}, s_{h+1}^{b,i}) = (s,a,s')\}$
$N_h^b(s,a) \leftarrow \sum_{s' \in \mathcal{S}} N_h^b(s,a,s')$
7 Estimate the transition model:
for $(s,a,h) \in \widehat{\mathcal{Z}}^{p,\pi^b}$ **do**
8 $\quad$ **for** $s' \in \mathcal{S}$ **do**
9 $\quad\quad \widehat{p}_h(s'|s,a) \leftarrow \frac{N_h^b(s,a,s')}{\max\{1, N_h^b(s,a)\}}$
10 $\quad$ **end**
11 **end**

IRLO
Compute $\mathcal{R}^{\cap}_{\widehat{p},\widehat{\pi}^E}$ and $\mathcal{R}^{\cup}_{\widehat{p},\widehat{\pi}^E}$ with Definition 3.3
using $\widehat{p}$, $\widehat{\pi}^E$, $\widehat{\mathcal{Z}}^{p,\pi^b}$, and $\widehat{\mathcal{S}}^{p,\pi^E}$
12 **return** $(\mathcal{R}^{\cap}_{\widehat{p},\widehat{\pi}^E}, \mathcal{R}^{\cup}_{\widehat{p},\widehat{\pi}^E})$

PIRLO
Compute the confidence set $\mathcal{C}(\widehat{p}, b)$ via Eq. (7)
Compute $\widetilde{\mathcal{R}}^{\cap}_{\widehat{p},\widehat{\pi}^E}$ and $\widetilde{\mathcal{R}}^{\cup}_{\widehat{p},\widehat{\pi}^E}$ with Eq. (9)
using $\widehat{p}$, $\widehat{\pi}^E$, $\widehat{\mathcal{Z}}^{p,\pi^b}$, and $\widehat{\mathcal{S}}^{p,\pi^E}$
13 **return** $(\widetilde{\mathcal{R}}^{\cap}_{\widehat{p},\widehat{\pi}^E}, \widetilde{\mathcal{R}}^{\cup}_{\widehat{p},\widehat{\pi}^E})$

and it outputs the *estimated sub- and super-feasible sets* $\widehat{\mathcal{R}}^{\cap}$ and $\widehat{\mathcal{R}}^{\cup}$ as estimates of $\mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E}$, respectively. IRLO leverages $\mathcal{D}^E$ to compute the empirical estimates of the expert's support $\mathcal{S}^{p,\pi^E}$ and policy π^E, denoted by $\widehat{\mathcal{S}}^{p,\pi^E}$ and $\widehat{\pi}^E$ (lines 1-3), and it uses $\mathcal{D}^b$ to compute the empirical estimates of the behavioral policy support $\mathcal{Z}^{p,\pi^b}$, and of the transition model p, denoted by $\widehat{\mathcal{Z}}^{p,\pi^b}$ and $\widehat{p}$ (lines 5-9). Finally, it returns the sub- and super-feasible sets computed with the estimated supports, expert's policy, and transition model: $\widehat{\mathcal{R}}^{\cap} = \mathcal{R}^{\cap}_{\widehat{p},\widehat{\pi}^E}$ and $\widehat{\mathcal{R}}^{\cup} = \mathcal{R}^{\cup}_{\widehat{p},\widehat{\pi}^E}$ (line 12).

Computationally Efficient Implementation In Algorithm 1, IRLO outputs the estimated feasible sets $\widehat{\mathcal{R}}^{\cup}$ and $\widehat{\mathcal{R}}^{\cap}$ obtained by computing the intersection and the union of a continuous set of transition models (Definition 3.3). To show the computational efficiency of IRLO, we provide in Appendix G (Algorithm 2, IRLO box) a *polynomial-time membership checker* that tests whether a candidate reward function $r \in \mathfrak{R}$ belongs to $\widehat{\mathcal{R}}^{\cup}$ and/or $\widehat{\mathcal{R}}^{\cap}$. We apply *extended value iteration* (EVI, Auer et al., 2008) to compute an upper bound Q^+ and a lower bound Q^- of the Q-function induced by the candidate reward r and varying the transition model in a set $\mathcal{C}$. For the IRLO algorithm, $\mathcal{C}$ corresponds to

the equivalence class of the empirical estimate $\widehat{p}$ induced by the empirical support $\widehat{\mathcal{Z}}^{p,\pi^b}$, i.e., $[\widehat{p}]_{\equiv_{\widehat{\mathcal{Z}}^{p,\pi^b}}}$:

$$\mathcal{C} := \left\{ p' \in \mathcal{P} \,|\, \forall (s,a,h) \in \widehat{\mathcal{Z}}^{p,\pi^b} : p'_h(\cdot|s,a) = \widehat{p}_h(\cdot|s,a) \right\}. \quad (6)$$

The algorithm has a time complexity of order $\mathcal{O}(HS^2A)$.

Sample Complexity Analysis We now show that the **IRLO** algorithm is statistically efficient. The following theorem provides a *polynomial* upper bound to its sample complexity.

Theorem 5.1. *Let $\mathcal{M}$ be an MDP without reward and let π^E be the expert's policy. Let $\mathcal{D}^E$ and $\mathcal{D}^b$ be two datasets of τ^E and τ^b trajectories collected with policies π^E and π^b in $\mathcal{M}$, respectively. Under Assumption 2.1, **IRLO** is (ϵ,δ)-PAC for d-IRL with a sample complexity at most:*

$$\tau^b \leqslant \widetilde{\mathcal{O}}\left(\frac{H^3 Z^{p,\pi^b} \ln\frac{1}{\delta}}{\epsilon^2} \left(\ln\frac{1}{\delta} + S^{p,\pi^b}_{\max} \right) + \frac{\ln\frac{1}{\delta}}{\ln \frac{1}{1-\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min}}} \right),$$

$$\tau^E \leqslant \widetilde{\mathcal{O}}\left(\frac{\ln\frac{1}{\delta}}{\ln \frac{1}{1-\rho^{\pi^E,\mathcal{Z}^{p,\pi^E}}_{\min}}} \right).$$

Some comments are in order. First, we observe that the sample complexity for the expert's dataset τ^E is constant and depends on the minimum non-zero value of the visitation distribution $\rho^{\pi^E,\mathcal{Z}^{p,\pi^E}}_{\min} > 0$, but it does not depend on the desired accuracy ϵ. This accounts for the minimum number of samples to have $\widehat{\mathcal{S}}^{p,\pi^E} = \mathcal{S}^{p,\pi^E}$, with high probability. Second, the sample complexity for the behavioral policy dataset τ^b displays a tight dependence on the desired accuracy ϵ and a dependence of order H^4 on the horizon since in the worst case, $Z^{p,\pi^b} \leqslant SAH$. Moreover, we notice the two-regime behavior represented by $\ln(1/\delta) + S^{p,\pi^b}_{\max}$ (i.e., small and large δ) as in previous works (Kaufmann et al., 2021; Metelli et al., 2023). This term is multiplied by an additional $\ln(1/\delta)$ term, which always appears in offline (forward) RL (Xie et al., 2021) and it is needed to control the minimum number of samples collected from every reachable state-action pair. Finally, we observe a dependence analogous to that of τ^E on the minimum non-zero value of the visitation distribution $\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min} > 0$, to ensure that $\widehat{\mathcal{Z}}^{p,\pi^b} = \mathcal{Z}^{p,\pi^b}$. Note that when $\pi^b = \pi^E$, Assumption 2.1 is fulfilled, and the sample complexity reduces to:

$$\tau^E \leqslant \widetilde{\mathcal{O}}\left(\frac{H^3 S^{p,\pi^E} \ln\frac{1}{\delta}}{\epsilon^2} \left(\ln\frac{1}{\delta} + S^{p,\pi^E}_{\max} \right) + \frac{\ln\frac{1}{\delta}}{\ln \frac{1}{1-\rho^{\pi^E,\mathcal{Z}^{p,\pi^E}}_{\min}}} \right).$$

Since $S^{p,\pi^E} \leqslant SH$, the dependence on the number of actions is no longer present. An analogous result holds for d_∞.

Theorem 5.2. *Under the conditions of Theorem 5.1, **IRLO** is (ϵ,δ)-PAC for d_∞-IRL with a sample complexity at most:*

$$\tau^b \leqslant \widetilde{\mathcal{O}}\left(\frac{H^4 \ln\frac{1}{\delta}}{\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min} \epsilon^2} \left(\ln\frac{1}{\delta} + S^{p,\pi^b}_{\max} \right) + \frac{\ln\frac{1}{\delta}}{\ln \frac{1}{1-\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min}}} \right),$$

and τ^E is bounded as in Theorem 5.1.

We note that, since $1/\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min} \geqslant \mathcal{Z}^{p,\pi^b}$, Theorem 5.2 delivers a larger sample complexity w.r.t. Theorem 5.1. This is expected because of the relation $d(r,r') \leqslant 2 d_\infty(r,r')$ between the two semimetrics (see Proposition 4.1).

6. Pessimistic Inverse Reinforcement Learning for Offline data (PIRLO)

In this section, we present our main algorithm, **PIRLO** (Pessimistic Inverse Reinforcement Learning for Offline data). Beyond statistical and computational efficiency, **PIRLO** provides guarantees on the *inclusion monotonicity* of the proposed feasible sets by embedding a form of *pessimism*.[9]

Before presenting the algorithm, we formally introduce the notion of inclusion monotonicity and intuitively justify it. Thanks to the PAC property (Theorem 5.1), in the limit of *infinite samples* $\tau^b, \tau^E \to +\infty$, **IRLO** recovers exactly the sub- $\widehat{\mathcal{R}}^{\cap} \to \mathcal{R}^{\cap}_{p,\pi^E}$ and the super- $\widehat{\mathcal{R}}^{\cup} \to \mathcal{R}^{\cup}_{p,\pi^E}$ feasible sets, and, consequently, the property $\widehat{\mathcal{R}}^{\cap} \subseteq \mathcal{R}_{p,\pi^E} \subseteq \widehat{\mathcal{R}}^{\cup}$ holds. Because of the meaning of these sets, i.e., the *tightest learnable* subset $\mathcal{R}^{\cap}_{p,\pi^E}$ and superset $\mathcal{R}^{\cup}_{p,\pi^E}$ of the *feasible set* $\mathcal{R}_{p,\pi^E}$, it is desirable to ensure the property $\widehat{\mathcal{R}}^{\cap} \subseteq \mathcal{R}_{p,\pi^E} \subseteq \widehat{\mathcal{R}}^{\cup}$ (in high probability) in the *finite samples* regime $\tau^b, \tau^E \leqslant +\infty$ too. The following definition formalizes the property.

Definition 6.1 (Inclusion Monotonic Algorithm)**.** *Let $\delta \in (0,1)$. An algorithm $\mathfrak{A}$ outputting the estimated sub- and super-feasible sets $\widehat{\mathcal{R}}^{\cap}$ and $\widehat{\mathcal{R}}^{\cup}$ is δ-*inclusion monotonic *if:*

$$\mathbb{P}_{(p,\pi^E,\pi^b)}\left(\widehat{\mathcal{R}}^{\cap} \subseteq \mathcal{R}_{p,\pi^E} \subseteq \widehat{\mathcal{R}}^{\cup} \right) \geqslant 1-\delta.$$

Clearly, one can always choose $\widehat{\mathcal{R}}^{\cap} = \{\}$ and $\widehat{\mathcal{R}}^{\cup} = \mathfrak{R}$ to satisfy Definition 6.1. Thus, the inclusion monotonicity property will always be employed in combination with the PAC requirement (Definition 4.3). The importance of monotonicity will arise from a practical viewpoint in Section 7.

Algorithm The pseudocode of **PIRLO** is shown in Algorithm 1 (**PIRLO** box). The first part (lines 1-11) is analogous to **IRLO** and the main difference lies in the presence

[9] We remark on the substantial difference between our use of pessimism and that of Zhao et al. (2023). Indeed, we apply pessimism to *feasible sets* to ensure that the estimated set fulfills the *inclusion monotonicity* property, while Zhao et al. (2023) apply pessimism to ensure the *entry-wise monotonicity* of the reward function, i.e., $\widehat{r}(s,a) \preceq r(s,a)$, for all $\widehat{r} \in \widehat{\mathcal{R}}$ and $r \in \mathcal{R}$.

of the confidence set $\mathcal{C}(\widehat{p},b)\subseteq\mathcal{P}$ (line 13), containing the transition models in $\mathcal{P}$ close in ℓ_1-norm to the empirical estimate $\widehat{p}$, except the ones that are not compatible with expert's actions. Formally, $\mathcal{C}(\widehat{p},b)$ is defined as:[10]

$$\mathcal{C}(\widehat{p},b):=\Big\{p'\in\mathcal{P}\big|\ \forall(s,h)\in\widehat{\mathcal{S}}^{p,\pi^E},\, s'\notin\widehat{\mathcal{S}}^{p,\pi^E}_{h+1}:p'_h(s'|s,\widehat{\pi}^E_h(s))=0\quad\forall(s,a,h)\in\widehat{\mathcal{Z}}^{p,\pi^b}:\|p'_h(\cdot|s,a)-\widehat{p}_h(\cdot|s,a)\|_1\leqslant\widehat{b}_h(s,a)\Big\},\tag{7}$$

where $\widehat{b}_h(s,a)$ is defined in Equation (18). The intuition is that, with high probability, the true transition model p, and its equivalence class $[p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}$, will belong to $\mathcal{C}(\widehat{p},b)$.

Drawing inspiration from *pessimism* in RL, **PIRLO** "penalizes" the estimates of the feasible set by removing from $\widehat{\mathcal{R}}^{\cap}$ the rewards for which we are *not confident enough* of their membership to $\mathcal{R}^{\cap}_{p,\pi^E}$, and by adding to $\widehat{\mathcal{R}}^{\cup}$ the rewards for which we are *not confident enough* of their non-membership to $\mathcal{R}^{\cup}_{p,\pi^E}$, based on the confidence set $\mathcal{C}(\widehat{p},b)$ on the transition model. This translates into the following expressions:

$$\widehat{\mathcal{R}}^{\cap}=\bigcap_{p'\in\mathcal{C}(\widehat{p},b)}\mathcal{R}^{\cap}_{p',\widehat{\pi}^E},\qquad\widehat{\mathcal{R}}^{\cup}=\bigcup_{p'\in\mathcal{C}(\widehat{p},b)}\mathcal{R}^{\cup}_{p',\widehat{\pi}^E}.\tag{8}$$

This way, if $p\in\mathcal{C}(\widehat{p},b)$ and $\widehat{\pi}^E=\pi^E$ with high probability, we have that, simultaneously, $\widehat{\mathcal{R}}^{\cap}\subseteq\mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E}\subseteq\widehat{\mathcal{R}}^{\cup}$. This entails the inclusion monotonicity property (Definition 6.1) thanks to Definition 3.3.

Computationally Efficient Implementation Differently from **IRLO**, computing the set operations of Equation (8) cannot be directly carried out by EVI.[11] For this reason, we propose a *relaxation* which achieves the double objective of: (i) enabling a computationally efficient implementation of **PIRLO** (Algorithm 2, **PIRLO** box); and (ii) allowing for a simpler statistical analysis, preserving both the PAC and the inclusion monotonicity properties (details in Appendix G):

$$\widetilde{\mathcal{R}}^{\cap}:=\{r\in\mathfrak{R}|\forall\overline{\pi}\in[\widehat{\pi}^E]_{\equiv_{\widehat{\mathcal{S}}^{p,\pi^E}}},\forall(s,h)\in\widehat{\mathcal{S}}^{p,\pi^E},\forall a\in\mathcal{A}:\min_{p'\in\mathcal{C}(\widehat{p},b)}Q^{\widehat{\pi}^E}_h(s,\widehat{\pi}^E_h(s);p',r)\geqslant\max_{p''\in\mathcal{C}(\widehat{p},b)}Q^{\overline{\pi}}_h(s,a;p'',r)\},$$

$$\widetilde{\mathcal{R}}^{\cup}:=\{r\in\mathfrak{R}|\forall\overline{\pi}\in[\widehat{\pi}^E]_{\equiv_{\widehat{\mathcal{S}}^{p,\pi^E}}},\forall(s,h)\in\widehat{\mathcal{S}}^{p,\pi^E},\forall a\in\mathcal{A}:\max_{p'\in\mathcal{C}(\widehat{p},b)}Q^{\widehat{\pi}^E}_h(s,\widehat{\pi}^E_h(s);p',r)\geqslant\min_{p''\in\mathcal{C}(\widehat{p},b)}Q^{\overline{\pi}}_h(s,a;p'',r)\},$$

where the universal/existential quantification over the transition model of Definition 3.3 has been relaxed by the two $\max-\min$. In other words, we allow a choice of different transition models for the two Q-functions appearing in the two members of the inequality. Thus, $\widetilde{\mathcal{R}}^{\cap}\subseteq\widehat{\mathcal{R}}^{\cap}$ and $\widehat{\mathcal{R}}^{\cup}\subseteq\widetilde{\mathcal{R}}^{\cup}$, preserving the inclusion monotonicity. For the membership checking of a candidate reward $r\in\mathfrak{R}$, similarly to the **IRLO** case, we compute upper and lower bounds Q^+ and Q^- to the Q-function by using EVI varying the transition model in the confidence set $\mathcal{C}(\widehat{p},b)$ defined in Equation (7). Now, the confidence set is made of ℓ_1 *constraints* and the corresponding $\max$ and $\min$ programs can be solved by using the approach of (Auer et al., 2008, Figure 2). The overall time complexity is of order $\mathcal{O}(HS^2A\log S)$.

Sample Efficiency and Inclusion Monotonicity We now show that **PIRLO** is statistically efficient, with the additional guarantee (w.r.t. **IRLO**) of the inclusion monotonicity.

Theorem 6.1. *Let $\mathcal{M}$ be an MDP without reward and let π^E be the expert's policy. Let $\mathcal{D}^E$ and $\mathcal{D}^b$ be two datasets of τ^E and τ^b trajectories collected by executing policies π^E and π^b in $\mathcal{M}$. Under Assumption 2.1, **PIRLO** is (ϵ,δ)-PAC for d-IRL with a sample complexity at most:*

$$\tau^b\leqslant\widetilde{\mathcal{O}}\left(\frac{H^3Z^{p,\pi^b}\ln\frac{1}{\delta}}{\epsilon^2}\left(\ln\frac{1}{\delta}+S^{p,\pi^b}_{\max}\right)+\frac{H^6\ln\frac{1}{\delta}}{\rho^{\pi^b,\mathcal{Z}^{p,\pi^E}}_{\min}\epsilon^2}\left(\ln\frac{1}{\delta}+S^{p,\pi^b}_{\max}\right)+\frac{\ln\frac{1}{\delta}}{\ln\frac{1}{1-\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min}}}\right),$$

*and τ^E is bounded as in Theorem 5.1. Furthermore, **PIRLO** is inclusion monotonic.*

The price for the inclusion monotonicity is the additional term in the sample complexity which grows with H^6 and with $1/\rho^{\pi^b,\mathcal{Z}^{p,\pi^E}}_{\min}$. The latter represents the minimum non-zero visitation probability with which policy π^b covers $\mathcal{Z}^{p,\pi^E}$, i.e., the support of ρ^{p,π^E}. Intuitively, since the expert's policy is optimal, this additional term is due to a mismatch between optimal Q-functions under the different transition models of $\mathcal{C}(\widehat{p},b)$. Notice that, under Assumption 2.1, $\mathcal{Z}^{p,\pi^E}\subseteq\mathcal{Z}^{p,\pi^b}$, consequently, $\rho^{\pi^b,\mathcal{Z}^{p,\pi^E}}_{\min}\geqslant\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min}$. We can provide an analogous result for d_∞.

Theorem 6.2. *Under the conditions of Theorem 6.1, **PIRLO** is (ϵ,δ)-PAC for d_∞-IRL with a sample complexity at most:*

$$\tau^b\leqslant\widetilde{\mathcal{O}}\left(\frac{H^4\ln\frac{1}{\delta}}{\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min}\epsilon^2}\left(\ln\frac{1}{\delta}+S^{p,\pi^b}_{\max}\right)+\frac{H^6\ln\frac{1}{\delta}}{\rho^{\pi^b,\mathcal{Z}^{p,\pi^E}}_{\min}\epsilon^2}\left(\ln\frac{1}{\delta}+S^{p,\pi^b}_{\max}\right)+\frac{\ln\frac{1}{\delta}}{\ln\frac{1}{1-\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min}}}\right),$$

*and τ^E is bounded as in Theorem 5.1. Furthermore, **PIRLO** is inclusion monotonic.*

Notice that both bounds in Theorem 6.1 and Theorem 6.2 also hold for the objectives defined in Equation (8).[12]

[10] Actually, this definition does not take into account a corner case. See Appendix D.5 for details and a more precise definition.

[11] Membership testing can be here implemented with a *bilinear program*, which is, in general, a difficult problem (Appendix G).

[12] In Appendix F.5, we provide a tighter bound for the superset

7. Reward Sanity Check with PIRLO

In the literature, IRL algorithms (Ratliff et al., 2006; Ziebart et al., 2008) provide *criteria* to select a specific reward function from the feasible set. Our algorithm, PIRLO, thanks to the inclusion monotonicity property, provides a partition of the space of rewards $\mathfrak{R}$ in three sets: (i) rewards contained in the sub-feasible set $\widehat{\mathcal{R}}^{\cap}$ (i.e., feasible w.h.p.), (ii) rewards *not* contained in the super-feasible set $\mathfrak{R}\backslash\widehat{\mathcal{R}}^{\cup}$ (i.e., not feasible w.h.p.), and (iii) rewards that we cannot discriminate with the given confidence ($\widehat{\mathcal{R}}^{\cup}\backslash\widehat{\mathcal{R}}^{\cap}$). The situation is illustrated in Figure 1. Thus, PIRLO can be used both as a *sanity checker* on the rewards outputted by a specific IRL algorithm and for defining the set of rewards from which selecting one. To exemplify this application, we have run PIRLO using highway driving data from Likmeta et al. (2021) and some human-interpretable reward. We provide the experimental details and the results in Appendix K.

8. A Bitter Lesson

Up to now, we assumed to have two datasets $\mathcal{D}^E$ and $\mathcal{D}^b$ of trajectories collected by policies π^E and π^b, respectively. As already noted, this setting generalizes the most common IRL scenario where the only dataset $\mathcal{D}^E$ is collected by the deterministic expert's policy π^E and there is no possibility of collecting further data. A natural question arises: *Why not directly considering the setting with $\mathcal{D}^E$ only?* The reason lies in the following *negative result* showing that the reward functions that can be learned from a single expert's dataset $\mathcal{D}^E$ are not completely satisfactory.

Proposition 8.1. *Let $\mathcal{M}$ be the usual MDP without reward with $A \geqslant 2$ and let π^E be the deterministic expert's policy. Let $\mathcal{D}^E$ be a dataset of trajectories collected by following π^E in $\mathcal{M}$. Then, for any reward in $r \in \mathcal{R}^{\cap}_{p,\pi^E}$ it holds that:*

$$\forall (s,h) \in \mathcal{S}^{p,\pi^E}, \forall a \in \mathcal{A}: \quad r_h(s, \pi^E_h(s)) \geqslant r_h(s,a). \tag{9}$$

Thus, if we have no information about the transition model in non-expert's actions (as when we have $\mathcal{D}^E$ only), there exists no reward function r that simultaneously: (i) surely belongs to the sub-feasible set ($r \in \mathcal{R}^{\cap}_{p,\pi^E}$) and (ii) assigns to a non-expert's action a reward value greater than that assigned to the expert's action in the same (s,h) pair. This is clearly a property that is undesirable as it significantly limits the expressive power of the reward function, making IRL closer to behavioral cloning and, consequently, inheriting its limitations. As mentioned above, this issue can be overcome with a behavioral policy π^b that explores enough.

Proposition 8.2. *Under the conditions of Proposition 8.1, assume that $p_h(\cdot|s,a)$ is known, where $a \in \mathcal{A}$ is a non-expert's action in $(s,h) \in \mathcal{S}^{p,\pi^E}$. Then, if $p_h(\cdot|s,a) \neq p_h(\cdot|s,\pi^E_h(s))$, there exists a reward $r \in \mathcal{R}^{\cap}_{p,\pi^E}$ such that:*

$$r_h(s,\pi^E_h(s)) < r_h(s,a).$$

9. Conclusion

In this paper, we have introduced a novel notion of *feasible set* and an innovative *learning framework* for managing the intrinsic difficulties of the *offline* IRL setting. Furthermore, we have motivated the importance of *inclusion monotonicity*, and we have devised an original form of *pessimism* to achieve it. Then, we have presented two provably efficient algorithms, IRLO and PIRLO. We have shown that the latter provides guarantees of inclusion monotonicity and that it can be employed as a *reward sanity checker*. Finally, we have highlighted an *intrinsic limitation* of the offline IRL setting when samples from the experts are the only available.

Limitations and Future Works To understand whether our algorithms are *minimax optimal*, future works should focus on the derivation of sample complexity lower bounds for offline IRL. Moreover, it would be appealing to extend our framework to more challenging (non-tabular) environments.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

Funded by the European Union – Next Generation EU within the project NRPP M4C2, Investment 1.3 DD. 341 - 15 march 2022 – FAIR – Future Artificial Intelligence Research – Spoke 4 - PE00000013 - D53C22002380006.

References

Adams, S., Cody, T., and Beling, P. A. A survey of inverse reinforcement learning. *Artificial Intelligence Review*, 55: 4307–4346, 2022.

Arora, S. and Doshi, P. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2018.

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems 21 (NeurIPS)*, pp. 89–96, 2008.

Azar, M. G., Osband, I., and Munos, R. Minimax regret

$\widehat{\mathcal{R}}^{\cup}$ without using the relaxation. Moreover, in Appendix F.4, we prove a larger sample complexity upper bound, when including an additional useful requirement.

bounds for reinforcement learning. In *International Conference on Machine Learning 34 (ICML)*, pp. 263–272, 2017.

Boularias, A., Kober, J., and Peters, J. Relative entropy inverse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics 14 (AISTATS 2011)*, pp. 182–189, 2011.

Brafman, R. I. and Tennenholtz, M. R-Max - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3: 213–231, 2003.

Buckman, J., Gelada, C., and Bellemare, M. G. The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*, 2020.

Chatzigeorgiou, I. Bounds on the Lambert function and their application to the outage analysis of user cooperation. *IEEE Communications Letters*, 17:1505–1508, 2013.

Dann, C., Lattimore, T., and Brunskill, E. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 5717–5727, 2017.

Fagin, R. and Stockmeyer, L. Relaxing the triangle inequality in pattern matching. *International Journal of Computer Vision*, 30:219–231, 1998.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning 38 (ICML)*, pp. 5084–5096, 2021.

Kakade, S. M. and Langford, J. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning 19 (ICML)*, pp. 267–274, 2002.

Kaufmann, E., Menard, P., Domingues, O. D., Jonsson, A., Leurent, E., and Valko, M. Adaptive reward-free exploration. In *International Conference on Algorithmic Learning Theory 32 (ALT)*, pp. 865–891, 2021.

Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49:209–232, 2002.

Komanduru, A. and Honorio, J. A lower bound for the sample complexity of inverse reinforcement learning. In *International Conference on Machine Learning 38 (ICML)*, pp. 5676–5685, 2021.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pp. 1179–1191, 2020.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems, 2020.

Likmeta, A., Metelli, A. M., Ramponi, G., Tirinzoni, A., Giuliani, M., and Restelli, M. Dealing with multiple experts and non-stationarity in inverse reinforcement learning: an application to real-life problems. *Machine Learning*, 110(9):2541–2576, 2021.

Lindner, D., Krause, A., and Ramponi, G. Active exploration for inverse reinforcement learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pp. 5843–5853, 2022.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Provably good batch off-policy reinforcement learning without great exploration. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pp. 1264–1274, 2020.

Metelli, A. M., Ramponi, G., Concetti, A., and Restelli, M. Provably efficient learning of transferable rewards. In *International Conference on Machine Learning 38 (ICML)*, pp. 7665–7676, 2021.

Metelli, A. M., Lazzati, F., and Restelli, M. Towards theoretical understanding of inverse reinforcement learning. In *International Conference on Machine Learning 40 (ICML)*, pp. 24555–24591, 2023.

Munos, R. Performance bounds in ℓ_p-norm for approximate value iteration. *SIAM Journal on Control and Optimization*, 46:541–561, 2007.

Ng, A. Y. and Russell, S. J. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning 17 (ICML)*, pp. 663–670, 2000.

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., and Peters, J. An algorithmic perspective on imitation learning. *Foundations and Trends®in Robotics*, 7:1–179, 2018.

Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

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

Ratliff, N. D., Bagnell, J. A., and Zinkevich, M. A. Maximum margin planning. In *International Conference on Machine Learning 23 (ICML)*, pp. 729—736, 2006.

Rockafellar, R. T. and Wets, R. J. B. *Variational Analysis*. Springer Verlag, 1998.

Russell, S. Learning agents for uncertain environments (extended abstract). In *Conference on Computational Learning Theory 11 (COLT)*, pp. 101–103, 1998.

Steele, J. M. *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. Cambridge University Press, 2004.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. A Bradford Book, 2018.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy Finetuning: Bridging sample-efficient offline and online reinforcement learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pp. 27395–27407, 2021.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. MOPO: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pp. 14129–14142, 2020.

Yue, S., Wang, G., Shao, W., Zhang, Z., Lin, S., Ren, J., and Zhang, J. Clare: Conservative model-based reward learning for offline inverse reinforcement learning. In *International Conference on Learning Representations 11 (ICLR)*, 2023.

Zeng, S., Li, C., Garcia, A., and Hong, M. When demonstrations meet generative world models: A maximum likelihood framework for offline inverse reinforcement learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, volume 36, pp. 65531–65565, 2023.

Zhao, L., Wang, M., and Bai, Y. Is inverse reinforcement learning harder than standard reinforcement learning? *arXiv preprint arXiv:2312.00054*, 2023.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence 23 (AAAI)*, pp. 1433–1438, 2008.

	(Metelli et al., 2021)	(Metelli et al., 2023)	(Lindner et al., 2022)	(Zhao et al., 2023)	**Ours**
Setting	online generative model	online generative model	online forward model	offline with behavioral policy	offline with behavioral policy
Solution concepts	$\overline{\mathcal{R}}_{p,\pi^E}$	$\overline{\mathcal{R}}_{p,\pi^E}$	$\overline{\mathcal{R}}_{p,\pi^E}$	$\mathscr{R} \approx \overline{\mathcal{R}}_{p,\pi^E}$	$\mathcal{R}_{p,\pi^E}$
Monotonicity	No	No	No	$\widehat{r} \preccurlyeq r$	$\mathcal{R}^{\cap}_{p,\pi^E} \subseteq \mathcal{R}_{p,\pi^E} \subseteq \mathcal{R}^{\cup}_{p,\pi^E}$
Reward distance $r, \widehat{r}$	$\|Q^*(p,r) - Q^*(p,\widehat{r})\|_\infty$	$\|r - \widehat{r}\|_\infty$	$\sup_{\widehat{\pi}*} \|Q^*(p,r) - Q^{\widehat{\pi}^*}(p,r)\|_\infty$	$\sup_{h \in [\![H]\!]} \mathbb{E}_{p,\overline{\pi}} \lvert V^{\overline{\pi}}(p,r) - V^{\overline{\pi}}(p,\widehat{r})\rvert$	$\sum_{h \in [\![H]\!]} \mathbb{E}_{p,\pi^b} \lvert r - \widehat{r}\rvert$
Feasible set distance	$\mathcal{H}$	$\mathcal{H}$	$\mathcal{H}$	$\sup_{V,A}$	$\mathcal{H}$
Lower bound?	No	Yes	No	Yes	No
Matching upper bound?	?	Yes	?	?	?
Trick	$\|Q\|_\infty \leqslant H$	$r = r'/(1+\epsilon)$	$\|Q\|_\infty \leqslant H$	$r = r(V,A)$	$d(r,\widehat{r}) = d'(r,\widehat{r}) / \max\{\|r\|_\infty, \|\widehat{r}\|_\infty\}$
Recover exact objective?	Yes	Yes	Yes	No	Yes

Table 1. Comparison of our paper with the main related works. *Setting* refers to whether the work analyses the online generative model, the online forward model, or the offline setting. To be precise, Lindner et al. (2022) provides some insights also for the online generative model, and Zhao et al. (2023) analyses also the online forward model in their setting. However, to keep the table clear and concise, we avoid inserting them. *Solution concepts* indicates the learning objective. By $\mathscr{R}$ we denote the concept of reward mapping, i.e., parametrization of the reward function using V, A, introduced by Zhao et al. (2023); in practice, it is equivalent to the "old" notion of feasible set $\overline{\mathcal{R}}_{p,\pi^E}$, which is more suitable to the online setting. *Monotonicity* specifies whether the work adopts a notion of monotonicity according to some partial order. By $\preccurlyeq$ we mean the entrywise order among vectors. We stress that we are the first to devise the concept of *inclusion monotonicity*, which permits applications such as the sanity checker. *Reward distance* refers to the distance adopted in the space of rewards $\mathfrak{R}$. We denote $\|\cdot\|_\infty := \max_{s,a,h} |\cdot|$ and we neglect the dependence on s, a, h for simplicity. While the works about the online setting can afford to control the error with an ℓ_∞ distance, for the offline setting an ℓ_1 objective is more suitable. It should be remarked that, while Zhao et al. (2023) compare the induced V-functions for a single fixed policy $\overline{\pi}$, and so they are not able to recover the exact objective, we compare directly the distance between rewards, and we do not suffer from this issue. *Feasible set distance* indicates the distance among sets. All the works adopt the Hausdorff distance $\mathcal{H}$ except for Zhao et al. (2023), which considers the worst parameters V, A; from a technical perspective, this is equivalent to $\mathcal{H}$. Metelli et al. (2023) provide a matching *lower bound*, while Zhao et al. (2023) computes a lower bound that does not depend on the confidence δ, and so, it is not tight. By *trick*, we mean the solution adopted to cope with the intrinsic unboundedness of the space of rewards. Observe that the constraint on the Q-function $\|Q\|_\infty \leqslant H$ imposes strange non-box constraints on the rewards; the normalization by $1+\epsilon$ proposed by Metelli et al. (2023) (see their Lemma B.1) is "hard-coded" for their ℓ_∞ objective and does not generalize to other notions of distance. Parametrizing the rewards by the bounded pair V-function, advantage function V, A, thus defining a reward mapping $\mathscr{R}$ as in (Zhao et al., 2023), is, in practice, equivalent to the constraint on the Q-function $\|Q\|_\infty \leqslant H$ and, thus, it does not solve the problem. Instead, notice that our normalization trick is an effective way to cope with this issue, and it can be applied to the settings of all other works, simplifying the sample complexity analysis. Finally, by *recover exact objective* we mean whether the setting analyzed is compatible with the solution concept adopted, i.e., if the solution concept can be retrieved exactly in that setting. Notice that this is not true for Zhao et al. (2023), which cannot improve beyond an ϵ-estimate due to the partial coverage of the data and the monotonicity notion that they adopt. Instead, our novel definition of feasible set $\mathcal{R}_{p,\pi^E}$ and our notion of inclusion monotonicity overcome this issue.

A. Related Works

Related works can be distinguished in theoretical IRL works and works about Reinforcement Learning (RL) in the *offline* setting. Since the former group of papers is more closely connected to the subjects of this work, we focus on it here, and we refer to Appendix A for a presentation of the remaining literature.

The notion of *feasible set* has been introduced implicitly by Ng & Russell (2000). More recently, Metelli et al. (2021) build upon previous works to define the *feasible set* explicitly. They propose two algorithms for the estimation of *feasible set* in the *online* setting with generative model. Moreover, the authors analyze the sample complexity of the algorithms and prove the first upper bound to the number of samples required for the estimation of the *feasible set* in the discounted infinite-horizon setting. Such bound is in the order of $\widetilde{\mathcal{O}}\big(\frac{SA}{(1-\gamma)^4\epsilon^2}\big)$, where S and A denote, respectively, the cardinality of the state and action spaces, γ is the discount factor and ϵ is the accuracy, measured as distance in max norm between the induced Q-functions. Later, Lindner et al. (2022) proposes an algorithm, named AceIRL, to estimate the *feasible set* in the *online* setting with forward model. The result is an upper bound of $\widetilde{\mathcal{O}}\big(\frac{H^5SA}{\epsilon^2}\big)$ to the number of trajectories required in the episodic finite-horizon setting. The first lower bound to the sample complexity of IRL has been devised by Komanduru & Honorio (2021). However, their setting concerns state-only rewards in tabular Markov Decision Processes (MDPs) with only two actions, resulting in a lower bound in the order of $\Omega(S \ln S)$. Metelli et al. (2023) analyzes the *online* setting with generative model by measuring the accuracy using the max norm directly between rewards. By adopting two different constructions for the hard instances, it proves a lower bound, along with a matching upper bound, for the sample complexity of estimating the *feasible set*. The number of samples is in the order of $\widetilde{\Theta}\big(\frac{H^3SA}{\epsilon^2}(\ln\frac{1}{\delta}+S)\big)$, where δ is the confidence. It is worth mentioning the work of Zhao et al. (2023), which analyze the sample complexity of estimating the *reward mapping*, a concept analogous to that of *feasible set*, in the context of offline IRL. They propose algorithms for solving the problem in both the *offline* and *online* settings, and analyze the sample complexity, obtaining an upper bound of $\widetilde{\mathcal{O}}\big(\frac{H^4S^2C^*}{\epsilon^2}\big)$ for the *offline* setting, where C^* is a concentrability coefficient. However, Zhao et al. (2023) adopts a solution concept which is intrinsically connected to the coverage of the state-action-stage space, which is a strong requirement in the *offline* context. As a consequence, the entrywise reward-based pessimistic approach proposed by the authors is not able to recover the solution concept exactly. We also mention (Yue et al., 2023) and (Zeng et al., 2023) as two additional IRL works that adopt pessimism but with different settings than ours.

For a clear comparison of these works, see Table 1.

Additional Related Works It is worth mentioning also the works that focus on (forward) RL in the *offline* setting, because they share with our topic some important concepts and technical tools. We provide a brief overview in this section.

The principle of *optimism* in the face of uncertainty is a well-established tool for favoring exploration in the context of *online* bandits (Lattimore & Szepesvári, 2020) and *online* (forward) RL (Kearns & Singh, 2002; Brafman & Tennenholtz, 2003; Azar et al., 2017; Dann et al., 2017). However, in the *offline* setting, the learning agent is given a batch dataset and is not allowed to interact with the environment, thus the adoption of *optimism* does not improve the performances. Moreover, one of the biggest challenges of the *offline* RL setting is that the given dataset might suffer from an insufficient coverage of the space (Levine et al., 2020). To improve the performances of algorithms for solving the *offline* policy optimization problem, the commonly adopted mechanism is the *pessimism* principle: "*Behave as though the world was plausibly worse than you observed it to be*" (Buckman et al., 2020). As opposed to the principle of *optimism* in the face of uncertainty which favors exploration, the *pessimism* principle favors exploitation. This tool has been adopted in a variety of works for devising algorithms to solve the *offline* policy optimization problem (Yu et al., 2020; Kumar et al., 2020; Liu et al., 2020). From a theoretical perspective, Buckman et al. (2020) proposes a unified framework for the study of this kind of algorithms, revealing the reasons why the *pessimism* principle can demonstrate good performance even when the dataset is not informative of every policy. Moreover, Jin et al. (2021) proposes a *pessimistic* variant of the value iteration algorithm, named PEVI, and it shows that the *pessimism* principle is not only provably efficient, but also minimax optimal. Another line of research more closely related to the *offline* IRL setting is that introduced by Rashidinejad et al. (2021). They analyse the *offline* policy optimization problem in the novel setting in which the composition of the batch dataset can be located at any point in the range between *expert data* and *uniform coverage data*. *Expert data* means that the dataset has been collected by an expert, and the problem reduces to the imitation learning problem (Osa et al., 2018), while *uniform coverage data* refers to a dataset that guarantees a uniform coverage of the space, which is a common setting in which it is usually required the existence of a uniformly bounded concentrability coefficient (Munos, 2007). Specifically, Rashidinejad et al. (2021) proposes a new framework that smoothly interpolates between the two extremes of data composition, and analyses the

information-theoretic limits of LCB, the algorithm they propose, in three different settings. Finally, Xie et al. (2021) builds upon Rashidinejad et al. (2021) and devises *policy finetuning*, a framework that interpolates between online and offline RL. Remarkably, Xie et al. (2021) designs a novel algorithm, PEVI-Adv, which achieves a sample complexity of at most $\widetilde{\mathcal{O}}\left(\frac{H^3SC^*}{\epsilon^2}\right)$ episodes in the finite-horizon setting, where C^* is the single-policy concentrability coefficient. Also notice that the authors prove a matching lower bound.

B. A Framework for the "old" Feasible Set

In this section, we apply the framework we presented in Section 3 to the "old" notion of *feasible set* (Definition 3.1). In addition, we present a rather negative result on the kind of reward functions contained in the subset of $\overline{\mathcal{R}}_{p,\pi^E}$.

Let us begin by adapting Definition 3.3 to $\overline{\mathcal{R}}_{p,\pi^E}$. Recall that we use $\mathcal{D}^E$ to estimate π^E, and $\mathcal{D}^b$ to estimate p.

Definition B.1 (Subset and Superset of $\overline{\mathcal{R}}_{p,\pi^E}$)**.** *Let $\mathcal{M}=\langle \mathcal{S},\mathcal{A},p,\mu_0,H\rangle$ be an MDP without reward and let π^E be the deterministic expert's policy and π^b the behavioral policy. Then, we define the* subset $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$ *and the* superset $\overline{\mathcal{R}}^{\cup}_{p,\pi^E}$ *of the feasible set $\overline{\mathcal{R}}_{p,\pi^E}$ as:*

$$\overline{\mathcal{R}}^{\cap}_{p,\pi^E} := \bigcap_{p'\in[p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}} \bigcap_{\pi'\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}} \overline{\mathcal{R}}_{p',\pi'},$$
$$\overline{\mathcal{R}}^{\cup}_{p,\pi^E} := \bigcup_{p'\in[p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}} \bigcup_{\pi'\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}} \overline{\mathcal{R}}_{p',\pi'}.$$

Clearly, since $p\in[p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}$ and $\pi^E\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}$, it holds that $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}\subseteq\overline{\mathcal{R}}_{p,\pi^E}\subseteq\overline{\mathcal{R}}^{\cup}_{p,\pi^E}$. Also notice that when $\mathcal{Z}^{p,\pi^b}=\mathcal{S}\times\mathcal{A}\times[\![H]\!]$ (and so, because of Assumption 2.1, $\mathcal{S}^{p,\pi^E}=\mathcal{S}\times[\![H]\!]$), then $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}=\overline{\mathcal{R}}_{p,\pi^E}=\overline{\mathcal{R}}^{\cup}_{p,\pi^E}$. The intuition underlying the definition is analogous to that for $\mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E}$.

The following theorem shows that $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$ and $\overline{\mathcal{R}}^{\cup}_{p,\pi^E}$ are "well-defined".

Theorem (Informal) B.1. *Let $\mathcal{M}=\langle \mathcal{S},\mathcal{A},p,\mu_0,H\rangle$ be an MDP without reward and let $\mathcal{D}^E$ and $\mathcal{D}^b$ be two datasets of trajectories collected by policies π^E and π^b. Then, subset $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$ and superset $\overline{\mathcal{R}}^{\cup}_{p,\pi^E}$ are the* tightest learnable *subset and superset of $\overline{\mathcal{R}}_{p,\pi^E}$ from $\mathcal{D}^E$ and $\mathcal{D}^b$.*

In Appendix C, we enunciate this result formally and we provide a proof.

The following theorem shows a negative result on the kind of reward functions contained in $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$ under reasonable conditions. The intuition is that, since $\overline{\mathcal{R}}_{p,\pi^E}$ requires the knowledge of the optimal (expert's) action in the entire $\mathcal{S}\times[\![H]\!]$, then if there are pairs $(s,h)\in\mathcal{S}\times[\![H]\!]$ in which we cannot have this information, then we are forced to make all actions optimal there (this is exactly what the intersection over policies makes in $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$). This imposes many constraints on the structure of the rewards, resulting in the following theorem. It should be remarked that, if we had only dataset $\mathcal{D}^b$ and if Assumption 2.1 was violated, then we might not know the expert's action in some $(s,h)\in\mathcal{S}^{p,\pi^E}$, and a similar result would also hold for $\mathcal{R}_{p,\pi^E}$; however, that would be a non-realistic setting for IRL.

Theorem B.2. *Let $\mathcal{M}$ be an MDP without reward and let π^E be the expert's policy and π^b be the behavioral policy. If for any stage $h\in[\![H]\!]$ there is at least a state, say s_h, for which $(s_h,h)\notin\mathcal{S}^{p,\pi^b}_h$, then $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$ is made of "almost-constant" rewards. Formally: $r\in\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$ if and only if there exists a sequence $\{k_h\}_h$ of H real numbers and a set $\{\bar{r}_s\}_s$ with as many real numbers as the cardinality of the support of μ_0, such that, for any $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]$:*

$$\begin{cases} r_h(s,a)=x(h,s) & \text{if } (s,h)\in\mathcal{S}^{p,\pi^b}_h \wedge a=\pi^E_h(s) \\ r_h(s,a)\leqslant x(h,s) & \text{if } (s,h)\in\mathcal{S}^{p,\pi^b}_h \wedge a\neq\pi^E_h(s) \\ r_h(s,a)=k_h & \text{if } (s,h)\notin\mathcal{S}^{p,\pi^b}_h \end{cases},$$

where $x(h,s):=\bar{r}_s$ if $h=1$, and $x(h,s):=k_h$ otherwise.

This theorem states that, under reasonable conditions, i.e., if at every stage h there is at least one state not reached by the behavioral policy π^b, then all the rewards of the subset $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$ have a trivial form, which, i.a., does not depend on the specific

value of the transition model, but only on its "support" $\mathcal{S}^{p,\pi^b}$. In practice, this means that $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$ does not represent an interesting target for learning.

Proof of Theorem B.2. The theorem expresses a necessary and sufficient condition, thus two proofs are required. In the following, recall that $\mathcal{S}^{p,\pi^b} \supseteq \mathcal{S}^{p,\pi^E}$ because of Assumption 2.1, and so the existence of $(s_h,h) \notin \mathcal{S}^{p,\pi^b}_h$ for all $h \in [\![H]\!]$ entails the existence of $(s_h,h) \notin \mathcal{S}^{p,\pi^E}_h$ for all $h \in [\![H]\!]$. Moreover, we consider the representation of $\overline{\mathcal{R}}_{p,\pi^E}$ as provided in Eq. 13. We begin with the proof of the sufficiency.

The proof of the sufficiency proceeds in four steps. First, we show that, outside $\mathcal{S}^{p,\pi^E}$, the definition of $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$ enforces all actions to be optimal. Then, we use this condition to show that, for any $h \geqslant 2$, irrespective of the transition model, all states take on the same value function value $\{\bar{V}_h\}_h$, and that all the actions in (s,h) outside $\mathcal{S}^{p,\pi^E}$ have the same reward value $\{k_h\}_h$. The subsequent steps build upon these findings to show that expert's actions take on constant reward value $\{k_h\}_h$ for $h \geqslant 2$, and that non-expert's actions are smaller than the corresponding expert's action.

Let us begin by showing that, outside $\mathcal{S}^{p,\pi^E}$, all actions shall be optimal. By definition of $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$, we have:

$$
\begin{aligned}
\overline{\mathcal{R}}^{\cap}_{p,\pi^E} &:= \bigcap_{p' \in [p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}} \bigcap_{\pi' \in [\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}} \overline{\mathcal{R}}_{p',\pi'} \\
&= \{r \in \mathfrak{R} \,|\, \forall p' \in [p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}, \forall \pi' \in [\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}, \forall (s,h) \in \mathcal{S} \times [\![H]\!] : \mathop{\mathbb{E}}_{a \sim \pi'_h(\cdot|s)} Q^*_h(s,a;p',r) = \max_{a' \in \mathcal{A}} Q^*_h(s,a';p',r)\} \\
&= \{r \in \mathfrak{R} \,|\, \forall p' \in [p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}, \forall (s,h) \in \mathcal{S} \times [\![H]\!], \forall \pi' \in [\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}} : \mathop{\mathbb{E}}_{a \sim \pi'_h(\cdot|s)} Q^*_h(s,a;p',r) = \max_{a' \in \mathcal{A}} Q^*_h(s,a';p',r)\} \\
&\overset{(1)}{=} \{r \in \mathfrak{R} \,|\, \forall p' \in [p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}} : \big(\forall (s,h) \in \mathcal{S}^{p,\pi^E} : Q^*_h(s,\pi^E_h(s);p',r) = \max_{a \in \mathcal{A}} Q^*_h(s,a;p',r) \wedge \\
&\qquad\qquad \forall (s,h) \notin \mathcal{S}^{p,\pi^E} : \forall \pi'_h(\cdot|s) \in \Delta^{\mathcal{A}} : \mathop{\mathbb{E}}_{a \sim \pi'_h(\cdot|s)} Q^*_h(s,a;p',r) = \max_{a' \in \mathcal{A}} Q^*_h(s,a';p',r)\big)\} \\
&= \{r \in \mathfrak{R} \,|\, \forall p' \in [p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}} : \big(\forall (s,h) \in \mathcal{S}^{p,\pi^E} : Q^*_h(s,\pi^E_h(s);p',r) = \max_{a \in \mathcal{A}} Q^*_h(s,a;p',r) \wedge \\
&\qquad\qquad \forall (s,h) \notin \mathcal{S}^{p,\pi^E} : \forall a \in \mathcal{A} : Q^*_h(s,a;p',r) = \max_{a' \in \mathcal{A}} Q^*_h(s,a';p',r)\big)\},
\end{aligned}
$$

where, at (1), we have partitioned $\mathcal{S} \times [\![H]\!]$ using $\mathcal{S}^{p,\pi^E}$ and we have applied the definition of $\equiv_{\mathcal{S}^{p,\pi^E}}$. This shows that, outside $\mathcal{S}^{p,\pi^E}$, all actions must be optimal.

Now, we show that, for any reward $r \in \overline{\mathcal{R}}^{\cap}_{p,\pi^E}$, there exists a sequence of value functions[13] $\{\bar{V}_h\}_{h \in [\![2,H]\!]}$ induced by r which does not depend neither on the transition model nor on the state considered: $V^*_h(s;p',r) = \bar{V}_h$ for any $s \in \mathcal{S}, h \in [\![2,H]\!]$, and for any p' of $[p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}$. So, let r be a reward of $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$ and let (s,h) be a pair not in $\mathcal{S}^{p,\pi^b} \supseteq \mathcal{S}^{p,\pi^E}$, for $h \in [\![H-1]\!]$. Notice that the existence of pair (s,h) is guaranteed by hypothesis. For what we have seen at the previous step, for any pair of actions a_1, a_2, for any $p' \in [p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}$, it holds that $Q^*_h(s,a_1;p',r) = Q^*_h(s,a_2;p',r)$. Through the Bellman's equation, we can write:

$$
r_h(s,a_1) + \mathop{\mathbb{E}}_{s' \sim p'_h(\cdot|s,a_1)} [V^*_{h+1}(s';p',r)] = r_h(s,a_2) + \mathop{\mathbb{E}}_{s' \sim p'_h(\cdot|s,a_2)} [V^*_{h+1}(s';p',r)]. \tag{10}
$$

Let $s_1, s_2 \in \mathcal{S}$. By definition of $\equiv_{\mathcal{Z}^{p,\pi^b}}$, this condition must hold for any $p'_h(\cdot|s,a_1) \in \Delta^{\mathcal{S}}$ and $p'_h(\cdot|s,a_2) \in \Delta^{\mathcal{S}}$. In particular, let us take two transition models $p^1, p^2 \in [p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}$ such that :

$$
p^1 = \begin{cases} p^1_h(s_1|s,a_1) = 1 \\ p^1_h(s_1|s,a_2) = 1 \end{cases},
$$

$$
p^2 = \begin{cases} p^2_h(s_1|s,a_1) = 1 \\ p^2_h(s_1|s,a_2) = 0 \end{cases}.
$$

[13] The value function of policy π at (s,h) under transition model p and reward r is defined as: $V^\pi_h(s;p,r) := \sum_{a \sim \pi_h(\cdot|s)} Q^\pi_h(s,a;p,r)$. The optimal value function is defined as: $V^*_h(s;p,r) := \max_{a \in \mathcal{A}} Q^*_h(s,a;p,r)$.

Inserting p^1 into Eq. 10, we get:

$$r_h(s,a_1)+V^*_{h+1}(s_1;p^1,r)=r_h(s,a_2)+V^*_{h+1}(s_1,p^1,r)\Longrightarrow r_h(s,a_1)=r_h(s,a_2).$$

Because a_1 and a_2 are arbitrary, this holds for any action $a\in\mathcal{A}$ in $(s,h)\notin\mathcal{S}^{p,\pi^b}$. Therefore, we have shown that (the condition at H is trivial since $V^*_{H+1}(s;p',r)=0$):

$$r\in\overline{\mathcal{R}}^{\cap}_{p,\pi^E}\Longrightarrow\exists\{k_h\}_{h\in[\![H]\!]}:\forall(s,h)\notin\mathcal{S}^{p,\pi^b},\forall a\in\mathcal{A}:r_h(s,a)=k_h.$$

Inserting p^2 into Eq. 10, we obtain:

$$r_h(s,a_1)+V^*_{h+1}(s_1;p^2,r)=r_h(s,a_2)+V^*_{h+1}(s_2;p^2,r)\Longrightarrow V^*_{h+1}(s_1;p^2,r)=V^*_{h+1}(s_2;p^2,r),$$

since $r_h(s,a_1)=r_h(s,a_2)$. Because p^2 (and so the next state s_1) can be chosen arbitrarily in $h+1\geqslant 2$, then we have proved that:

$$r\in\overline{\mathcal{R}}^{\cap}_{p,\pi^E}\Longrightarrow\exists\{\bar{V}_h\}_{h\in[\![2,H]\!]}:\forall p'\in[p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}},\forall(s,h)\in\mathcal{S}\times[\![H]\!]:V^*_h(s;p',r)=\bar{V}_h.$$

In a similar manner, we can prove the same result also for $(s,h)\in\mathcal{S}^{p,\pi^b}\setminus\mathcal{S}^{p,\pi^E}$.

The next step of the proof consists in showing that, for any $h\in[\![2,H]\!]$, for any $s\in\mathcal{S}^{p,\pi^E}_h$, the reward value assigned to expert's action coincides with k_h. Let $(s,h)\in\mathcal{S}^{p,\pi^E}$ with $h\geqslant 2$. For any $p'\in[p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}$, it holds:

$$\begin{aligned}\bar{V}_h=V^*_h(s;p',r)=Q^*_h(s,a^E;p',r)&=r_h(s,a^E)+\mathop{\mathbb{E}}_{s'\sim p'_h(\cdot|s,a^E)}[V^*_{h+1}(s';p',r)]\\&=r_h(s,a^E)+\mathop{\mathbb{E}}_{s'\sim p'_h(\cdot|s,a^E)}[\bar{V}_{h+1}]\\&=r_h(s,a^E)+\bar{V}_{h+1}.\end{aligned}\tag{11}$$

By hypothesis, there exists $s'\notin\mathcal{S}^{p,\pi^b}_h$ such that, for any $a\in\mathcal{A}$:

$$\begin{aligned}\bar{V}_h=V^*_h(s';p',r)=Q^*_h(s',a;p',r)&=r_h(s',a)+\mathop{\mathbb{E}}_{s''\sim p'_h(\cdot|s',a)}[V^*_{h+1}(s'';p',r)]\\&=k_h+\mathop{\mathbb{E}}_{s'\sim p'_h(\cdot|s',a)}[\bar{V}_{h+1}]\\&=k_h+\bar{V}_{h+1}.\end{aligned}\tag{12}$$

Comparing Eq. 11 and Eq. 12, we infer that $r_h(s,a)=k_h$.

With a similar reasoning, we can prove that the reward value of non-expert's action shall be at most k_h. For simplicity, set $\bar{V}_{H+1}=0$. Let (s,h) be any pair in $\mathcal{S}^{p,\pi^E}$ and let a be any non-expert's action. Then, for any $p'\in[p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}$, we have:

$$\begin{aligned}Q^*_h(s,a;p',r)&=r_h(s,a)+\mathop{\mathbb{E}}_{s'\sim p'_h(\cdot|s,a)}[V^*_{h+1}(s';p',r)]\\&=r_h(s,a)+\bar{V}_{h+1}\\&\leqslant V^*_h(s;p',r)\\&=r_h(s,\pi^E_h(s))+\bar{V}_{h+1}\\&=k_h+\bar{V}_{h+1}.\end{aligned}$$

From which it follows that $r_h(s,a)\leqslant k_h$. This concludes the proof of the sufficiency.

With regards to the necessity, we have to show that any reward r that can be expressed as in the statement of the theorem belongs to $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$. It is easy to notice that, irrespective of the transition model p', the optimal value function of any state $s\in\mathcal{S}$ at stage $h\geqslant 2$ is $V^*_h(s;p',r)=\sum_{i=h}^{H}k_i$, and that it is achieved by playing the expert's policy. At step $h=1$, since all next states take on the same optimal value function, the result is immediate. □

C. On the Learnability of the Feasible Set

In this section, we formalize the notion of (PAC-)*learnability* and we analyze the learnability properties of the various objects that we introduced in Section 3 (and in Appendix B). Specifically, we show that both the definitions of *feasible set* $\overline{\mathcal{R}}_{p,\pi^E}$ and $\mathcal{R}_{p,\pi^E}$ are not *learnable* in the setting of Section 3 unless the behavioral policy covers the entire $\mathcal{S}\times\mathcal{A}\times[\![H]\!]$ space. Next, we demonstrate that the framework we have introduced cannot be improved; simply put, we show that $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$ and $\overline{\mathcal{R}}^{\cup}_{p,\pi^E}$ are the *tightest learnable* bounds of $\overline{\mathcal{R}}_{p,\pi^E}$ according to partial order $\subseteq$, and that $\mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E}$ are the *tightest learnable* bounds of $\mathcal{R}_{p,\pi^E}$.

We give a definition of *learnability* in the context of the Probably Approximately Correct (PAC) framework since our main focus is on PAC bounds to the sample complexity in this work. Let $\phi\in\Phi$ be our target of learning, i.e., the quantity that we aim to estimate, and let μ be a certain distribution that provides us with N independent samples $X_1,X_2,\dots,X_N\sim\mu$. Intuitively, ϕ can be *learned* from μ if there exists a procedure able to use the samples $X_1,X_2,\dots,X_N$ of μ to create a "good-enough" estimate of ϕ, where the "goodness" is measured by a meaningful notion of distance. We formalize the intuition in the following definition.

Definition C.1 (PAC-learnability)**.** *A quantity $\phi\in\Phi$ is PAC-learnable from a distribution μ if there exists a semimetric d in Φ, that satisfies a ρ-relaxed triangle inequality with finite ρ, and an algorithm $\mathfrak{A}$ such that, for any $\epsilon,\delta\in(0,1)$, there exists a finite $N\in\mathbb{N}$ for which:*

$$\mathbb{P}_\mu\big(d(\phi,\widehat{\phi}\leqslant\epsilon)\big)\geqslant 1-\delta,$$

where $\widehat{\phi}\in\Phi$ is the estimate of ϕ computed by $\mathfrak{A}$ using at least N samples, and $\mathbb{P}_\mu$ is the probability measure induced by μ.

Simply put, ϕ is *PAC-learnable* if the samples from μ leak "enough" information about ϕ. Notice that any metric satisfies a ρ-relaxed triangle inequality with $\rho=1$. See Appendix I for an in-depth analysis of the semimetrics used in this work.

In the context of Section 3, we identify as quantity of interest ϕ the *feasible set* $\overline{\mathcal{R}}_{p,\pi^E}$ ($\mathcal{R}_{p,\pi^E}$), and as distribution generating samples[14] $\mu=\mathbb{P}_{p,\pi^b}$. We have the following result of *non-learnability* of the *feasible set* in the *offline* setting.

Theorem C.1. *Let $\mathcal{M}=\langle\mathcal{S},\mathcal{A},p,\mu_0,H\rangle$ be an MDP without reward and let π^E be the deterministic expert policy. Assume to know π^E in all $(s,h)\in\mathcal{S}\times[\![H]\!]$ and also to know $\mathcal{S}^{p,\pi^E}$, i.e., there is no need to learn them. Let $\mathcal{Z}^{p,\pi^b}$ denote the portion of space covered by a behavioral distribution π^b in $\mathcal{M}$. If $\mathcal{Z}^{p,\pi^b}\neq\mathcal{S}\times\mathcal{A}\times[\![H]\!]$, then $\overline{\mathcal{R}}_{p,\pi^E}$ is not PAC-learnable from $\mathbb{P}_{p,\pi^b}$. Moreover, if $\mathcal{Z}^{p,\pi^b}\neq\mathcal{S}\times\mathcal{A}\times[\![H]\!]$ and for at least one $(s,h)\notin\mathcal{S}^{p,\pi^b}$ there exists a policy $\pi\in\Pi$ such that $\mathbb{P}_{p,\pi}(s_h=s)>0$, then not even $\mathcal{R}_{p,\pi^E}$ is PAC-learnable from $\mathbb{P}_{p,\pi^b}$.*

Proof. Let us start with $\overline{\mathcal{R}}_{p,\pi^E}$. The idea is to construct two problem instances whose feasible sets lie at a fixed non-zero distance and such that samples do not allow to discriminate between them.

We start with the construction of the two instances. Let π^E be the expert's policy and let $\mathcal{M}_1=\langle\mathcal{S},\mathcal{A},p^1,\mu_0,H\rangle$ be an MDP without reward in which policy π^b induces the distribution over trajectories $\mathbb{P}_{p^1,\pi^b}$. By hypothesis, there exists triple $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]$ not in $\mathcal{Z}^{p,\pi^b}$, i.e., such that $\mathbb{P}_{p^1,\pi^b}(s_h=s,a_h=a)=0$. Let us construct another problem instance $\mathcal{M}_2=\langle\mathcal{S},\mathcal{A},p^2,\mu_0,H\rangle$ such that $p^2\equiv_{\mathcal{Z}^{p,\pi^b}}p^1$. This is possible by simply setting, at triple (s,a,h), the condition $p^1_h(\cdot|s,a)\neq p^2_h(\cdot|s,a)$, and equality elsewhere. Observe that, because of this choice, $\mathbb{P}_{p^1,\pi^b}=\mathbb{P}_{p^2,\pi^b}$. Let $\overline{\mathcal{R}}_{p^1,\pi^E}$ and $\overline{\mathcal{R}}_{p^2,\pi^E}$ denote, respectively, the feasible sets of instances $\mathcal{M}_1$ and $\mathcal{M}_2$ with expert's policy π^E.

Now we show that $\overline{\mathcal{R}}_{p^1,\pi^E}\neq\overline{\mathcal{R}}_{p^2,\pi^E}$. To do so, we claim the existence a reward $r\in\overline{\mathcal{R}}_{p^1,\pi^E}$ such that $r\notin\overline{\mathcal{R}}_{p^2,\pi^E}$. W.l.o.g., assume a be a non-expert's action[15]. By definition of $\overline{\mathcal{R}}_{p^1,\pi^E}$, at triple (s,a,h), it holds that $Q^{\pi^E}_h(s,a;p^1,r)\leqslant Q^{\pi^E}_h(s,\pi^E_h(s);p^1,r)$. Since p^2 coincides with p^1 everywhere except for triple (s,a,h), the constraint is equivalent to $Q^{\pi^E}_h(s,a;p^1,r)\leqslant Q^{\pi^E}_h(s,\pi^E_h(s);p^2,r)$. Similarly, we have that $r\in\overline{\mathcal{R}}_{p^2,\pi^E}$ if $Q^{\pi^E}_h(s,a;p^2,r)\leqslant Q^{\pi^E}_h(s,\pi^E_h(s);p^2,r)$.

[14] Observe that, to avoid mentioning the creation of a mixture of distributions $\mathbb{P}_{p,\pi^b}$ and $\mathbb{P}_{p,\pi^E}$ (because of the two datasets), we assume that π^E and $\mathcal{S}^{p,\pi^E}$ are given and need not to be learned. While this simplifies the learning problem, notice that even the estimation of the transition model alone can end up in non-learnability issues.

[15] Otherwise, we can make the same construction with any non-expert's action $a'\in\mathcal{A}$, and we can show that the constraints of the feasible sets of $\mathcal{M}_1$ and $\mathcal{M}_2$ have the same lower bounds $Q^{\pi^E}_h(s,a';p^1,r)=Q^{\pi^E}_h(s,a';p^2,r)$, but different upper bounds $Q^{\pi^E}_h(s,\pi^E_h(s);p^1,r)\neq Q^{\pi^E}_h(s,\pi^E_h(s);p^2,r)$.

Clearly, both the constraints have the same upper bound $Q_h^{\pi^E}(s,\pi_h^E(s);p^2,r)$, and since $Q_{h+1}^{\pi^E}(s',\pi_{h+1}^E(s');p^1,r)=Q_{h+1}^{\pi^E}(s',\pi_{h+1}^E(s');p^2,r)$ for any $s'\in\mathcal{S}$, then $p_h^1(\cdot|s,a)\neq p_h^2(\cdot|s,a)$ entails $Q_h^{\pi^E}(s,a;p^1,r)\neq Q_h^{\pi^E}(s,a;p^2,r)$. Therefore, we can find $r\in\overline{\mathcal{R}}_{p^1,\pi^E}$ such that $r\notin\overline{\mathcal{R}}_{p^2,\pi^E}$, thus $\overline{\mathcal{R}}_{p^1,\pi^E}\neq\overline{\mathcal{R}}_{p^2,\pi^E}$.

We proceed by contradiction. Let us assume that the feasible set $\overline{\mathcal{R}}_{p,\pi^E}$ is PAC-learnable in both $\mathcal{M}_1$ and $\mathcal{M}_2$. By definition of learnability, there exists a semi-metric d and an algorithm $\mathfrak{A}$ with certain properties. By definition of semi-metric, since $\overline{\mathcal{R}}_{p^1,\pi^E}\neq\overline{\mathcal{R}}_{p^2,\pi^E}$, then there exists a certain $c>0$ such that $d(\overline{\mathcal{R}}_{p^1,\pi^E},\overline{\mathcal{R}}_{p^2,\pi^E})=c$. Moreover, by ρ-relaxed triangle inequality, we know that a set of rewards $\widetilde{\mathcal{R}}$ such that $d(\widetilde{\mathcal{R}},\overline{\mathcal{R}}_{p^1,\pi^E})<c/(2\rho)$ and $d(\widetilde{\mathcal{R}},\overline{\mathcal{R}}_{p^2,\pi^E})<c/(2\rho)$ at the same time does not exist, thus the two events $\left\{d(\widetilde{\mathcal{R}},\overline{\mathcal{R}}_{p^1,\pi^E})<c/(2\rho)\right\}$ and $\left\{d(\widetilde{\mathcal{R}},\overline{\mathcal{R}}_{p^2,\pi^E})<c/(2\rho)\right\}$ are disjoint. By the choices $\epsilon<c/(2\rho)$ and $\delta<1/2$, algorithm $\mathfrak{A}$ must satisfy

$$\mathop{\mathbb{P}}_{p^1,\pi^b}\left(d(\widehat{\mathcal{R}}^{\mathfrak{A}},\overline{\mathcal{R}}_{p^1,\pi^E})<\frac{c}{2\rho}\right)>\frac{1}{2}\qquad\wedge\qquad\mathop{\mathbb{P}}_{p^2,\pi^b}\left(d(\widehat{\mathcal{R}}^{\mathfrak{A}},\overline{\mathcal{R}}_{p^2,\pi^E})<\frac{c}{2\rho}\right)>\frac{1}{2}.$$

By construction, we have $\mathbb{P}_{p^1,\pi^b}=\mathbb{P}_{p^2,\pi^b}$. In other words, samples do not allow to discriminate between instances $\mathcal{M}_1$ and $\mathcal{M}_2$, and so between $\overline{\mathcal{R}}_{p^1,\pi^E}$ and $\overline{\mathcal{R}}_{p^2,\pi^E}$. Therefore, when faced with $\mathcal{M}_1$, independently on the number N of samples, algorithm $\mathfrak{A}$ outputs $\widehat{\mathcal{R}}^{\mathfrak{A}}$ such that:

$$\begin{aligned}&\mathop{\mathbb{P}}_{p^1,\pi^b}\left(\left\{d(\widehat{\mathcal{R}}^{\mathfrak{A}},\overline{\mathcal{R}}_{p^1,\pi^E})<\frac{c}{2\rho}\right\}\cup\left\{d(\widehat{\mathcal{R}}^{\mathfrak{A}},\overline{\mathcal{R}}_{p^2,\pi^E})<\frac{c}{2\rho}\right\}\right)\\&\quad=\underbrace{\mathop{\mathbb{P}}_{p^1,\pi^b}\left(d(\widehat{\mathcal{R}}^{\mathfrak{A}},\overline{\mathcal{R}}_{p^1,\pi^E})<\frac{c}{2\rho}\right)}_{>1/2}+\underbrace{\mathop{\mathbb{P}}_{p^1,\pi^b}\left(d(\widehat{\mathcal{R}}^{\mathfrak{A}},\overline{\mathcal{R}}_{p^2,\pi^E})<\frac{c}{2\rho}\right)}_{>1/2}>1,\end{aligned}$$

where we have used that the two events are disjoint. This is clearly a contradiction, thus the statement of the theorem holds for the notion of feasible set in Definition 3.1.

With regards to the novel notion of feasible set $\mathcal{R}_{p,\pi^E}$, the proof is analogous. The only difference is in how to show that $\mathcal{R}_{p^1,\pi^E}\neq\mathcal{R}_{p^2,\pi^E}$ when the triple $(s,a,h)\notin\mathcal{Z}^{p,\pi^b}$ is such that $(s,h)\notin\mathcal{S}^{p,\pi^b}$. Indeed, by Definition 3.2, in such (s,h) there is no constraint on which action shall be optimal. However, by the hypothesis contained in the statement of the theorem, there exists a policy that brings to (s,h), so since π^b does not reach (s,h), then there exists another triple (s',a',h'), with $h'<h$, such that $(s',a',h')\notin\mathcal{Z}^{p,\pi^b}$ and $(s',h')\in\mathcal{S}^{p,\pi^b}$. Therefore, the same passages adopted to show that $\mathcal{R}_{p^1,\pi^E}\neq\mathcal{R}_{p^2,\pi^E}$ can be used to show also that $\mathcal{R}_{p^1,\pi^E}\neq\mathcal{R}_{p^2,\pi^E}$. It should be remarked that the hypothesis $\mathcal{Z}^{p,\pi^b}\neq\mathcal{S}\times\mathcal{A}\times[\![H]\!]$ alone is not sufficient[16] for the non-learnability of $\mathcal{R}_{p,\pi^E}$. This concludes the proof. □

The following theorem demonstrates that the solution concepts (subset and superset) that we propose in our framework are the *tightest learnable*. Observe that Theorem C.2 entails Theorem C.1. However, since in the proof of Theorem C.2 we make use of the construction introduced in the proof of Theorem C.1, we prefer to keep the two theorems separated.

Theorem C.2. *Let $\mathcal{M}=\langle\mathcal{S},\mathcal{A},p,\mu_0,H\rangle$ be an MDP without reward and let π^E be the deterministic expert policy. Assume to know π^E in all $(s,h)\in\mathcal{S}\times[\![H]\!]$ and also to know $\mathcal{S}^{p,\pi^E}$, i.e., there is no need to learn them. Let $\mathcal{S}^{p,\pi^b},\mathcal{Z}^{p,\pi^b}$ denote the portion of space covered by a behavioral distribution π^b in $\mathcal{M}$. Then, $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$ and $\overline{\mathcal{R}}^{\cup}_{p,\pi^E}$ are, respectively, the* tightest *subset and superset of $\overline{\mathcal{R}}_{p,\pi^E}$ that can be* learned *from $\mathbb{P}_{p,\pi^b}$. Moreover, $\mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E}$ are, respectively, the* tightest *subset and superset of $\mathcal{R}_{p,\pi^E}$ that can be* learned *from $\mathbb{P}_{p,\pi^b}$.*

Proof. The theorem states that the considered quantities are the tightest learnable. Thus, we split the proof in two parts. First, we prove that such quantities are PAC-learnable, then we show that there is no other object that is at the same time learnable and tighter.

Let us begin with $\mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E}$. By Definition C.1, these quantities are PAC-learnable if we can find a semi-metric d between sets of rewards and an algorithm $\mathfrak{A}$ such that, for any arbitrarily small choice of the accuracy ϵ and confidence

[16] Consider for instance the MDP without reward $\mathcal{M}$ in which $\mathcal{S}\times[\![H]\!]\setminus\mathcal{S}^{p,\pi^b}=\{(\bar{s},1)\}$ and $\mathcal{S}\times\mathcal{A}\times[\![H]\!]\setminus\mathcal{Z}^{p,\pi^b}=\{(\bar{s},a_1,1),\dots,(\bar{s},a_A,1)\}$, i.e., that π^b covers the entire space except for a state $\bar{s}$ at stage $h=1$ ($\mu_0(\bar{s})=0$). Clearly, such state does not appear in the constraints defining the feasible set, and the feasible set $\mathcal{R}_{p,\pi^E}$ is learnable by $\mathbb{P}_{p,\pi^b}$!

δ, we can always find a *finite* number of samples that algorithm $\mathfrak{A}$ can use to compute a set of rewards ϵ-close, according to semi-metric d, to $\mathcal{R}^{\cap}_{p,\pi^E}$ (or to $\mathcal{R}^{\cup}_{p,\pi^E}$) w.h.p.. This is exactly what, for instance, Theorem 5.1 states: Algorithm 1 requires a finite number of samples to compute an ϵ-correct estimate of $\mathcal{R}^{\cap}_{p,\pi^E}$ (or of $\mathcal{R}^{\cup}_{p,\pi^E}$) w.h.p. according to any of the semi-metrics presented in Definition 4.1 (for which we prove in Appendix I that a ρ-relaxed triangle inequality holds). We are not going to show that an analogous of Theorem 5.1 holds also for $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$ and $\overline{\mathcal{R}}^{\cup}_{p,\pi^E}$.

Now we show that these quantities are the tightest learnable. Let us start with $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$, and then we will move to $\overline{\mathcal{R}}^{\cup}_{p,\pi^E}$, $\mathcal{R}^{\cap}_{p,\pi^E}$, and $\mathcal{R}^{\cup}_{p,\pi^E}$.

The idea is to construct by contradiction another concept $\overline{R}^{\cap}_{p,\pi^E}$ (non calligraphic) of subset of $\overline{\mathcal{R}}_{p,\pi^E}$ which is tighter than $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$, and then show that we can construct a problem instance in which the newly defined concept $\overline{R}^{\cap}_{p,\pi^E}$ fails at being a subset of $\overline{\mathcal{R}}_{p,\pi^E}$. Thus, by contradiction, let us assume that there exists a problem instance $\mathcal{M}=\langle\mathcal{S},\mathcal{A},p,\mu_0,H\rangle$ with expert's policy π^E and distribution generating samples $\mathbb{P}_{p,\pi^b}$, in which there exists a PAC-learnable set $\overline{R}^{\cap}_{p,\pi^E}$ from $\mathbb{P}_{p,\pi^b}$ such that $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}\subset\overline{R}^{\cap}_{p,\pi^E}\subseteq\overline{\mathcal{R}}_{p,\pi^E}$. If $\mathcal{Z}^{p,\pi^b}=\mathcal{S}\times\mathcal{A}\times[\![H]\!]$, then $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}=\overline{\mathcal{R}}_{p,\pi^E}$, so we consider the case in which $\mathcal{Z}^{p,\pi^b}\subset\mathcal{S}\times\mathcal{A}\times[\![H]\!]$. Let $\bar{r}$ be a reward of $\overline{R}^{\cap}_{p,\pi^E}$ which is not present in $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$. By definition of $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$, we have that $\bar{r}\notin\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$ if and only if there exists $p'\in[p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}$ and $\pi'\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^b}}}$ such that $\bar{r}\notin\mathcal{R}_{p',\pi'}$. Therefore, similarly to the proof of Theorem C.1, we can construct a new problem instance $\mathcal{M}'=\langle\mathcal{S},\mathcal{A},p',\mu_0,H\rangle\cup\{\pi'\}$ such that, since $\mathcal{Z}^{p,\pi^b}\subset\mathcal{S}\times\mathcal{A}\times[\![H]\!]$, $\overline{\mathcal{R}}_{p,\pi^E}\neq\mathcal{R}_{p',\pi'}$. By definition of p', we know that $\mathbb{P}_{p',\pi^b}=\mathbb{P}_{p,\pi^b}$, thus any algorithm $\mathfrak{A}$ estimating the new concept of subset $\overline{R}^{\cap}_{p,\pi^E}$ fails to distinguish instances $\mathcal{M}$ and $\mathcal{M}'$. This means that we can choose ϵ,δ so that if $\mathfrak{A}$ returns in $\mathcal{M}$ a set containing $\bar{r}$, then with high probability it will return it also in $\mathcal{M}'$. However, $\bar{r}\notin\mathcal{R}_{p',\pi'}$, so we get a contradiction.

The proof for $\overline{\mathcal{R}}^{\cup}_{p,\pi^E}$ is analogous. By contradiction, we claim the existence of a set $\overline{R}^{\cup}_{p,\pi^E}$ such that $\overline{\mathcal{R}}_{p,\pi^E}\subseteq\overline{R}^{\cup}_{p,\pi^E}\subset\overline{\mathcal{R}}^{\cup}_{p,\pi^E}$. The contradiction will be shown by considering a reward $\bar{r}$ of $\mathcal{R}^{\cup}_{p,\pi^E}$ which is not in $\overline{R}^{\cup}_{p,\pi^E}$, and then constructing the problem instance in which the feasible set contains exactly that reward, but set $\overline{R}^{\cup}_{p,\pi^E}$ does not (w.h.p.).

As far as $\mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E}$ are concerned, the proofs are analogous to those presented above. However, there is a detail that has to be explained. Specifically, we have seen in the proof of Theorem C.1 that the condition $\mathcal{Z}^{p,\pi^b}\subset\mathcal{S}\times\mathcal{A}\times[\![H]\!]$ is not a sufficient condition for $\mathcal{R}_{p^1,\pi^E}\neq\mathcal{R}_{p^2,\pi^E}$. Therefore, in principle, the proof for $\overline{\mathcal{R}}^{\cap}_{p,\pi^E}$ ($\overline{\mathcal{R}}^{\cup}_{p,\pi^E}$) cannot be adapted directly to $\mathcal{R}^{\cap}_{p,\pi^E}$ ($\mathcal{R}^{\cup}_{p,\pi^E}$). However, we observe that $\mathcal{R}^{\cap}_{p,\pi^E}\subset\mathcal{R}_{p,\pi^E}$ (respectively, $\mathcal{R}_{p,\pi^E}\subset\mathcal{R}^{\cup}_{p,\pi^E}$) holds strictly if $\mathcal{Z}^{p,\pi^b}\neq\mathcal{S}\times\mathcal{A}\times[\![H]\!]$ and for at least one $(s,h)\notin\mathcal{S}^{p,\pi^b}$ there exists a policy $\pi\in\Pi$ such that $\mathbb{P}_{p,\pi}(s_h=s)>0$. Otherwise, it holds that $\mathcal{R}^{\cap}_{p,\pi^E}=\mathcal{R}_{p,\pi^E}=\mathcal{R}^{\cup}_{p,\pi^E}$, as explained in Section D.4. This is exactly the condition required in Theorem C.1 for proving $\mathcal{R}_{p^1,\pi^E}\neq\mathcal{R}_{p^2,\pi^E}$. Therefore, by using this observation, we can prove the statement of the theorem also for $\mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E}$. □

D. Further considerations

In this appendix, we collect a variety of considerations and remarks about the learning framework introduced, about the need of two datasets, alternative representations of the feasible set, and some others.

D.1. About the need of two datasets

We presented **IRLO** (and, subsequently, **PIRLO**) in the case two datasets $\mathcal{D}^b$ and $\mathcal{D}^E$ collected with π^b and π^E, respectively, are available. This scenario is common in previous IRL works (Boularias et al., 2011) but, although convenient for our analysis, it is not strictly necessary to achieve a meaningful sample complexity. Indeed, we remark that the expert's dataset is employed for estimating the expert's support $\mathcal{S}^{p,\pi^E}$ and policy π^E. This task can be anyway achieved under Assumption 2.1 using just one dataset $\mathcal{D}=\{\langle s_1^{b,i},a_1^{b,i},a_1^{E,i},\dots,s_{H-1}^{b,i},a_{H-1}^{b,i},a_{H-1}^{E,i},s_H^{b,i}\rangle\}_{i\in[\![\tau]\!]}$ playing the behavioral policy π^b and keeping track of the expert's actions too. In such a case, we must require that every transition $(s,\pi_h^E(s),s')$ is exercised at least once in $\mathcal{D}$. This leads to a sample complexity bound which is larger in the last constant term in which $\rho_{\min}^{\pi^E,\mathcal{Z}^{p,\pi^E}}$ is replaced with:

$$\min\left\{\rho_{\min}^{\pi^E,\mathcal{Z}^{p,\pi^E}},\ \min_{\substack{(s,h)\in\mathcal{S}^{p,\pi^E},\\ s'\in\mathcal{S}:p_h(s'|s,\pi_h^E(s))>0}}\rho_h^{p,\pi^b}(s,a)p_h(s'|s,\pi_h^E(s))\right\}.$$

D.2. About the dependence on $\rho_{\min}$

The majority of the results presented for the d_∞ semimetric in this paper are characterized by a dependence on the minimum non-zero visitation probability $\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}}$ of the behavioral policy π^b. This is expected since we are targeting as solution concept the *tightest learnable* subset and supersets of the feasible set. Clearly, one can further relax this requirement, accepting to target *non-tightest learnable* sets with a benefit in the sample complexity. Consider a minimum-visitation threshold $\overline{\rho}$, we define $\mathcal{Z}_{\overline{\rho}}^{p,\pi^b}=\{(s,a,h)\colon \rho_h^{p,\pi^b}(s,a)>\overline{\rho}\}$ as the set of triples (s,a,h) that are visited by at least $\overline{\rho}$ probability (notice that $\mathcal{Z}^{p,\pi^b}=\mathcal{Z}_0^{p,\pi^b}$). We can use this set to employ suitable equivalence $\equiv_{\mathcal{Z}_{\overline{\rho}}^{p,\pi^b}}$ relations over transition models to group together those that differ in triples (s,a,h) visited with probability smaller than $\overline{\rho}$. This allows to redefine the sub- and super-feasible sets as follows:

$$\mathcal{R}^{\cap}_{p,\pi^E,\overline{\rho}}:=\bigcap_{p'\in[p]_{\equiv_{\mathcal{Z}_{\overline{\rho}}^{p,\pi^b}}}}\mathcal{R}_{p',\pi^E},\quad \mathcal{R}^{\cup}_{p,\pi^E,\overline{\rho}}:=\bigcup_{p'\in[p]_{\equiv_{\mathcal{Z}_{\overline{\rho}}^{p,\pi^b}}}}\mathcal{R}_{p',\pi^E}.$$

Obviously, by the definition of the equivalence relation, we have that $\mathcal{R}^{\cap}_{p,\pi^E,\overline{\rho}}\subseteq\mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E}\subseteq\mathcal{R}^{\cup}_{p,\pi^E,\overline{\rho}}$. Under the assumption that $\overline{\rho}\leqslant\rho_h^{p,\pi^b}(s,\pi_h^E(s))$ for every (s,h), these sets are clearly *learnable*, but lose the property of being the *tightest* ones. The advantage of targeting these feasible sets is that we can reproduce the same proofs done for the original $\mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E}$ obtaining a smaller sample complexity that scales with $\overline{\rho}$ instead of $\rho_{\min}^{p,\pi^b}$.

D.3. Equivalent definitions of the feasible sets

In both Definition 3.1 and Theorem 3.1, we represent the set of constraints defining the (old) feasible set using the Q-function of policy π^E or of some policy $\overline{\pi}\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}$. However, it is possible to provide an alternative equivalent representation based on the optimal Q-function Q^*. It is easy to notice that the old feasible set $\overline{\mathcal{R}}_{p,\pi^E}$ can be rewritten as:

$$\overline{\mathcal{R}}_{p,\pi^E}=\{r\in\mathfrak{R}\,|\,\forall(s,h)\in\mathcal{S}\times[\![H]\!],\forall a\in\mathcal{A}\colon Q_h^*(s,\pi_h^E(s);p,r)\geqslant Q_h^*(s,a;p,r)\}. \tag{13}$$

Moreover, thanks to Lemma E.1, the new feasible set $\mathcal{R}_{p,\pi^E}$ can be rewritten as:

$$\mathcal{R}_{p,\pi^E}=\{r\in\mathfrak{R}\,|\,\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\colon Q_h^*(s,\pi_h^E(s);p,r)\geqslant Q_h^*(s,a;p,r)\}.$$

We prefer to work with the representations presented in the main paper because the relaxations (see Section 6) of those representations are "better" (See Appendix G) than the relaxations of the representations just introduced.

As a direct consequence of Theorem 3.1, we have the following corollary (see Appendix E for the proof).

Corollary D.1. *In the setting of Definition 3.2 the feasible reward set $\mathcal{R}_{p,\pi^E}$ satisfies:*

$$\mathcal{R}_{p,\pi^E}=\bigcup_{\pi'\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}}\overline{\mathcal{R}}_{p,\pi'}.$$

This corollary provides the explicit relationship between the old $\overline{\mathcal{R}}_{p,\pi^E}$ and new $\mathcal{R}_{p,\pi^E}$ definitions of feasible set. Clearly $\overline{\mathcal{R}}_{p,\pi^E}\subseteq\mathcal{R}_{p,\pi^E}$. By using Corollary D.1, we can rewrite $\mathcal{R}^{\cap}_{p,\pi^E}$ as $\mathcal{R}^{\cap}_{p,\pi^E}=\bigcap_{p'\in[p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}}\bigcup_{\pi'\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}}\overline{\mathcal{R}}_{p',\pi'}$. Observe that in general $\mathcal{R}^{\cap}_{p,\pi^E}\neq\bigcup_{\pi'\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}}\bigcap_{p'\in[p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}}\overline{\mathcal{R}}_{p',\pi'}$, because the union of the intersection is different from the intersection of the union. Furthermore, because of the different definitions of $\overline{\mathcal{R}}_{p,\pi^E}$ and $\mathcal{R}_{p,\pi^E}$, we have that, in general, $\mathcal{R}^{\cap}_{p,\pi^E}\nsubseteq\overline{\mathcal{R}}_{p,\pi^E}$ i.e., the subset for the new notion of *feasible set* is not a subset of $\overline{\mathcal{R}}_{p,\pi^E}$. Differently, with regard to the superset, it holds that $\mathcal{R}^{\cup}_{p,\pi^E}\supseteq\overline{\mathcal{R}}_{p,\pi^E}$.

It should be remarked that Corollary D.1 and Theorem 3.1 are not in contradiction. Indeed, looking at the union over policies in Corollary D.1, one might expect an existential quantifier inside Theorem 3.1, but we find a universal quantifier. By using Corollary D.1 and Eq. 13, we can "transform" the union into an existential quantifier to obtain:

$$\mathcal{R}_{p,\pi^E}=\{r\in\mathfrak{R}\,|\,\exists\overline{\pi}\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}\colon\forall(s,h)\in\mathcal{S}\times[\![H]\!],\forall a\in\mathcal{A}\colon Q_h^*(s,\overline{\pi}_h(s);p,r)\geqslant Q_h^*(s,a;p,r)\},$$

i.e., we are representing the feasible set $\mathcal{R}_{p,\pi^E}$ as the set of all the rewards that induce an optimal policy in $[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}$. From Theorem 3.1, we have:

$$\mathcal{R}_{p,\pi^E} = \{r \in \mathfrak{R} \,|\, \forall \overline{\pi} \in [\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}, \forall (s,h) \in \mathcal{S}^{p,\pi^E}, \forall a \in \mathcal{A} : Q_h^{\overline{\pi}}(s, \pi_h^E(s); p, r) \geqslant Q_h^{\overline{\pi}}(s, a; p, r)\},$$

i.e., we are representing the feasible set $\mathcal{R}_{p,\pi^E}$ as the set of all the rewards for which playing the expert's action in $\mathcal{S}^{p,\pi^E}$ is the optimal strategy irrespective of the optimal action outside $\mathcal{S}^{p,\pi^E}$. To put it simple, Corollary D.1 uses the existential quantifier because it says that a certain policy in $[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}$ is optimal, while Theorem 3.1 uses the universal quantifier because it does not care about which policy in $[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}$ is optimal, but only that the expert's action is played in $\mathcal{S}^{p,\pi^E}$. The $\exists$ gives the optimal policy, while the $\forall$ says that one of the policies in $[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}$ is optimal, without telling which. Clearly, there is no contradiction.

D.4. A remark about non reachable states

The strict condition $\mathcal{Z}^{p,\pi^b} \subset \mathcal{S} \times \mathcal{A} \times [\![H]\!]$ alone is not a sufficient condition to have $\mathcal{R}^{\cap}_{p,\pi^E} \neq \mathcal{R}_{p,\pi^E} \neq \mathcal{R}^{\cup}_{p,\pi^E}$. Indeed, if the portion of $\mathcal{S} \times \mathcal{A} \times [\![H]\!]$ not contained into $\mathcal{Z}^{p,\pi^b}$ is made only of $(s,h) \in \mathcal{S} \times [\![H]\!]$ for which there is no $\pi \in \Pi$ such that $(s,h) \in \mathcal{S}^{p,\pi}$ for the given $p \in \mathcal{P}$, then neither the policy nor the transition model in (s,h) appears in the constraints of $\mathcal{R}^{\cap}_{p,\pi^E}, \mathcal{R}_{p,\pi^E}$, or $\mathcal{R}^{\cup}_{p,\pi^E}$ (when viewed using Theorem 3.1). In practice, the values of the rewards r of $\mathcal{R}^{\cap}_{p,\pi^E}$ (and $\mathcal{R}_{p,\pi^E}$, and $\mathcal{R}^{\cup}_{p,\pi^E}$) in such $(s,a,h) \notin \mathcal{Z}^{p,\pi^b}$ can be chosen arbitrarily, irrespective of the reward in any other $(s',a',h') \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]$, and therefore we have $\mathcal{R}^{\cap}_{p,\pi^E} = \mathcal{R}_{p,\pi^E} = \mathcal{R}^{\cup}_{p,\pi^E}$.

D.5. An annoying corner case

To cope with the bitter lesson of Section 8, we work with two datasets. As aforementioned, we use $\mathcal{D}^E$ to estimate $\mathcal{S}^{p,\pi^E}$ and π^E, and we use $\mathcal{D}^b$ to estimate p. However, it might happen the following situation. Let $(s,h) \in \widehat{\mathcal{S}}^{p,\pi^E}$ (where $\widehat{\mathcal{S}}^{p,\pi^E}$ is the estimate of $\mathcal{S}^{p,\pi^E}$ computed from $\mathcal{D}^E$), and let $\widehat{\mathcal{S}}^{p,\pi^E}_{h+1} = \{\overline{s}\}$, i.e., dataset $\mathcal{D}^E$ tells us that the support of the expert's policy at $h+1$ is made of state $\overline{s}$ only. Let $a^E := \widehat{\pi}_h^E(s)$. By using dataset $\mathcal{D}^b$, we might come up with the estimate of the transition model at (s, a^E, h):

$$\begin{cases} \widehat{p}_h(\overline{s} | s, a^E) > 0 \\ \widehat{p}_h(s' | s, a^E) > 0 \end{cases},$$

where $s' \in \mathcal{S}$ is some other state not in $\widehat{\mathcal{S}}^{p,\pi^E}_{h+1}$. Clearly, this means that $s' \in \mathcal{S}^{p,\pi^E}_{h+1}$; however, due to finite data, dataset $\mathcal{D}^E$ does not provide us with this information. This fact provides a contradiction between $\widehat{p}$ and $\widehat{\mathcal{S}}^{p,\pi^E}$. To avoid issues in the implementation of **PIRLO**, we define the confidence set $\mathcal{C}(\widehat{p}, b)$ (see Eq. 7) by allowing the support of the transition model of expert's actions to be compatible with the estimate provided by $\mathcal{D}^b$, i.e., we set:

$$\mathcal{C}(\widehat{p}, b) := \Big\{ p' \in \mathcal{P} \,|\, \forall (s,a,h) \in \widehat{\mathcal{Z}}^{p,\pi^b} : \|p'_h(\cdot|s,a) - \widehat{p}_h(\cdot|s,a)\|_1 \leqslant b_h(s,a) \wedge \\ \forall (s,h) \in \widehat{\mathcal{S}}^{p,\pi^E}, \forall s' \notin \big(\widehat{\mathcal{S}}^{p,\pi^E}_{h+1} \cup \operatorname{supp} \widehat{p}_h(\cdot|s, \widehat{\pi}_h^E(s))\big) : p'_h(s'|s, \widehat{\pi}_h^E(s)) = 0 \Big\}.$$

Observe that the union over the support of $\widehat{p}_h(\cdot|s,a^E)$ solves the potential issue created by the corner case described in this section. It should be remarked that, under good event $\mathcal{E}$ (see Appendix F), it holds that $\widehat{\mathcal{S}}^{p,\pi^E} = \mathcal{S}^{p,\pi^E}$, and therefore $\widehat{\mathcal{S}}^{p,\pi^E}_{h+1} \cup \operatorname{supp} \widehat{p}_h(\cdot|s, \widehat{\pi}_h^E(s)) = \widehat{\mathcal{S}}^{p,\pi^E}_{h+1} = \mathcal{S}^{p,\pi^E}_{h+1}$.

D.6. Distances d and d_∞ control the distance between value functions

We provide the proof of the proposition reported in Section 4.

Proposition 4.2. *For any $r, r' \in \mathfrak{R}$, it holds that:*

$$d^G_{V*}(r, \widehat{r}) \leqslant 2 d_\infty(r, \widehat{r}) \leqslant \frac{2 d(r, \widehat{r})}{\rho^{\pi^b, \mathcal{Z}^{p,\pi^b}}_{\min}}.$$

Proof. The proof is similar to that of Theorem 4.1 of (Metelli et al., 2023). For any s, h and policy $\widehat{\pi}^*$ optimal in some $\widehat{r}$, we can write:

$$\begin{aligned}
V_h^*(s;r) - V_h^{\widehat{\pi}^*}(s;r) &= V_h^*(s;r) - V_h^{\widehat{\pi}^*}(s;r) \pm V_h^{\widehat{\pi}^*}(s;\widehat{r}) \\
&= \left(V_h^*(s;r) - V_h^{\widehat{\pi}^*}(s;\widehat{r})\right) + \left(V_h^{\widehat{\pi}^*}(s;\widehat{r}) - V_h^{\widehat{\pi}^*}(s;r)\right) \\
&\overset{(1)}{\leqslant} \left(V_h^*(s;r) - V_h^{\pi^*}(s;\widehat{r})\right) + \left(V_h^{\widehat{\pi}^*}(s;\widehat{r}) - V_h^{\widehat{\pi}^*}(s;r)\right) \\
&= \sum_{l=h}^{H} \sum_{(s',a') \in \mathcal{S}\times\mathcal{A}} \mathbb{P}_{p,\pi^*}(s_l = s', a_l = a' | s_h = s)(r_l(s',a') - \widehat{r}_l(s',a')) \\
&\quad + \sum_{l=h}^{H} \sum_{(s',a') \in \mathcal{S}\times\mathcal{A}} \mathbb{P}_{p,\widehat{\pi}^*}(s_l = s', a_l = a' | s_h = s)(r_l(s',a') - \widehat{r}_l(s',a')) \\
&\leqslant 2 \sum_{l=h}^{H} \|r_l - \widehat{r}_l\|_\infty,
\end{aligned}$$

where at (1) we have used that $\widehat{\pi}^*$ is optimal under $\widehat{r}$.

Multiplying both sides by $1/M(r,\widehat{r})$ concludes the proof, and noticing that ρ^{p,π^b} permits to bound d_∞ by d, we get the result. □

E. Proofs of Section 3 and Section 4

In this section, we provide the missing proofs of Section 3 and Section 4.

To prove Theorem 3.1, it is useful to introduce the following lemma.

Lemma E.1. *In the setting of Definition 3.2, the feasible reward set $\mathcal{R}_{p,\pi^E}$ satisfies:*

$$\mathcal{R}_{p,\pi^E} = \{r \in \mathfrak{R} \,|\, \forall (s,h) \in \mathcal{S}^{p,\pi^E}, \forall a \in \mathcal{A} \colon Q_h^*(s, \pi_h^E(s); p, r) \geqslant Q_h^*(s,a;p,r)\}.$$

Proof. The statement of the theorem is equivalent to the necessary and sufficient condition:

$$J(\pi^E;\mu_0,p,r) = \max_{\pi\in\Pi} J(\pi;\mu_0,p,r) \iff \forall (s,h) \in \mathcal{S}^{p,\pi^E} : Q_h^*(s,\pi_h^E(s);p,r) = \max_{a\in\mathcal{A}} Q_h^*(s,a;p,r).$$

We split the proof in two parts. First we show the sufficiency, then the necessity.

Let us start with the sufficiency. Let r be any reward in $\mathfrak{R}$ and p any transition model in $\mathcal{P}$. By contradiction, suppose that there exists a policy $\pi' \in \arg\max_{\pi\in\Pi} J(\pi;\mu_0,p,r)$ for which there exists a (s',h') in the union of the supports of the $h \in [\![H]\!]$ distributions $\rho_h^{p,\pi'}(\cdot)$ in which $Q_{h'}^*(s', \pi'_{h'}(s');p,r) < \max_{a'\in\mathcal{A}} Q_{h'}^*(s',a';p,r)$ (the notation refers to a deterministic π' but it can be taken stochastic by computing the expected value). Let $\pi^* \in \arg\max_{\pi\in\Pi} V_h^\pi(s;p,r)\, \forall (s,h) \in \mathcal{S} \times [\![H]\!]$ be an auxiliary optimal policy whose existence is a widely-known result in RL (see Puterman, 1994). By hypothesis, it holds that:

$$J(\pi';\mu_0,p,r) = \max_{\pi\in\Pi} J(\pi;\mu_0,p,r) = J(\pi^*;\mu_0,p,r).$$

From the performance difference lemma (Kakade & Langford, 2002), by denoting the advantage function by $A_h^\pi(s,a;p,r) := Q_h^\pi(s,a;p,r) - V_h^\pi(s;p,r)$, we can write:

$$\begin{aligned}
J(\pi';\mu_0,p,r) - J(\pi^*;\mu_0,p,r) &= \sum_{h\in[\![H]\!]} \mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi'}(\cdot,\cdot)} \left[A_h^{\pi^*}(s,a;p,r)\right] \\
&\overset{(1)}{=} \rho_{h'}^{p,\pi'}(s') A_{h'}^{\pi^*}(s', \pi'_{h'}(s');p,r) \\
&\overset{(2)}{<} 0,
\end{aligned}$$

where at (1) we have used that in all $(s,h)\in\mathcal{S}\times[\![H]\!]\setminus\{(s',h')\}$ the policy π' prescribes the action greedy w.r.t. Q^*, and thus the advantage is 0, and (2) holds by (contradiction) hypothesis. However, by hypothesis, we know that $J(\pi^*;\mu_0,p,r)-J(\pi';\mu_0,p,r)=0$, thus we have obtained a contradiction, so the sufficiency holds.

As far as the necessity is concerned, let us consider again an auxiliary optimal policy π^* and a policy π' such that $Q^*_{h'}(s',\pi'_{h'}(s');p,r)=\max_{a'\in\mathcal{A}}Q^*_{h'}(s',a';p,r)$ in the support of the $h\in[\![H]\!]$ distributions $\rho_h^{p,\pi'}(\cdot)$. By applying the performance difference lemma, we can write:

$$\begin{aligned}J(\pi';\mu_0,p,r)-J(\pi^*;\mu_0,p,r)&=\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{s\sim\rho_h^{p,\pi'}(\cdot)}\left[A_h^{\pi^*}(s,a;p,r)\right]\\&=\sum_{h\in[\![H]\!]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}:\rho_h^{p,\pi'}(s,a)>0}\rho_h^{p,\pi'}(s,a)\underbrace{A_h^{\pi^*}(s,a;p,r)}_{=0}\\&=0,\end{aligned}$$

where we have simply used the hypothesis.

By setting $\pi^E\equiv\pi'$, we get the result. □

Now, we are ready to prove Theorem 3.1:

Theorem 3.1. *In the setting of Definition 3.2, the feasible reward set $\mathcal{R}_{p,\pi^E}$ satisfies:*

$$\begin{aligned}\mathcal{R}_{p,\pi^E}=\{r\in\mathfrak{R}\,|\,\forall\overline{\pi}\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}},\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}:\\Q_h^{\overline{\pi}}(s,\pi_h^E(s);p,r)\geqslant Q_h^{\overline{\pi}}(s,a;p,r)\}.\end{aligned}\tag{4}$$

Proof. Thanks to Lemma E.1, to prove the statement of the theorem we have to show the equivalence of the constraints:

$$\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}:\quad Q_h^*(s,\pi_h^E(s);p,r)\geqslant Q_h^*(s,a;p,r)\tag{14}$$

$$\Longleftrightarrow$$

$$\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A},\forall\overline{\pi}\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}:\quad Q_h^{\overline{\pi}}(s,\pi_h^E(s);p,r)\geqslant Q_h^{\overline{\pi}}(s,a;p,r),\tag{15}$$

where we have exchanged the order of the quantifiers (because they all are of the same type). Observe that Eq. 14 can be rewritten as:

$$\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}:\quad Q_h^{\pi^*}(s,\pi_h^E(s);p,r)\geqslant Q_h^{\pi^*}(s,a;p,r),$$

because of the existence of some optimal policy π^* (see Puterman, 1994). Now, by induction over $h\in[\![H]\!]$, it is easy to show that Eq. 14 entails the existence of an optimal policy $\pi^*\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}$. Therefore, we can rewrite the constraint as:

$$\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}:\quad Q_h^{\pi^E}(s,\pi_h^E(s);p,r)\geqslant Q_h^{\pi^*}(s,a;p,r),$$

since playing π^* from $\mathcal{S}^{p,\pi^E}$ brings again into $\mathcal{S}^{p,\pi^E}$. By definition of π^*, we have:

$$Q_h^{\pi^*}(s,a;p,r)=Q_h^*(s,a;p,r):=\max_{\pi\in\Pi}Q_h^{\pi}(s,a;p,r)\geqslant Q_h^{\overline{\pi}}(s,a;p,r),$$

for all $\overline{\pi}\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}$. Since $\pi^*\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}$, then we have shown that the two conditions in Eq. 14 and Eq. 15 are equivalent. □

As a direct consequence of Theorem 3.1, we have the following corollary.

Corollary D.1. *In the setting of Definition 3.2 the feasible reward set $\mathcal{R}_{p,\pi^E}$ satisfies:*

$$\mathcal{R}_{p,\pi^E}=\bigcup_{\pi'\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}}\overline{\mathcal{R}}_{p,\pi'}.$$

Proof. Let $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]$ and let p be a transition model in $\mathcal{P}$. Define:

$$\mathfrak{R}_a := \{r\in\mathfrak{R}\,|\,Q_h^*(s,a;p,r)=\max_{a'\in\mathcal{A}}Q_h^*(s,a';p,r)\},$$

i.e., the set of rewards satisfying the constraint on the optimality of action a in a single (s,h) pair. It is well known (see Puterman, 1994) that, given a reward function and a transition model, there always exists an optimal policy whose Q-function coincides, by definition, with the optimal Q-function. In other words, for any $p\in\mathcal{P}$ and any $r\in\mathfrak{R}$, the optimal Q-function is "well-defined". Therefore, it holds that:

$$\bigcup_{a\in\mathcal{A}}\mathfrak{R}_a=\mathfrak{R},$$

because *we are making the union of the rewards that induce action a to be optimal in (s,h) for any $a\in\mathcal{A}$.* To put it simple, if we add the constraint that at pair $(s,h)\in\mathcal{S}\times[\![H]\!]$ there *exists* an optimal action, we are not actually adding a constraint. Notice that we can do the same with policies $\pi\in\Pi$ instead of actions $a\in\mathcal{A}$. Thanks to Lemma E.1 and the property just highlighted, we can write:

$$\begin{aligned}\mathcal{R}_{p,\pi^E}&=\{r\in\mathfrak{R}\,|\,\forall(s,h)\in\mathcal{S}^{p,\pi^E}:Q_h^*(s,\pi_h^E(s);p,r)=\max_{a\in\mathcal{A}}Q_h^*(s,a;p,r)\}\\&=\{r\in\mathfrak{R}\,|\,\exists\pi'\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}:\forall(s,h)\in\mathcal{S}\times[\![H]\!]:Q_h^*(s,\pi'_h(s);p,r)=\max_{a\in\mathcal{A}}Q_h^*(s,a;p,r)\}\\&=\bigcup_{\pi'\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}}\overline{\mathcal{R}}_{p,\pi'}.\end{aligned}$$

In this way, the constraints are defined only for $(s,h)\in\mathcal{S}^{p,\pi^E}$. □

With regards to Section 4, we provide the following proposition.

Proposition 4.1. *For any $r,r'\in\mathfrak{R}$, it holds that:*

$$d(r,r')\leqslant 2d_\infty(r,r')\leqslant\frac{2}{\rho_{min}^{\pi^b,\mathcal{Z}^{p,\pi^b}}}d(r,r').$$

Proof. The first inequality is straightforward. For the second, observe that:

$$\begin{aligned}d_\infty(r,\widehat{r}):&=\frac{1}{M}\sum_{h\in[H]}\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}|r_h(s,a)-\widehat{r}_h(s,a)|\\&=\frac{1}{M}\sum_{h\in[H]}\max\Big\{\max_{(s,a)\in\mathcal{Z}_h^{p,\pi^b}}|r_h(s,a)-\widehat{r}_h(s,a)|,\max_{(s,a)\notin\mathcal{Z}_h^{p,\pi^b}}|r_h(s,a)-\widehat{r}_h(s,a)|\Big\}\\&=\frac{1}{\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}}}\frac{1}{M}\sum_{h\in[H]}\max\Big\{\max_{(s,a)\in\mathcal{Z}_h^{p,\pi^b}}\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}}|r_h(s,a)-\widehat{r}_h(s,a)|,\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}}\max_{(s,a)\notin\mathcal{Z}_h^{p,\pi^b}}|r_h(s,a)-\widehat{r}_h(s,a)|\Big\}\\&\overset{(1)}{\leqslant}\frac{1}{\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}}}\frac{1}{M}\sum_{h\in[H]}\max\Big\{\max_{(s,a)\in\mathcal{Z}_h^{p,\pi^b}}\rho_h^{p,\pi^b}(s,a)|r_h(s,a)-\widehat{r}_h(s,a)|,\max_{(s,a)\notin\mathcal{Z}_h^{p,\pi^b}}|r_h(s,a)-\widehat{r}_h(s,a)|\Big\}\\&\overset{(2)}{\leqslant}\frac{1}{\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}}}\frac{1}{M}\sum_{h\in[H]}\max\Big\{\sum_{(s,a)\in\mathcal{Z}_h^{p,\pi^b}}\rho_h^{p,\pi^b}(s,a)|r_h(s,a)-\widehat{r}_h(s,a)|,\max_{(s,a)\notin\mathcal{Z}_h^{p,\pi^b}}|r_h(s,a)-\widehat{r}_h(s,a)|\Big\}\\&=\frac{1}{\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}}}\frac{1}{M}\sum_{h\in[H]}\max\Big\{\mathbb{E}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}|r_h(s,a)-\widehat{r}_h(s,a)|,\max_{(s,a)\notin\mathcal{Z}_h^{p,\pi^b}}|r_h(s,a)-\widehat{r}_h(s,a)|\Big\}\\&\overset{(3)}{\leqslant}\frac{1}{\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}}}\frac{1}{M}\sum_{h\in[H]}\left(\mathbb{E}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}|r_h(s,a)-\widehat{r}_h(s,a)|+\max_{(s,a)\notin\mathcal{Z}_h^{p,\pi^b}}|r_h(s,a)-\widehat{r}_h(s,a)|\right)\\&=:\frac{1}{\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}}}d(r,\widehat{r}),\end{aligned}$$

where at (1) we have upper bounded $\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}}\leqslant\rho_h^{p,\pi^b}(s,a)$ for $(s,a)\in\mathcal{Z}_h^{p,\pi^b}$, and $\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}}\leqslant 1$, and at (2) and (3) we have used that $\max\{a,b\}\leqslant a+b$ for $a,b\geqslant 0$. □

F. Sample Complexity Analysis

In this section, we present our results on the sample complexity of **IRLO** (Algorithm 1 - **IRLO** box) and **PIRLO** (Algorithm 1 - **PIRLO** box) with both distances d and d_∞.

The section is organized into various subsections. We begin with Section F.1, in which we present the concentration results that will be used in all the sample complexity proofs. Section F.2 contains the proofs of sample complexity of **IRLO** w.r.t. distances d and d_∞. Analogously, Section F.3 contains the proofs of sample complexity of **PIRLO** w.r.t. distances d and d_∞. In Section F.4, we present additional sample complexity results for **PIRLO** w.r.t. d, d_∞ under additional requirements. We conclude with Section F.5, in which we present a result of sample complexity on the estimation problem of the superset only, as defined in Equation (8).

F.1. Concentration Lemmas

We define *good event* $\mathcal{E}$ as the intersection of four events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4$. Events $\mathcal{E}_1$ and $\mathcal{E}_2$ allow to obtain exact estimates of $\mathcal{Z}^{p,\pi^b}$ and $\mathcal{S}^{p,\pi^E}$ w.h.p., while events $\mathcal{E}_3$ and $\mathcal{E}_4$ allow to concentrate the estimates of the transition models around their means.

Lemma F.1 (Coverage Events)**.** *Let $\mathcal{M}$ be an MDP without reward and let π^E be the expert's policy. Let $\mathcal{D}^b = \{\langle s_h^{b,i}, a_h^{b,i}\rangle_{h\in[\![H]\!]}\}_{i\in[\![\tau^b]\!]}$ and $\mathcal{D}^E = \{\langle s_h^{E,j}, a_h^{E,j}\rangle_{h\in[\![H]\!]}\}_{j\in[\![\tau]\!]^E}$ be datasets of τ^b and τ^E trajectories collected by executing policies π^b and π^E in $\mathcal{M}$. Denote with $N_h^b(s,a)$ the visitation count of triple $(s,a,h)\in\mathcal{Z}^{p,\pi^b}$ computed using $\mathcal{D}^b$, and by $N_h^E(s,a)$ the analogous for $\mathcal{D}^E$. For any $\delta\in(0,1)$, define events $\mathcal{E}_1, \mathcal{E}_2$ as:*

$$\mathcal{E}_1 := \left\{ N_h^b(s,a) \geqslant 1,\ \forall (s,a,h)\in\mathcal{Z}^{p,\pi^b} \quad \textit{when } \tau^b \geqslant c_1 \frac{\ln \frac{|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\ln \frac{1}{1-\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}}}} \right\},$$

$$\mathcal{E}_2 := \left\{ N_h^E(s,a) \geqslant 1,\ \forall (s,a,h)\in\mathcal{Z}^{p,\pi^E} \quad \textit{when } \tau^E \geqslant c_2 \frac{\ln \frac{|\mathcal{S}^{p,\pi^E}|}{\delta}}{\ln \frac{1}{1-\rho_{\min}^{\pi^E,\mathcal{Z}^{p,\pi^E}}}} \right\},$$

where c_1 and c_2 are universal constants. Then, event $\mathcal{E}_1\cap\mathcal{E}_2$ holds with probability at least $1-\delta/2$.

Proof. Let us begin with event $\mathcal{E}_1$. We observe that $N_h^b(s,a)\sim \text{Bin}(\tau^b, \rho_h^{p,\pi^b}(s,a))$. In an analogous manner as Lemma E.5 of Metelli et al. (2023), we can write:

$$\begin{aligned}
\mathop{\mathbb{P}}_{p,\pi^b}(\mathcal{E}_1^\complement) &= \mathop{\mathbb{P}}_{p,\pi^b}(\exists (s,a,h)\in\mathcal{Z}^{p,\pi^b} : N_h^b(s,a)=0) \\
&\overset{(1)}{\leqslant} \sum_{(s,a,h)\in\mathcal{Z}^{p,\pi^b}} \mathop{\mathbb{P}}_{p,\pi^b}(N_h^b(s,a)=0) \\
&= \sum_{(s,a,h)\in\mathcal{Z}^{p,\pi^b}} (1-\rho_h^{p,\pi^b}(s,a))^{\tau^b} \\
&\overset{(2)}{\leqslant} \sum_{(s,a,h)\in\mathcal{Z}^{p,\pi^b}} (1-\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}})^{\tau^b} \\
&= |\mathcal{Z}^{p,\pi^b}|(1-\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}})^{\tau^b} \leqslant \frac{\delta}{4},
\end{aligned}$$

where at (1) we used a union bound, and at (2) the definition of $\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}} := \min_{(s,a,h)\in\mathcal{Z}^{p,\pi^b}} \rho_h^{p,\pi^b}(s,a)$. Solving w.r.t. τ^b we get:

$$\tau^b \geqslant \frac{\ln \frac{4|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\ln \frac{1}{1-\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}}}},$$

from which the bound for event $\mathcal{E}_1$ holds for some uninteresting constant c_1.

Observe that, by following a similar reasoning, we can prove the bound for event $\mathcal{E}_2$, by recalling that, by hypothesis, the expert's policy is deterministic, so $|\mathcal{Z}^{p,\pi^E}| = |\mathcal{S}^{p,\pi^E}|$. The statement of the theorem follows by an application of the union bound. □

Before presenting the next lemma, we introduce the symbols:

$$b_h(s,a) := \sqrt{\frac{2\beta(N_h^b(s,a),\delta)}{\max\{N_h^b(s,a),1\}}}, \tag{16}$$

and

$$\beta(n,\delta) := \ln(4Z^{p,\pi^b}/\delta) + (S_{\max}^{p,\pi^b} - 1)\ln(e(1+n/(S_{\max}^{p,\pi^b}-1))), \tag{17}$$

with $Z^{p,\pi^b} := |\mathcal{Z}^{p,\pi^b}|$ and $S_{\max}^{p,\pi^b} := \max_{h\in[\![H]\!]} |\mathcal{S}_h^{p,\pi^b}|$. The corresponding counterparts with the estimated quantities are given as follows:

$$\widehat{b}_h(s,a) := \sqrt{\frac{2\widehat{\beta}(N_h^b(s,a),\delta)}{\max\{N_h^b(s,a),1\}}}, \tag{18}$$

and

$$\widehat{\beta}(n,\delta) := \ln(4\widehat{Z}^{p,\pi^b}/\delta) + (\widehat{S}_{\max}^{p,\pi^b} - 1)\ln(e(1+n/(\widehat{S}_{\max}^{p,\pi^b}-1))), \tag{19}$$

with $\widehat{Z}^{p,\pi^b} := |\widehat{\mathcal{Z}}^{p,\pi^b}|$ and $\widehat{S}_{\max}^{p,\pi^b} := \max_{h\in[\![H]\!]} |\widehat{\mathcal{S}}_h^{p,\pi^b}|$. Clearly, under the good event $\mathcal{E}_1 \cap \mathcal{E}_2$, the two versions coincide.

Lemma F.2 (Concentration)**.** *Let $\mathcal{M}$ be an MDP without reward and let π^E be the expert's policy. Let $\mathcal{D}^b = \{\langle s_h^{b,i}, a_h^{b,i}\rangle_{h\in[\![H]\!]}\}_{i\in[\![\tau^b]\!]}$ and $\mathcal{D}^E = \{\langle s_h^{E,j}, a_h^{E,j}\rangle_{h\in[\![H]\!]}\}_{j\in[\![\tau]\!]^E}$ be datasets of τ^b and τ^E trajectories collected by executing policies π^b and π^E in $\mathcal{M}$. Denote with $\widehat{p}_h(\cdot|s,a)$ the empirical transition model of triple $(s,a,h)\in\mathcal{Z}^{p,\pi^b}$ computed using $\mathcal{D}^b$. For any $\delta\in(0,1)$, define events $\mathcal{E}_3, \mathcal{E}_4$ as:*

$$\mathcal{E}_3 := \left\{ N_h^b(s,a) KL(\widehat{p}_h(\cdot|s,a)\|p_h(\cdot|s,a)) \leqslant \beta(N_h^b(s,a),\delta), \quad \forall \tau^b\in\mathbb{N}, \forall (s,a,h)\in\mathcal{Z}^{p,\pi^b} \right\},$$

$$\mathcal{E}_4 := \left\{ \frac{1}{N_h^b(s,a)\vee 1} \leqslant c_4 \frac{\ln\frac{|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\tau^b \rho_h^{p,\pi^b}(s,a)}, \quad \forall (s,a,h)\in\mathcal{Z}^{p,\pi^b} \right\},$$

where c_4 is a universal constant. Then, event $\mathcal{E}_3\cap\mathcal{E}_4$ holds with probability at least $1-\delta/2$.

Proof. We show that both events $\mathcal{E}_3, \mathcal{E}_4$ hold with probability at least $1-\frac{\delta}{4}$. The thesis follows through the application of a union bound.

Let us begin with event $\mathcal{E}_3$. Similarly to the proof of Lemma 10 in Kaufmann et al. (2021), we apply Lemma J.2 and a union bound to get:

$$\begin{aligned}
\mathop{\mathbb{P}}_{p,\pi^b}(\mathcal{E}_3^{\complement}) &= \mathop{\mathbb{P}}_{p,\pi^b}\left(\exists (s,a,h)\in\mathcal{Z}^{p,\pi^b}, \exists \tau^b\in\mathbb{N}\colon N_h^b(s,a) KL(\widehat{p}_h(\cdot|s,a)\|p_h(\cdot|s,a)) > \beta(N_h^b(s,a),\delta) \right) \\
&\leqslant \sum_{(s,a,h)\in\mathcal{Z}^{p,\pi^b}} \mathop{\mathbb{P}}_{p,\pi^b}\left(\exists \tau^b\in\mathbb{N}\colon N_h^b(s,a) KL(\widehat{p}_h(\cdot|s,a)\|p_h(\cdot|s,a)) > \beta(N_h^b(s,a),\delta) \right) \\
&\leqslant \sum_{(s,a,h)\in\mathcal{Z}^{p,\pi^b}} \frac{\delta}{4|\mathcal{Z}^{p,\pi^b}|} = \frac{\delta}{4}.
\end{aligned}$$

It should be remarked that, in the definition of β (Equation 17), we have used $|\mathcal{S}_{\max}^{p,\pi^b}|$ instead of $\mathcal{S}$ because it better represents the support of the transition model in triples $(s,a,h)\in\mathcal{Z}^{p,\pi^b}$.

As far as event $\mathcal{E}_4$ is concerned, consider an arbitrary triple $(s,a,h)\in\mathcal{Z}^{p,\pi^b}$. Observe that the visitation count $N_h^b(s,a)$ is binomially distributed, i.e., $N_h^b(s,a)\sim\text{Bin}(\tau,\rho_h^{p,\pi^b}(s,a))$. Therefore, similarly to Lemma B.1 of (Xie et al., 2021), by applying Lemma J.1 with confidence $\delta/(4|\mathcal{Z}^{p,\pi^b}|)$, we can concentrate the binomial as:

$$\frac{\rho_h^{p,\pi^b}(s,a)}{N_h^b(s,a)\vee 1}\leqslant\frac{8\ln\frac{4|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\tau},$$

from which we get:

$$\mathop{\mathbb{P}}_{p,\pi^b}\left(\frac{1}{N_h^b(s,a)\vee 1}\leqslant\frac{8\ln\frac{4|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\tau\rho_h^{p,\pi^b}(s,a)}\right)\geqslant 1-\frac{\delta}{4|\mathcal{Z}^{p,\pi^b}|}.$$

We can perform a union bound over $(s,a,h)\in\mathcal{Z}^{p,\pi^b}$ to get:

$$\mathop{\mathbb{P}}_{p,\pi^b}\left(\exists(s,a,h)\in\mathcal{Z}^{p,\pi^b}:\frac{1}{N_h^b(s,a)\vee 1}>\frac{8\ln\frac{4|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\tau\rho_h^{p,\pi^b}(s,a)}\right)\leqslant\frac{\delta}{4}.$$

By choosing c_4 appropriately, we get the result. □

Since $\mathcal{E}:=\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\cap\mathcal{E}_4$, then, by combining Lemma F.1 with Lemma F.2 through a union bound, we get that $\mathcal{E}$ holds w.p. $1-\delta$.

F.2. Proof of Theorem 5.1 and Theorem 5.2

The next lemmas show that, for any reward in $\mathcal{R}^{\cap}_{p,\pi^E}$ ($\mathcal{R}^{\cup}_{p,\pi^E}$), it is possible to find a "similar" reward in the estimate $\mathcal{R}^{\cap}_{\widehat{p},\widehat{\pi}^E}$ ($\mathcal{R}^{\cup}_{\widehat{p},\widehat{\pi}^E}$). Notice that, under events $\mathcal{E}_1$ and $\mathcal{E}_2$ we have that, respectively, $\widehat{\mathcal{Z}}^{p,\pi^b}=\mathcal{Z}^{p,\pi^b}$ and $\widehat{\mathcal{S}}^{p,\pi^E}=\mathcal{S}^{p,\pi^E}$. For the sake of simplicity, we provide here the (recursive) definitions of $p^m,p^M\in[p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}$ for any $r\in\mathfrak{R}$:

$$\begin{aligned}
p^M&:=\begin{cases}p_h^M(\cdot|s,a)=p_h(\cdot|s,a), & \forall(s,a,h)\in\mathcal{Z}^{p,\pi^b}\\ p_h^M(\cdot|s,a)=\mathbb{1}\{\cdot=\arg\max_{s'\in\mathcal{S}}V_{h+1}^{\pi^M}(s';p^M,r\}, & \text{otherwise}\end{cases},\\
p^m&:=\begin{cases}p_h^m(\cdot|s,a)=p_h(\cdot|s,a), & \forall(s,a,h)\in\mathcal{Z}^{p,\pi^b}\\ p_h^m(\cdot|s,a)=\mathbb{1}\{\cdot=\arg\min_{s'\in\mathcal{S}}V_{h+1}^{\pi^m}(s';p^m,r\}, & \text{otherwise}\end{cases},
\end{aligned}\tag{20}$$

where we have used the following (recursive) policy definitions $\pi^M,\widetilde{\pi}^m\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}$:

$$\begin{aligned}
\pi^M&:=\begin{cases}\pi_h^M(s)=\pi_h^E(s),\text{ if }(s,h)\in\mathcal{S}^{p,\pi^E}\\ \pi_h^M(\cdot|s)=\mathbb{1}\{\cdot=\arg\max_{a\in\mathcal{A}}Q_h^{\pi^M}(s,a;p^M,r)\},\text{ if }(s,h)\notin\mathcal{S}^{p,\pi^E}\end{cases},\\
\pi^m&:=\begin{cases}\pi_h^m(s)=\pi_h^E(s),\text{ if }(s,h)\in\mathcal{S}^{p,\pi^E}\\ \pi_h^m(\cdot|s)=\mathbb{1}\{\cdot=\arg\max_{a\in\mathcal{A}}Q_h^{\pi^m}(s,a;p^m,r)\},\text{ if }(s,h)\notin\mathcal{S}^{p,\pi^E}\end{cases}.
\end{aligned}\tag{21}$$

Thanks to these definitions, we can rewrite $\mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E}$ as:

$$\begin{aligned}
\mathcal{R}^{\cap}_{p,\pi^E}&=\{r\in\mathfrak{R}\,|\,\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\backslash\{a^E\}:Q_h^{\pi^E}(s,a^E;p,r)\geqslant Q_h^{\pi^M}(s,a;p^M,r)\},\\
\mathcal{R}^{\cup}_{p,\pi^E}&=\{r\in\mathfrak{R}\,|\,\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\backslash\{a^E\}:Q_h^{\pi^E}(s,a^E;p,r)\geqslant Q_h^{\pi^m}(s,a;p^m,r)\}.
\end{aligned}\tag{22}$$

We will denote by $\widehat{p}^M,\widehat{p}^m,\widehat{\pi}^M,\widehat{\pi}^m$ the transition models and policies defined as in Eq. 20 and Eq. 21 but for transition model $\widehat{p}$.

Lemma F.3 (Reward Choice Subset)**.** *Let $\mathcal{R}^{\cap}_{p,\pi^E}$ be the subset of the feasible set $\mathcal{R}_{p,\pi^E}$ estimated through $\mathcal{R}^{\cap}_{\widehat{p},\widehat{\pi}^E}$ outputted by Algorithm 1. Under event $\mathcal{E}$, for any $r \in \mathcal{R}^{\cap}_{p,\pi^E}$, the reward $\widehat{r}$ constructed as:*

$$\begin{cases} \widehat{r}_h(s,a) = r_h(s,a) + \sum_{s' \in \mathcal{S}} p_h(s'|s,a) V^{\pi^M}_{h+1}(s'; p^M, r) - \sum_{s' \in \mathcal{S}} \widehat{p}_h(s'|s,a) V^{\widehat{\pi}^M}_{h+1}(s'; \widehat{p}^M, \widehat{r}), & \forall (s,a,h) \in \mathcal{Z}^{p,\pi^b} \\ \widehat{r}_h(s,a) = r_h(s,a), & \forall (s,a,h) \notin \mathcal{Z}^{p,\pi^b} \end{cases},$$

belongs to $\mathcal{R}^{\cap}_{\widehat{p},\widehat{\pi}^E}$. Moreover, for any reward $\widehat{r} \in \mathcal{R}^{\cap}_{\widehat{p},\widehat{\pi}^E}$, we can construct in the same manner a reward r that belongs to $\mathcal{R}^{\cap}_{p,\pi^E}$.

Proof. The idea of the proof is to show that $Q^{\pi^M}_h(s,a;p^M,r) = Q^{\widehat{\pi}^M}_h(s,a;\widehat{p}^M,\widehat{r})$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]$. Indeed, in this way, since $r \in \mathcal{R}^{\cap}_{p,\pi^E}$, then it holds that:

$$\forall (s,h) \in \mathcal{S}^{p,\pi^E}, \forall a \in \mathcal{A} \setminus \{a^E\} : Q^{\pi^E}_h(s,a^E;\widehat{p},\widehat{r}) - Q^{\widehat{\pi}^M}_h(s,a;\widehat{p}^M,\widehat{r}) = Q^{\pi^E}_h(s,a^E;p,r) - Q^{\pi^M}_h(s,a;p^M,r) \geqslant 0.$$

Let us begin with any $(s,a,h) \in \mathcal{Z}^{p,\pi^b}$. By definition of $\widehat{r}$ and by rearranging the terms, we obtain:

$$\begin{aligned} &\widehat{r}_h(s,a) + \sum_{s' \in \mathcal{S}} \widehat{p}_h(s'|s,a) V^{\widehat{\pi}^M}_{h+1}(s'; \widehat{p}^M, \widehat{r}) = r_h(s,a) + \sum_{s' \in \mathcal{S}} p_h(s'|s,a) V^{\pi^M}_{h+1}(s'; p^M, r) \\ &\iff Q^{\widehat{\pi}^M}_h(s,a;\widehat{p}^M,\widehat{r}) = Q^{\pi^M}_h(s,a;p^M,r). \end{aligned} \tag{23}$$

In particular, observe that, by Assumption 2.1, it holds $\mathcal{Z}^{p,\pi^E} \subseteq \mathcal{Z}^{p,\pi^b}$; moreover, by definition of $\widehat{p}^M$ and p^M, playing an expert's action from $\mathcal{S}^{p,\pi^E}$ brings again into $\mathcal{S}^{p,\pi^E}$, therefore, for $(s,a^E,h) \in \mathcal{Z}^{p,\pi^E}$, this means:

$$Q^{\pi^E}_h(s,a^E;\widehat{p},\widehat{r}) = Q^{\pi^E}_h(s,a^E;p,r).$$

Now, we show by induction that, for any $(s,a,h) \notin \mathcal{Z}^{p,\pi^b}$, it holds that:

$$Q^{\widehat{\pi}^M}_h(s,a;\widehat{p}^M,\widehat{r}) = Q^{\pi^M}_h(s,a;p^M,r).$$

As case base, consider stage H. Clearly, for any $(s,a) \notin \mathcal{Z}^{p,\pi^b}_H$, we have:

$$\begin{aligned} Q^{\widehat{\pi}^M}_H(s,a;\widehat{p}^M,\widehat{r}) &= \widehat{r}_H(s,a) \\ &= r_H(s,a) \\ &= Q^{\pi^M}_h(s,a;p^M,r), \end{aligned}$$

where we have used the definition of $\widehat{r}$. Make the inductive hypothesis that, at stage $h+1$, for any $(s,a) \notin \mathcal{Z}^{p,\pi^b}_{h+1}$, it holds that $Q^{\widehat{\pi}^M}_{h+1}(s,a;\widehat{p}^M,\widehat{r}) = Q^{\pi^M}_{h+1}(s,a;p^M,r)$, and consider stage h:

$$\begin{aligned} Q^{\widehat{\pi}^M}_h(s,a;\widehat{p}^M,\widehat{r}) &= \widehat{r}_h(s,a) + \sum_{s' \in \mathcal{S}} \widehat{p}^M_h(s'|s,a) V^{\widehat{\pi}^M}_{h+1}(s';\widehat{p}^M,\widehat{r}) \\ &\overset{(1)}{=} r_h(s,a) + \sum_{s' \in \mathcal{S}} \widehat{p}^M_h(s'|s,a) V^{\widehat{\pi}^M}_{h+1}(s';\widehat{p}^M,\widehat{r}) \\ &\overset{(2)}{=} r_h(s,a) + \max_{s' \in \mathcal{S}} V^{\widehat{\pi}^M}_{h+1}(s';\widehat{p}^M,\widehat{r}) \\ &= r_h(s,a) + \max\Big\{ \max_{s' \in \mathcal{S}^{p,\pi^b}_{h+1}} V^{\widehat{\pi}^M}_{h+1}(s';\widehat{p}^M,\widehat{r}), \max_{s' \notin \mathcal{S}^{p,\pi^b}_{h+1}} V^{\widehat{\pi}^M}_{h+1}(s';\widehat{p}^M,\widehat{r}) \Big\} \\ &\overset{(3)}{=} r_h(s,a) + \max\Big\{ \max_{s' \in \mathcal{S}^{p,\pi^b}_{h+1}} V^{\pi^M}_{h+1}(s';p^M,r), \max_{s' \notin \mathcal{S}^{p,\pi^b}_{h+1}} V^{\widehat{\pi}^M}_{h+1}(s';\widehat{p}^M,\widehat{r}) \Big\} \\ &\overset{(4)}{=} r_h(s,a) + \max\Big\{ \max_{s' \in \mathcal{S}^{p,\pi^b}_{h+1}} V^{\pi^M}_{h+1}(s';p^M,r), \max_{s' \notin \mathcal{S}^{p,\pi^b}_{h+1}} V^{\pi^M}_{h+1}(s';p^M,r) \Big\} \end{aligned}$$

$$
\begin{aligned}
&= r_h(s,a) + \max_{s' \in \mathcal{S}} V_{h+1}^{\pi^M}(s'; p^M, r) \\
&= Q_h^{\pi^M}(s,a; p^M, r),
\end{aligned}
$$

where at (1) we use the definition of $\hat{r}$ outside $\mathcal{Z}^{p,\pi^b}$, at (2) we use the definition of $\hat{p}^M$, at (3) we use Eq. 23, and at (4) we use the inductive hypothesis.

Notice that the same passages can be carried out if we exchanged p and $\hat{p}$. This concludes the proof. □

Notice that the reward function chosen in Lemma F.3 can be interpreted, in an analogous manner as in the proof of Theorem 3.1 of Metelli et al. (2021), as the reward that provides, in transition model $\hat{p}$, the same Q-function provided by the given reward in p.

We can prove an analogous result for the superset $\mathcal{R}^{\cup}_{\hat{p},\hat{\pi}^E}$.

Lemma F.4 (Reward Choice Superset)**.** *Let $\mathcal{R}^{\cup}_{p,\pi^E}$ be the subset of the feasible set $\mathcal{R}_{p,\pi^E}$ estimated through $\mathcal{R}^{\cup}_{\hat{p},\hat{\pi}^E}$ outputted by Algorithm 1. Under event $\mathcal{E}$, for any $r \in \mathcal{R}^{\cup}_{p,\pi^E}$, the reward $\hat{r}$ constructed as:*

$$
\begin{cases}
\hat{r}_h(s,a) = r_h(s,a) + \sum_{s' \in \mathcal{S}} p_h(s'|s,a) V_{h+1}^{\pi^m}(s'; p^m, r) - \sum_{s' \in \mathcal{S}} \hat{p}_h(s'|s,a) V_{h+1}^{\hat{\pi}^m}(s'; \hat{p}^m, \hat{r}), & \forall (s,a,h) \in \mathcal{Z}^{p,\pi^b} \\
\hat{r}_h(s,a) = r_h(s,a), & \forall (s,a,h) \notin \mathcal{Z}^{p,\pi^b}
\end{cases},
$$

belongs to $\mathcal{R}^{\cup}_{\hat{p},\hat{\pi}^E}$. Moreover, for any reward $\hat{r} \in \mathcal{R}^{\cup}_{\hat{p},\hat{\pi}^E}$, we can construct in the same manner a reward r that belongs to $\mathcal{R}^{\cup}_{p,\pi^E}$.

Proof. The idea of the proof is analogous to that of Lemma F.3, and is reported here for completeness. We aim to show that $Q_h^{\pi^m}(s,a; p^m, r) = Q_h^{\hat{\pi}^m}(s,a; \hat{p}^m, \hat{r})$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]$. Indeed, in this way, since $r \in \mathcal{R}^{\cup}_{p,\pi^E}$, then it holds that:

$$
\forall (s,h) \in \mathcal{S}^{p,\pi^E}, \forall a \in \mathcal{A} \setminus \{a^E\} : Q_h^{\pi^E}(s, a^E; \hat{p}, \hat{r}) - Q_h^{\hat{\pi}^m}(s,a; \hat{p}^m, \hat{r}) = Q_h^{\pi^E}(s, a^E; p, r) - Q_h^{\pi^m}(s,a; p^m, r) \geqslant 0.
$$

Let us begin with any $(s,a,h) \in \mathcal{Z}^{p,\pi^b}$. By definition of $\hat{r}$ and by rearranging the terms, we obtain:

$$
\begin{aligned}
&\hat{r}_h(s,a) + \sum_{s' \in \mathcal{S}} \hat{p}_h(s'|s,a) V_{h+1}^{\hat{\pi}^m}(s'; \hat{p}^m, \hat{r}) = r_h(s,a) + \sum_{s' \in \mathcal{S}} p_h(s'|s,a) V_{h+1}^{\pi^m}(s'; p^m, r) \\
&\iff Q_h^{\hat{\pi}^m}(s,a; \hat{p}^m, \hat{r}) = Q_h^{\pi^m}(s,a; p^m, r).
\end{aligned} \tag{24}
$$

In particular, observe that, by Assumption 2.1, it holds $\mathcal{Z}^{p,\pi^E} \subseteq \mathcal{Z}^{p,\pi^b}$; moreover, by definition of $\hat{p}^m$ and p^m, playing an expert's action from $\mathcal{S}^{p,\pi^E}$ brings again into $\mathcal{S}^{p,\pi^E}$, therefore, for $(s, a^E, h) \in \mathcal{Z}^{p,\pi^E}$, this means:

$$
Q_h^{\pi^E}(s, a^E; \hat{p}, \hat{r}) = Q_h^{\pi^E}(s, a^E; p, r).
$$

Now, we show by induction that, for any $(s,a,h) \notin \mathcal{Z}^{p,\pi^b}$, it holds that:

$$
Q_h^{\hat{\pi}^m}(s,a; \hat{p}^m, \hat{r}) = Q_h^{\pi^m}(s,a; p^m, r).
$$

As case base, consider stage H. Clearly, for any $(s,a) \notin \mathcal{Z}_H^{p,\pi^b}$, we have:

$$
\begin{aligned}
Q_H^{\hat{\pi}^m}(s,a; \hat{p}^m, \hat{r}) &= \hat{r}_H(s,a) \\
&= r_H(s,a) \\
&= Q_h^{\pi^m}(s,a; p^m, r),
\end{aligned}
$$

where we have used the definition of $\hat{r}$. Make the inductive hypothesis that, at stage $h+1$, for any $(s,a) \notin \mathcal{Z}_{h+1}^{p,\pi^b}$, it holds that $Q_{h+1}^{\hat{\pi}^m}(s,a; \hat{p}^m, \hat{r}) = Q_{h+1}^{\pi^m}(s,a; p^m, r)$, and consider stage h:

$$
Q_h^{\hat{\pi}^m}(s,a; \hat{p}^m, \hat{r}) = \hat{r}_h(s,a) + \sum_{s' \in \mathcal{S}} \hat{p}_h^m(s'|s,a) V_{h+1}^{\hat{\pi}^m}(s'; \hat{p}^m, \hat{r})
$$

$$
\begin{aligned}
&\overset{(1)}{=} r_h(s,a) + \sum_{s' \in \mathcal{S}} \widehat{p}_h^m(s'|s,a) V_{h+1}^{\widehat{\pi}^m}(s'; \widehat{p}^m, \widehat{r}) \\
&\overset{(2)}{=} r_h(s,a) + \min_{s' \in \mathcal{S}} V_{h+1}^{\widehat{\pi}^m}(s'; \widehat{p}^m, \widehat{r}) \\
&= r_h(s,a) + \min\Big\{ \min_{s' \in \mathcal{S}_{h+1}^{p,\pi^b}} V_{h+1}^{\widehat{\pi}^m}(s'; \widehat{p}^m, \widehat{r}), \min_{s' \notin \mathcal{S}_{h+1}^{p,\pi^b}} V_{h+1}^{\widehat{\pi}^m}(s'; \widehat{p}^m, \widehat{r}) \Big\} \\
&\overset{(3)}{=} r_h(s,a) + \min\Big\{ \min_{s' \in \mathcal{S}_{h+1}^{p,\pi^b}} V_{h+1}^{\pi^m}(s'; p^m, r), \min_{s' \notin \mathcal{S}_{h+1}^{p,\pi^b}} V_{h+1}^{\widehat{\pi}^m}(s'; \widehat{p}^m, \widehat{r}) \Big\} \\
&\overset{(4)}{=} r_h(s,a) + \min\Big\{ \min_{s' \in \mathcal{S}_{h+1}^{p,\pi^b}} V_{h+1}^{\pi^m}(s'; p^m, r), \min_{s' \notin \mathcal{S}_{h+1}^{p,\pi^b}} V_{h+1}^{\pi^m}(s'; p^m, r) \Big\} \\
&= r_h(s,a) + \max_{s' \in \mathcal{S}} V_{h+1}^{\pi^m}(s'; p^m, r) \\
&= Q_h^{\pi^m}(s,a; p^m, r),
\end{aligned}
$$

where at (1) we use the definition of $\hat{r}$ outside $\mathcal{Z}^{p,\pi^b}$, at (2) we use the definition of $\widehat{p}^m$, at (3) we use Eq. 24, and at (4) we use the inductive hypothesis.

Notice that the same passages can be carried out if we exchanged p and $\widehat{p}$. This concludes the proof. □

From the proofs of Lemma F.3 and Lemma F.4, we notice that proving the result for the superset is "easier", because we simply have to consider a single transition model; instead, for the subset, we have to consider all the transition models in the equivalence class. Luckily, we can single out a "worst" transition model from this class and provide the proof only for it. We will see in the proofs of the results with pessimism how to cope with the trickier problem in which there exist many "worst" transition models, and thus the recursion cannot be applied directly.

Thanks to Lemma F.3 and Lemma F.4, we can upper bound the distance between sets of rewards by a term that depends on the distance between the transition models. We do not have error for the policy because, under good event $\mathcal{E}$, we have that $\widehat{\pi}^E = \pi^E$ in $\widehat{\mathcal{S}}^{p,\pi^E} = \mathcal{S}^{p,\pi^E}$.

Lemma F.5 (Performance Decomposition Subset and Superset)**.** *Let $\widehat{\mathcal{R}}^{\cap} := \mathcal{R}^{\cap}_{\widehat{p},\widehat{\pi}^E}$ and $\widehat{\mathcal{R}}^{\cup} := \mathcal{R}^{\cup}_{\widehat{p},\widehat{\pi}^E}$ be the output of IRLO (Algorithm 1). Under the good event $\mathcal{E}$, it holds that:*

$$
\mathcal{H}_d(\mathcal{R}^{\cap}_{p,\pi^E}, \widehat{\mathcal{R}}^{\cap}) \leqslant H \sum_{h \in [\![H]\!]} \mathop{\mathbb{E}}_{(s,a) \sim \rho_h^{p,\pi^b}(\cdot,\cdot)} b_h(s,a),
$$

and:

$$
\mathcal{H}_d(\mathcal{R}^{\cup}_{p,\pi^E}, \widehat{\mathcal{R}}^{\cup}) \leqslant H \sum_{h \in [\![H]\!]} \mathop{\mathbb{E}}_{(s,a) \sim \rho_h^{p,\pi^b}(\cdot,\cdot)} b_h(s,a).
$$

Proof. By Definition 4.2, we can write:

$$
\begin{aligned}
\mathcal{H}(\mathcal{R}^{\cap}_{p,\pi^E}, \widehat{\mathcal{R}}^{\cap}) &:= \max\Big\{ \sup_{r \in \mathcal{R}^{\cap}_{p,\pi^E}} \inf_{\hat{r} \in \widehat{\mathcal{R}}^{\cap}} d(r,\hat{r}), \sup_{\hat{r} \in \widehat{\mathcal{R}}^{\cap}} \inf_{r \in \mathcal{R}^{\cap}_{p,\pi^E}} d(r,\hat{r}) \Big\} \\
&\overset{(1)}{\leqslant} \max\Big\{ \sup_{r \in \mathcal{R}^{\cap}_{p,\pi^E}} d(r, \widetilde{r}^1), \sup_{\hat{r} \in \widehat{\mathcal{R}}^{\cap}} d(\widetilde{r}^2, \hat{r}) \Big\} \\
&\overset{(2)}{=} \max\Bigg\{ \sup_{r \in \mathcal{R}^{\cap}_{p,\pi^E}} \frac{1}{M} \sum_{h \in [\![H]\!]} \Bigg(\mathop{\mathbb{E}}_{(s,a) \sim \rho_h^{p,\pi^b}(\cdot,\cdot)} \big| r_h(s,a) - \widetilde{r}_h^1(s,a) \big| + \underbrace{\max_{(s,a) \notin \mathcal{Z}_h^{p,\pi^b}} \big| r_h(s,a) - \widetilde{r}_h^1(s,a) \big|}_{=0} \Bigg), \\
&\qquad \sup_{\hat{r} \in \widehat{\mathcal{R}}^{\cap}} \frac{1}{M} \sum_{h \in [\![H]\!]} \Bigg(\mathop{\mathbb{E}}_{(s,a) \sim \rho_h^{p,\pi^b}(\cdot,\cdot)} \big| \hat{r}_h(s,a) - \widetilde{r}_h^2(s,a) \big| + \underbrace{\max_{(s,a) \notin \mathcal{Z}_h^{p,\pi^b}} \big| \hat{r}_h(s,a) - \widetilde{r}_h^2(s,a) \big|}_{=0} \Bigg) \Bigg\}
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(3)}{=}\max\Bigg\{\sup_{r\in\mathcal{R}^{\cap}_{p,\pi^E}}\frac{1}{M}\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}\Big|\sum_{s'\in\mathcal{S}}(p_h(s'|s,a)-\widehat{p}_h(s'|s,a))V_{h+1}^{\pi^M}(s';p^M,r)\Big|,\\
&\qquad\sup_{\widehat{r}\in\widehat{\mathcal{R}}^{\cap}}\frac{1}{M}\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}\Big|\sum_{s'\in\mathcal{S}}(\widehat{p}_h(s'|s,a)-p_h(s'|s,a))V_{h+1}^{\widehat{\pi}^M}(s';\widehat{p}^M,\widehat{r})\Big|\Bigg\}\\
&\overset{(4)}{\leqslant}\max\Bigg\{\sup_{r\in\mathcal{R}^{\cap}_{p,\pi^E}}\frac{1}{M}\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}\sum_{s'\in\mathcal{S}}\Big|(p_h(s'|s,a)-\widehat{p}_h(s'|s,a))V_{h+1}^{\pi^M}(s';p^M,r)\Big|,\\
&\qquad\sup_{\widehat{r}\in\widehat{\mathcal{R}}^{\cap}}\frac{1}{M}\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}\sum_{s'\in\mathcal{S}}\Big|(\widehat{p}_h(s'|s,a)-p_h(s'|s,a))V_{h+1}^{\widehat{\pi}^M}(s';\widehat{p}^M,\widehat{r})\Big|\Bigg\}\\
&\overset{(5)}{\leqslant}\max\Bigg\{\sup_{r\in\mathcal{R}^{\cap}_{p,\pi^E}}\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}\sum_{s'\in\mathcal{S}}\Big|(p_h(s'|s,a)-\widehat{p}_h(s'|s,a))H\Big|,\\
&\qquad\sup_{\widehat{r}\in\widehat{\mathcal{R}}^{\cap}}\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}\sum_{s'\in\mathcal{S}}\Big|(\widehat{p}_h(s'|s,a)-p_h(s'|s,a))H\Big|\Bigg\}\\
&=\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}\sum_{s'\in\mathcal{S}}\Big|(p_h(s'|s,a)-\widehat{p}_h(s'|s,a))H\Big|\\
&=\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}H\big\|p_h(\cdot|s,a)-\widehat{p}_h(\cdot|s,a)\big\|_1\\
&\overset{(6)}{\leqslant}\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}H\sqrt{2KL\big(p_h(\cdot|s,a)\|\widehat{p}_h(\cdot|s,a)\big)}\\
&\overset{(7)}{\leqslant}\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}H\sqrt{2\frac{\beta(N_h^b(s,a),\delta)}{N_h^b(s,a)}}\\
&=H\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}b_h(s,a),
\end{aligned}
$$

where at (1) we have applied Lemma F.3, denoting by $\widetilde{r}^1\in\widehat{\mathcal{R}}^{\cap}$ and $\widetilde{r}^2\in\mathcal{R}^{\cap}_{p,\pi^E}$ the choices of rewards, and used that $\inf_{x\in\mathcal{X}}f(x)\leqslant f(\bar{x})$ for any $\bar{x}\in\mathcal{X}$; at (2) we have used the definition of distance d, at (3) we have inserted the definitions of $\widetilde{r}^1$ and $\widetilde{r}^2$ as provided by Lemma F.3, in particular using that $Q_h^{\pi^M}(s,a;p^M,r)=Q_h^{\widehat{\pi}^M}(s,a;\widehat{p}^M,\widehat{r})$, at (4) we have applied triangle inequality, to bring the absolute value inside the summation, at (5) we upper bound $V_{h+1}^{\pi^M}(s';p^M,r)\leqslant H\|r\|_\infty$ and $V_{h+1}^{\widehat{\pi}^M}(s';\widehat{p}^M,\widehat{r})\leqslant H\|\widehat{r}\|_\infty$, and since $M:=1/\max\{\|r\|_\infty,\|\widehat{r}\|_\infty\}$, we obtain $H\|r\|_\infty/M\leqslant H$ and $H\|\widehat{r}\|_\infty/M\leqslant H$; at (6) we have applied Pinsker's inequality, and at (7) we have applied the bound of $\mathcal{E}_3\supseteq\mathcal{E}$ noticing that $N_h^b(s,a)\geqslant 1$ because of event $\mathcal{E}_1\supseteq\mathcal{E}$.

A similar procedure can be carried out also for the supersets, using Lemma F.4 instead of Lemma F.3. □

Finally, we have all the tools we need to prove Theorem 5.1.

Theorem 5.1. *Let $\mathcal{M}$ be an MDP without reward and let π^E be the expert's policy. Let $\mathcal{D}^E$ and $\mathcal{D}^b$ be two datasets of τ^E and τ^b trajectories collected with policies π^E and π^b in $\mathcal{M}$, respectively. Under Assumption 2.1, **IRLO** is (ϵ,δ)-PAC for d-IRL with a sample complexity at most:*

$$
\tau^b\leqslant\widetilde{\mathcal{O}}\left(\frac{H^3Z^{p,\pi^b}\ln\frac{1}{\delta}}{\epsilon^2}\left(\ln\frac{1}{\delta}+S_{\max}^{p,\pi^b}\right)+\frac{\ln\frac{1}{\delta}}{\ln\frac{1}{1-\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}}}}\right),
$$

$$
\tau^E\leqslant\widetilde{\mathcal{O}}\left(\frac{\ln\frac{1}{\delta}}{\ln\frac{1}{1-\rho_{\min}^{\pi^E,\mathcal{Z}^{p,\pi^E}}}}\right).
$$

Proof. First, observe that, thanks to Lemma F.1 and Lemma F.2, we have that good event $\mathcal{E}$ holds w.p. $1-\delta$ with a number of trajectories τ^E and τ^b upper bounded as in events $\mathcal{E}_1$ and $\mathcal{E}_2$. Now, under good event $\mathcal{E}$, the idea of the proof is to compute

the number of trajectories τ^b needed to have a distance between sets of rewards smaller than ϵ. Then, we combine it with the number of trajectories required by event $\mathcal{E}$ through $\max\{a,b\} \leqslant a+b$ for $a,b \geqslant 0$.

Let us begin with the subset. Thanks to Lemma F.5, we can write:

$$
\begin{aligned}
\mathcal{H}_d(\mathcal{R}^{\cap}_{p,\pi^E}, \widehat{\mathcal{R}}^{\cap}) &\leqslant H \sum_{h\in[\![H]\!]} \mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} b_h(s,a) \\
&= H \sum_{h\in[\![H]\!]} \sum_{(s,a)\in\mathcal{Z}_h^{p,\pi^b}} \rho_h^{p,\pi^b}(s,a) b_h(s,a) \\
&= H \sum_{h\in[\![H]\!]} \sum_{(s,a)\in\mathcal{Z}_h^{p,\pi^b}} \rho_h^{p,\pi^b}(s,a) \sqrt{2\frac{\beta(N_h^b(s,a),\delta)}{N_h^b(s,a)}} \\
&\overset{(1)}{\leqslant} \sqrt{2} H \sum_{h\in[\![H]\!]} \sum_{(s,a)\in\mathcal{Z}_h^{p,\pi^b}} \rho_h^{p,\pi^b}(s,a) \sqrt{\frac{\beta(\tau^b,\delta)}{N_h^b(s,a)}} \\
&= \sqrt{2\beta(\tau^b,\delta)} H \sum_{h\in[\![H]\!]} \sum_{(s,a)\in\mathcal{Z}_h^{p,\pi^b}} \rho_h^{p,\pi^b}(s,a) \sqrt{\frac{1}{N_h^b(s,a)}} \\
&\overset{(2)}{\leqslant} \sqrt{2\beta(\tau^b,\delta)} H \sum_{h\in[\![H]\!]} \sum_{(s,a)\in\mathcal{Z}_h^{p,\pi^b}} \rho_h^{p,\pi^b}(s,a) \sqrt{c_4 \frac{\ln\frac{|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\tau^b \rho_h^{p,\pi^b}(s,a)}} \\
&\overset{(3)}{=} c_5 \sqrt{\frac{\beta(\tau^b,\delta)\ln\frac{|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\tau^b}} H \sum_{h\in[\![H]\!]} \sum_{(s,a)\in\mathcal{Z}_h^{p,\pi^b}} \sqrt{\rho_h^{p,\pi^b}(s,a)} \\
&\overset{(4)}{\leqslant} c_5 \sqrt{\frac{\beta(\tau^b,\delta)\ln\frac{|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\tau^b}} H \sum_{h\in[\![H]\!]} \sqrt{|\mathcal{Z}_h^{p,\pi^b}|} \sqrt{\sum_{(s,a)\in\mathcal{Z}_h^{p,\pi^b}} \rho_h^{p,\pi^b}(s,a)} \\
&= c_5 \sqrt{\frac{\beta(\tau^b,\delta)\ln\frac{|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\tau^b}} H \sum_{h\in[\![H]\!]} \sqrt{|\mathcal{Z}_h^{p,\pi^b}|} \\
&\overset{(5)}{\leqslant} c_5 \sqrt{\frac{\beta(\tau^b,\delta)\ln\frac{|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\tau^b}} H \sqrt{H|\mathcal{Z}^{p,\pi^b}|} \leqslant \epsilon,
\end{aligned}
$$

where at (1) we have used that $\tau^b \geqslant N_h^b(s,a)$ for all $(s,a,h) \in \mathcal{Z}^{p,\pi^b}$, and that function $\beta(\cdot,\delta)$ is monotonically increasing; at (2) we have applied the result in Lemma F.2 for event $\mathcal{E}_4$, at (3) we define constant $c_5 := \sqrt{2c_4}$, and at (4) and (5) we have applied the Cauchy-Schwarz's inequality.

To compute an upper bound to the number of trajectories required to have $\mathcal{H}_d(\mathcal{R}^{\cap}_{p,\pi^E}, \widehat{\mathcal{R}}^{\cap}) \leqslant \epsilon$, we compute the smallest τ^b that satisfies:

$$
c_5 \sqrt{\frac{\beta(\tau^b,\delta)\ln\frac{|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\tau^b}} H \sqrt{H|\mathcal{Z}^{p,\pi^b}|} \leqslant \epsilon.
$$

By using the definition of $\beta(\tau^b,\delta)$ from Eq. 17 and rearranging the terms, this is equivalent to finding the smallest τ^b such that:

$$
\tau^b \geqslant c_6 \frac{H^3 |\mathcal{Z}^{p,\pi^b}| \ln\frac{|\mathcal{Z}^{p,\pi^b}|}{\delta} \left(\ln\frac{4|\mathcal{Z}^{p,\pi^b}|}{\delta} + (|\mathcal{S}^{p,\pi^b}_{\max}|-1)\ln\left(e(1+\tau^b/(|\mathcal{S}^{p,\pi^b}_{\max}|-1))\right)\right)}{\epsilon^2},
$$

where $c_6 := c_5^2$. If we define:

$$a := c_6 \frac{H^3 |\mathcal{Z}^{p,\pi^b}| \ln \frac{|\mathcal{Z}^{p,\pi^b}|}{\delta} \ln \frac{4|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\epsilon^2},$$
$$b := c_6 \frac{H^3 |\mathcal{Z}^{p,\pi^b}| \ln \frac{|\mathcal{Z}^{p,\pi^b}|}{\delta} (|\mathcal{S}^{p,\pi^b}_{\max}| - 1)}{\epsilon^2},$$
$$c := \frac{e}{|\mathcal{S}^{p,\pi^b}_{\max}| - 1},$$
$$d := e,$$

then we can rewrite the previous expression as:

$$\tau^b \geqslant a + b \ln(c\tau^b + d).$$

To solve it, we can notice that $a, b, c, d > 0$ and $2bc > e$, thus we can apply Lemma J.3[17] to obtain:

$$\tau^b \geqslant 2a + 3b \ln(2bc) + d/c.$$

Replacing a, b, c, d with their values, we get:

$$\begin{aligned}
\tau^b &\geqslant 2c_6 \frac{H^3 |\mathcal{Z}^{p,\pi^b}| \ln \frac{|\mathcal{Z}^{p,\pi^b}|}{\delta} \ln \frac{4|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\epsilon^2} + \left(|\mathcal{S}^{p,\pi^b}_{\max}| - 1\right) \\
&\quad + 3c_6 \frac{H^3 |\mathcal{Z}^{p,\pi^b}| \ln \frac{|\mathcal{Z}^{p,\pi^b}|}{\delta} (|\mathcal{S}^{p,\pi^b}_{\max}| - 1)}{\epsilon^2} \ln \left(2c_6 e \frac{H^3 |\mathcal{Z}^{p,\pi^b}| \ln \frac{|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\epsilon^2} \right) \\
&\leqslant \widetilde{\mathcal{O}} \left(\frac{H^3 |\mathcal{Z}^{p,\pi^b}| \ln \frac{1}{\delta}}{\epsilon^2} \left(\ln \frac{1}{\delta} + |\mathcal{S}^{p,\pi^b}_{\max}| \right) \right).
\end{aligned}$$

Now, observe that an upper bound to the number of trajectories needed to have $\mathcal{H}_d(\mathcal{R}^{\cup}_{p,\pi^E}, \widehat{\mathcal{R}}^{\cup}) \leqslant \epsilon$ can be obtained with an identical derivation, and, thus, it is of the same order. The statement of the theorem follows by the considerations at the beginning of the proof. □

We provide here the proof of Theorem 5.2. The proof is analogous to that of Theorem 5.1.

Theorem 5.2. *Under the conditions of Theorem 5.1,* ***IRLO*** *is (ϵ, δ)-PAC for d_∞-IRL with a sample complexity at most:*

$$\tau^b \leqslant \widetilde{\mathcal{O}} \left(\frac{H^4 \ln \frac{1}{\delta}}{\rho^{\pi^b, \mathcal{Z}^{p,\pi^b}}_{\min} \epsilon^2} \left(\ln \frac{1}{\delta} + S^{p,\pi^b}_{\max} \right) + \frac{\ln \frac{1}{\delta}}{\ln \frac{1}{1 - \rho^{\pi^b, \mathcal{Z}^{p,\pi^b}}_{\min}}} \right),$$

and τ^E is bounded as in Theorem 5.1.

Sketch of proof. The proof is exactly the same as that of Theorem 5.1. First, observe that it is possible to prove a lemma analogous to Lemma F.5, so that:

$$\mathcal{H}_\infty(\mathcal{R}^{\cap}_{p,\pi^E}, \widehat{\mathcal{R}}^{\cap}) \leqslant H \sum_{h \in [\![H]\!]} \max_{(s,a) \in \mathcal{Z}^{p,\pi^b}_h} b_h(s,a).$$

Next, when applying Lemma J.1, we simply notice that, for all $h \in [\![H]\!]$:

$$\max_{(s,a) \in \mathcal{Z}^{p,\pi^b}_h} \sqrt{\frac{1}{\rho^{p,\pi^b}_h(s,a)}} \leqslant \sqrt{\frac{1}{\rho^{\pi^b, \mathcal{Z}^{p,\pi^b}}_{\min}}}.$$

We can do the same also for the superset. Following the derivation in the proof of Theorem 5.1, the result can be obtained. □

[17] It should be remarked that the adoption of Lemma 15 of (Kaufmann et al., 2021) provides the same asymptotical dependence on the quantities of interest. However, Lemma J.3 is more concise.

F.3. Proof of Theorem 6.1 and Theorem 6.2

We will denote by $a^E := \pi^E_h(s)$ for all $(s,h) \in \mathcal{S}^{p,\pi^E}$, the expert's action. Given any reward $r \in \mathfrak{R}$, it is useful to define (recursively) the transition models $\widetilde{p}^M$ and $\widetilde{p}^m$ as:

$$\widetilde{p}^M := \begin{cases} \widetilde{p}^M_h(\cdot|s,a) = \underset{\substack{p': \|p'_h(\cdot|s,a) - \widehat{p}_h(\cdot|s,a)\|_1 \leqslant b_h(s,a) \\ \wedge \forall s' \notin \mathcal{S}^{p,\pi^E}_{h+1} : p'_h(s'|s,a) = 0}}{\arg\max} \underset{s' \sim p'_h(\cdot|s,a)}{\mathbb{E}} V^{\widetilde{\pi}^M}_{h+1}(s'; \widetilde{p}^M, r)\}, \text{ if } (s,a,h) \in \mathcal{Z}^{p,\pi^b} \\ \widetilde{p}^M_h(\cdot|s,a) = \mathbb{1}\{\cdot = \arg\max_{s' \in \mathcal{S}} V^{\widetilde{\pi}^M}_{h+1}(s'; \widetilde{p}^M, r)\}, \text{ if } (s,a,h) \notin \mathcal{Z}^{p,\pi^b} \end{cases}, \\ \widetilde{p}^m := \begin{cases} \widetilde{p}^m_h(\cdot|s,a) = \underset{\substack{p': \|p'_h(\cdot|s,a) - \widehat{p}_h(\cdot|s,a)\|_1 \leqslant b_h(s,a) \\ \wedge \forall s' \notin \mathcal{S}^{p,\pi^E}_{h+1} : p'_h(s'|s,a) = 0}}{\arg\min} \underset{s' \sim p'_h(\cdot|s,a)}{\mathbb{E}} V^{\widetilde{\pi}^m}_{h+1}(s'; \widetilde{p}^m, r)\}, \text{ if } (s,a,h) \in \mathcal{Z}^{p,\pi^b} \\ \widetilde{p}^m_h(\cdot|s,a) = \mathbb{1}\{\cdot = \arg\min_{s' \in \mathcal{S}} V^{\widetilde{\pi}^m}_{h+1}(s'; \widetilde{p}^m, r)\}, \text{ if } (s,a,h) \notin \mathcal{Z}^{p,\pi^b} \end{cases}, \tag{25}$$

where we have used the following (recursive) policy definitions $\widetilde{\pi}^M, \widetilde{\pi}^m \in [\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}$:

$$\widetilde{\pi}^M := \begin{cases} \widetilde{\pi}^M_h(s) = \pi^E_h(s), \text{ if } (s,h) \in \mathcal{S}^{p,\pi^E} \\ \widetilde{\pi}^M_h(\cdot|s) = \mathbb{1}\{\cdot = \arg\max_{a \in \mathcal{A}} Q^{\widetilde{\pi}^M}_h(s,a; \widetilde{p}^M, r)\}, \text{ if } (s,h) \notin \mathcal{S}^{p,\pi^E} \end{cases}, \\ \widetilde{\pi}^m := \begin{cases} \widetilde{\pi}^m_h(s) = \pi^E_h(s), \text{ if } (s,h) \in \mathcal{S}^{p,\pi^E} \\ \widetilde{\pi}^m_h(\cdot|s) = \mathbb{1}\{\cdot = \arg\max_{a \in \mathcal{A}} Q^{\widetilde{\pi}^m}_h(s,a; \widetilde{p}^m, r)\}, \text{ if } (s,h) \notin \mathcal{S}^{p,\pi^E} \end{cases}. \tag{26}$$

Thanks to these definitions, we can rewrite $\widetilde{\mathcal{R}}^\cap$ and $\widetilde{\mathcal{R}}^\cup$ as:

$$\begin{aligned} \widetilde{\mathcal{R}}^\cap &= \{r \in \mathfrak{R} \,|\, \forall (s,h) \in \mathcal{S}^{p,\pi^E}, \forall a \in \mathcal{A} \backslash \{a^E\} : Q^{\pi^E}_h(s,a^E; \widetilde{p}^m, r) \geqslant Q^{\widetilde{\pi}^M}_h(s,a; \widetilde{p}^M, r)\}, \\ \widetilde{\mathcal{R}}^\cup &= \{r \in \mathfrak{R} \,|\, \forall (s,h) \in \mathcal{S}^{p,\pi^E}, \forall a \in \mathcal{A} \backslash \{a^E\} : Q^{\pi^E}_h(s,a^E; \widetilde{p}^M, r) \geqslant Q^{\widetilde{\pi}^m}_h(s,a; \widetilde{p}^m, r)\}. \end{aligned} \tag{27}$$

Both Theorem 6.1 and Theorem 6.2 uses d_∞ instead of d use the same reward choice lemmas, but differ for the performance decomposition lemmas.

Lemma F.6 (Reward Choice Subset)**.** *For any $r \in \mathcal{R}^\cap_{p,\pi^E}$, the reward $\widehat{r}$ constructed as:*

$$\begin{cases} \widehat{r}_h(s,a^E) = r_h(s,a^E) + \sum_{s' \in \mathcal{S}} p_h(s'|s,a^E) V^{\pi^E}_{h+1}(s'; p, r) - \sum_{s' \in \mathcal{S}} \widetilde{p}^m_h(s'|s,a^E) V^{\pi^E}_{h+1}(s'; \widetilde{p}^m, \widehat{r}), & \forall (s,a^E,h) \in \mathcal{Z}^{p,\pi^E} \\ \widehat{r}_h(s,a) = r_h(s,a) + \sum_{s' \in \mathcal{S}} p^M_h(s'|s,a) V^{\pi^M}_{h+1}(s'; p^M, r) - \sum_{s' \in \mathcal{S}} \widetilde{p}^M_h(s'|s,a) V^{\widetilde{\pi}^M}_{h+1}(s'; \widetilde{p}^M, \widehat{r}), & \textit{otherwise}, \end{cases},$$

belongs to $\widetilde{\mathcal{R}}^\cap$.

Proof. Consider any $(s,a^E,h) \in \mathcal{Z}^{p,\pi^E}$. By definition of $\widehat{r}$, by rearranging the terms, we have that:

$$\begin{aligned} &\widehat{r}_h(s,a^E) + \sum_{s' \in \mathcal{S}} \widetilde{p}^m_h(s'|s,a^E) V^{\pi^E}_{h+1}(s'; \widetilde{p}^m, \widehat{r}) = r_h(s,a^E) + \sum_{s' \in \mathcal{S}} p_h(s'|s,a^E) V^{\pi^E}_{h+1}(s'; p, r) \\ &\iff Q^{\pi^E}_h(s,a^E; \widetilde{p}^m, \widehat{r}) = Q^{\pi^E}_h(s,a^E; p, r). \end{aligned} \tag{28}$$

Now, consider any other triple $(s,a,h) \notin \mathcal{Z}^{p,\pi^E}$. Similarly, by rearranging the terms, we obtain:

$$Q^{\widetilde{\pi}^M}_h(s,a; \widetilde{p}^M, \widehat{r}) = Q^{\pi^M}_h(s,a; p^M, r). \tag{29}$$

By hypothesis, $r \in \mathcal{R}^\cap_{p,\pi^E}$, therefore:

$$\forall (s,h) \in \mathcal{S}^{p,\pi^E}, \forall a \in \mathcal{A} \backslash \{a^E\} : Q^{\pi^E}_h(s,a^E; p, r) \geqslant Q^{\pi^M}_h(s,a; p^M, r),$$

from which it follows that:

$$\forall (s,h) \in \mathcal{S}^{p,\pi^E}, \forall a \in \mathcal{A} \backslash \{a^E\} : Q^{\pi^E}_h(s,a^E; \widetilde{p}^m, \widehat{r}) \geqslant Q^{\widetilde{\pi}^M}_h(s,a; \widetilde{p}^M, \widehat{r}).$$

□

For the superset, we have an analogous result.

Lemma F.7 (Reward Choice Superset)**.** *For any $\widehat{r}\in\widetilde{\mathcal{R}}^{\cup}$, the reward r constructed as:*

$$\begin{cases} r_h(s,a^E)=\widehat{r}_h(s,a^E)+\sum_{s'\in\mathcal{S}}\widetilde{p}_h^M(s'|s,a^E)V_{h+1}^{\pi^E}(s';\widetilde{p}^M,\widehat{r})-\sum_{s'\in\mathcal{S}}p_h(s'|s,a^E)V_{h+1}^{\pi^E}(s';p,r), & \forall(s,a^E,h)\in\mathcal{Z}^{p,\pi^E} \\ r_h(s,a)=\widehat{r}_h(s,a)+\sum_{s'\in\mathcal{S}}\widetilde{p}_h^m(s'|s,a)V_{h+1}^{\widetilde{\pi}^m}(s';\widetilde{p}^m,\widehat{r})-\sum_{s'\in\mathcal{S}}p_h^m(s'|s,a)V_{h+1}^{\pi^m}(s';p^m,r), & \textit{otherwise}, \end{cases},$$

belongs to $\mathcal{R}^{\cup}_{p,\pi^E}$.

Proof. Consider any $(s,a^E,h)\in\mathcal{Z}^{p,\pi^E}$. By definition of r, by rearranging the terms, we have that:

$$\begin{aligned} &\widehat{r}_h(s,a^E)+\sum_{s'\in\mathcal{S}}\widetilde{p}_h^M(s'|s,a^E)V_{h+1}^{\pi^E}(s';\widetilde{p}^M,\widehat{r})=r_h(s,a^E)+\sum_{s'\in\mathcal{S}}p_h(s'|s,a^E)V_{h+1}^{\pi^E}(s';p,r) \\ &\iff Q_h^{\pi^E}(s,a^E;\widetilde{p}^M,\widehat{r})=Q_h^{\pi^E}(s,a^E;p,r). \end{aligned} \tag{30}$$

Now, consider any other triple $(s,a,h)\notin\mathcal{Z}^{p,\pi^E}$. Similarly, by rearranging the terms, we obtain:

$$Q_h^{\widetilde{\pi}^m}(s,a;\widetilde{p}^m,\widehat{r})=Q_h^{\pi^m}(s,a;p^m,r). \tag{31}$$

By hypothesis, $\widehat{r}\in\widetilde{\mathcal{R}}^{\cup}$, therefore:

$$\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\backslash\{a^E\}:Q_h^{\pi^E}(s,a^E;\widetilde{p}^M,\widehat{r})\geqslant Q_h^{\widetilde{\pi}^m}(s,a;\widetilde{p}^m,\widehat{r}),$$

from which it follows that:

$$\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\backslash\{a^E\}:Q_h^{\pi^E}(s,a^E;p,r)\geqslant Q_h^{\pi^m}(s,a;p^m,r).$$

Since $p^m\in[p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}$ and π^m is the worst policy in $[\pi^E]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}$ for p^m, then we have $r\in\mathcal{R}^{\cup}_{p,\pi^E}$. □

F.3.1. Lemmas for Theorem 6.1

Lemma F.8 (Performance Decomposition Subset)**.** *Under good event $\mathcal{E}$, it holds that:*

$$\mathcal{H}_d(\mathcal{R}^{\cap}_{p,\pi^E},\widetilde{\mathcal{R}}^{\cap})\leqslant 2H\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}b_h(s,a)+8H^3\max_{(s,a,h)\in\mathcal{Z}^{p,\pi^E}}b_h(s,a).$$

Proof. Observe that:

$$\begin{aligned} \mathcal{H}_d(\mathcal{R}^{\cap}_{p,\pi^E},\widetilde{\mathcal{R}}^{\cap}) &:= \max\{\sup_{r\in\mathcal{R}^{\cap}_{p,\pi^E}}\inf_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cap}}d(r,\widetilde{r}),\sup_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cap}}\inf_{r\in\mathcal{R}^{\cap}_{p,\pi^E}}d(r,\widetilde{r})\} \\ &\overset{(1)}{=}\sup_{r\in\mathcal{R}^{\cap}_{p,\pi^E}}\inf_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cap}}d(r,\widetilde{r}) \\ &=:\sup_{r\in\mathcal{R}^{\cap}_{p,\pi^E}}\inf_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cap}}\frac{1}{M}\sum_{h\in[\![H]\!]}\Big(\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}\big|r_h(s,a)-\widetilde{r}_h(s,a)\big|+\max_{(s,a)\notin\mathcal{Z}_h^{p,\pi^b}}\big|r_h(s,a)-\widetilde{r}_h(s,a)\big|\Big) \\ &\overset{(2)}{\leqslant}\sup_{r\in\mathcal{R}^{\cap}_{p,\pi^E}}\frac{1}{M}\sum_{h\in[\![H]\!]}\Big(\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}\big|r_h(s,a)-\widehat{r}_h(s,a)\big|+\max_{(s,a)\notin\mathcal{Z}_h^{p,\pi^b}}\big|r_h(s,a)-\widehat{r}_h(s,a)\big|\Big), \end{aligned} \tag{32}$$

where at (1) we have used that, under event $\mathcal{E}$, we have $\widetilde{\mathcal{R}}^{\cap}\subseteq\mathcal{R}^{\cap}_{p,\pi^E}$, and at (2) we have applied Lemma F.6, denoting with $\widehat{r}$ the reward chosen from $\widetilde{\mathcal{R}}^{\cap}$.

Now, we consider the various triples $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]$ differently according to the definition of $\widehat{r}$ in Lemma F.6. Let us begin with any $(s,a^E,h)\in\mathcal{Z}^{p,\pi^E}$. Thanks to Eq. 28, we know that, for any $(s,h)\in\mathcal{S}^{p,\pi^E}$, it holds that $Q_h^{\pi^E}(s,a^E;p,r)=$

$Q_h^{\pi^E}(s,a^E;\widetilde{p}^m,\widehat{r})$. Since any expert's action when played from $\mathcal{S}^{p,\pi^E}$ brings to $\mathcal{S}^{p,\pi^E}$ (even under $\widetilde{p}^m$, by definition) , then we can write:

$$\begin{aligned}
\left|r_h(s,a^E)-\widehat{r}_h(s,a^E)\right| &= \Big|\sum_{s'\in\mathcal{S}} p_h(s'|s,a^E)V_{h+1}^{\pi^E}(s';p,r)-\sum_{s'\in\mathcal{S}}\widetilde{p}_h^m(s'|s,a^E)V_{h+1}^{\pi^E}(s';\widetilde{p}^m,\widehat{r})\Big| \\
&= \Big|\sum_{s'\in\mathcal{S}^{p,\pi^E}} p_h(s'|s,a^E)V_{h+1}^{\pi^E}(s';p,r)-\sum_{s'\in\mathcal{S}^{p,\pi^E}}\widetilde{p}_h^m(s'|s,a^E)V_{h+1}^{\pi^E}(s';\widetilde{p}^m,\widehat{r})\Big| \\
&\overset{(1)}{=} \Big|\sum_{s'\in\mathcal{S}^{p,\pi^E}} p_h(s'|s,a^E)V_{h+1}^{\pi^E}(s';p,r)-\sum_{s'\in\mathcal{S}^{p,\pi^E}}\widetilde{p}_h^m(s'|s,a^E)V_{h+1}^{\pi^E}(s';p,r)\Big| \\
&= \Big|\sum_{s'\in\mathcal{S}} (p_h(s'|s,a^E)-\widetilde{p}_h^m(s'|s,a^E))V_{h+1}^{\pi^E}(s';p,r)\Big| \\
&\overset{(2)}{\leqslant} \sum_{s'\in\mathcal{S}}\left|(p_h(s'|s,a^E)-\widetilde{p}_h^m(s'|s,a^E))HM\right| \\
&\leqslant HM\left\|p_h(\cdot|s,a^E)-\hat{p}_h(\cdot|s,a^E)\right\|_1+HM\left\|\widetilde{p}_h^m(\cdot|s,a^E)-\hat{p}_h(\cdot|s,a^E)\right\|_1 \\
&\overset{(3)}{\leqslant} 2MHb_h(s,a^E),
\end{aligned} \tag{33}$$

where at (1) we use Eq. 28, at (2) we use triangle inequality and we upper bound the value function by H times the maximum reward, and at (3) we use the event in Lemma F.2 twice.

Now, let us consider any triple $(s,a,h)\in\mathcal{Z}^{p,\pi^b}\setminus\mathcal{Z}^{p,\pi^E}$. Thanks to Lemma F.6, we can write:

$$\begin{aligned}
\left|r_h(s,a)-\widehat{r}_h(s,a)\right| &= \Big|\sum_{s'\in\mathcal{S}} p_h^M(s'|s,a)V_{h+1}^{\pi^M}(s';p^M,r)-\sum_{s'\in\mathcal{S}}\widetilde{p}_h^M(s'|s,a)V_{h+1}^{\tilde{\pi}^M}(s';\widetilde{p}^M,\widehat{r})\Big| \\
&\overset{(1)}{=} \Big|\sum_{s'\in\mathcal{S}} p_h(s'|s,a)V_{h+1}^{\pi^M}(s';p^M,r)-\sum_{s'\in\mathcal{S}}\widetilde{p}_h^M(s'|s,a)V_{h+1}^{\tilde{\pi}^M}(s';\widetilde{p}^M,\widehat{r}) \\
&\quad \pm\sum_{s'\in\mathcal{S}} p_h(s'|s,a)V_{h+1}^{\tilde{\pi}^M}(s';\widetilde{p}^M,\widehat{r})\Big| \\
&\overset{(2)}{\leqslant} \Big|\sum_{s'\in\mathcal{S}} (p_h(s'|s,a)-\widetilde{p}_h^M(s'|s,a))V_{h+1}^{\tilde{\pi}^M}(s';\widetilde{p}^M,\widehat{r})\Big| \\
&\quad +\Big|\sum_{s'\in\mathcal{S}} p_h(s'|s,a)\big(V_{h+1}^{\pi^M}(s';p^M,r)-V_{h+1}^{\tilde{\pi}^M}(s';\widetilde{p}^M,\widehat{r})\big)\Big| \\
&\overset{(3)}{\leqslant} 2MHb_h(s,a)+\Big|\sum_{s'\in\mathcal{S}} p_h(s'|s,a)\big(V_{h+1}^{\pi^M}(s';p^M,r)-V_{h+1}^{\tilde{\pi}^M}(s';\widetilde{p}^M,\widehat{r})\big)\Big| \\
&\overset{(4)}{\leqslant} 2MHb_h(s,a)+\Big|\sum_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} p_h(s'|s,a)\big(Q_{h+1}^{\pi^E}(s',a^E;p,r)-Q_{h+1}^{\pi^E}(s',a^E;\widetilde{p}^M,\widehat{r})\big)\Big| \\
&\quad +\Big|\sum_{s'\notin\mathcal{S}^{p,\pi^E}_{h+1}} p_h(s'|s,a)\big(\max_{a'\in\mathcal{A}}Q_{h+1}^{\pi^M}(s',a';p^M,r)-\max_{a''\in\mathcal{A}}Q_{h+1}^{\tilde{\pi}^M}(s',a'';\widetilde{p}^M,\widehat{r})\big)\Big| \\
&\overset{(5)}{\leqslant} 2MHb_h(s,a)+\Big|\sum_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} p_h(s'|s,a)\big(Q_{h+1}^{\pi^E}(s',a^E;p,r)-Q_{h+1}^{\pi^E}(s',a^E;\widetilde{p}^M,\widehat{r})\big)\Big| \\
&\quad +\Big|\sum_{s'\notin\mathcal{S}^{p,\pi^E}_{h+1}} p_h(s'|s,a)\underbrace{\big(\max_{a'\in\mathcal{A}}Q_{h+1}^{\pi^M}(s',a';p^M,r)-\max_{a''\in\mathcal{A}}Q_{h+1}^{\pi^M}(s',a'';p^M,r)\big)}_{=0}\Big| \\
&\leqslant 2MHb_h(s,a)+\sum_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} p_h(s'|s,a)\underbrace{\left|Q_{h+1}^{\pi^E}(s',a^E;p,r)-Q_{h+1}^{\pi^E}(s',a^E;\widetilde{p}^M,\widehat{r})\right|}_{=:X_{h+1}(s')},
\end{aligned}$$

where at (1) we have used that, since $(s,a,h)\in\mathcal{Z}^{p,\pi^b}$, then $p_h^M(\cdot|s,a)=p_h(\cdot|s,a)$, at (2) we have applied triangle inequality, at (3) we upper bound the value function by H times the maximum reward and we use the event in Lemma F.2 twice, at (4)

we use triangle inequality and the Bellman's equation and that the expert's action a^E is played by both π^M and $\widetilde{\pi}^M$ in any $(s,h)\in\mathcal{S}^{p,\pi^E}$, at (5) we apply Eq. 29.

Now, having defined terms $X_h(s)$ for all $(s,h)\in\mathcal{S}^{p,\pi^E}$ as above, we recursively bound term $X_{h+1}(s')$.

$$
\begin{aligned}
X_{h+1}(s') &:= \left|Q^{\pi^E}_{h+1}(s',a^E;p,r) - Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^M,\widehat{r})\right| \\
&\overset{(6)}{=} \left|r_{h+1}(s',a^E) - \widehat{r}_{h+1}(s',a^E)\right| \\
&\quad + \Big|\sum_{s''\in\mathcal{S}^{p,\pi^E}_{h+2}} p_{h+1}(s''|s',a^E)V^{\pi^E}_{h+2}(s'';p,r) - \sum_{s''\in\mathcal{S}^{p,\pi^E}_{h+2}} \widetilde{p}^M_{h+1}(s''|s',a^E)V^{\pi^E}_{h+2}(s'';\widetilde{p}^M,\widehat{r})\Big)\Big| \\
&\overset{(7)}{\leqslant} 2MHb_{h+1}(s',a^E) + \Big|\sum_{s''\in\mathcal{S}^{p,\pi^E}_{h+2}} p_{h+1}(s''|s',a^E)V^{\pi^E}_{h+2}(s'';p,r) - \sum_{s''\in\mathcal{S}^{p,\pi^E}_{h+2}} \widetilde{p}^M_{h+1}(s''|s',a^E)V^{\pi^E}_{h+2}(s'';\widetilde{p}^M,\widehat{r})\Big) \\
&\quad \pm \sum_{s''\in\mathcal{S}^{p,\pi^E}_{h+2}} p_{h+1}(s''|s',a^E)V^{\pi^E}_{h+2}(s'';\widetilde{p}^M,\widehat{r})\Big| \\
&\overset{(8)}{\leqslant} 2MHb_{h+1}(s',a^E) + \sum_{s''\in\mathcal{S}^{p,\pi^E}_{h+2}} \Big|\big(p_{h+1}(s''|s',a^E) - \widetilde{p}^M_{h+1}(s''|s',a^E)\big)V^{\pi^E}_{h+2}(s'';\widetilde{p}^M,\widehat{r})\Big| \\
&\quad + \sum_{s''\in\mathcal{S}^{p,\pi^E}_{h+2}} p_{h+1}(s''|s',a^E)\Big|V^{\pi^E}_{h+2}(s'';p,r) - V^{\pi^E}_{h+2}(s'';\widetilde{p}^M,\widehat{r})\Big| \\
&\overset{(9)}{\leqslant} 4MHb_{h+1}(s',a^E) + \sum_{s''\in\mathcal{S}^{p,\pi^E}_{h+2}} p_{h+1}(s''|s',a^E)\underbrace{\Big|Q^{\pi^E}_{h+2}(s'',a^E;p,r) - Q^{\pi^E}_{h+2}(s'',a^E;\widetilde{p}^M,\widehat{r})\Big|}_{=:X_{h+2}(s'')},
\end{aligned}
$$

where at (6) we use the definition of Q function, and we apply triangle inequality; at (7) we use again triangle inequality and Eq. 33, at (8) we apply triangle inequality, and at (9) we use again the event in Lemma F.2 twice. The recursion on the X terms tells us that:

$$
X_h(s) \leqslant 4MHb_h(s,a^E) + \mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)} X_{h+1}(s'). \tag{34}
$$

Therefore, we can upper bound the difference between rewards in $(s,a,h)\in\mathcal{Z}^{p,\pi^b}\setminus\mathcal{Z}^{p,\pi^E}$ as:

$$
\left|r_h(s,a) - \widehat{r}_h(s,a)\right| \leqslant 2MHb_h(s,a) + 4MH\sum_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} p_h(s'|s,a)\sum_{h'\in[\![h+1,H]\!]}\mathop{\mathbb{E}}_{s''\sim\rho^{p,\pi^E}_{h'}(\cdot|s_{h+1}=s')} b_{h'}(s'',a^E). \tag{35}
$$

Now, the only missing triples to consider are those $(s,a,h)\notin\mathcal{Z}^{p,\pi^b}$. Similarly to the triples just considered, we can write:

$$
\begin{aligned}
\left|r_h(s,a) - \widehat{r}_h(s,a)\right| &= \Big|\sum_{s'\in\mathcal{S}} p^M_h(s'|s,a)V^{\pi^M}_{h+1}(s';p^M,r) - \sum_{s'\in\mathcal{S}} \widetilde{p}^M_h(s'|s,a)V^{\widetilde{\pi}^M}_{h+1}(s';\widetilde{p}^M,\widehat{r})\Big| \\
&\overset{(1)}{=} \Big|\max_{s'\in\mathcal{S}} V^{\pi^M}_{h+1}(s';p^M,r) - \max_{s''\in\mathcal{S}} V^{\widetilde{\pi}^M}_{h+1}(s'';\widetilde{p}^M,\widehat{r})\Big| \\
&\overset{(2)}{\leqslant} \max_{s'\in\mathcal{S}}\Big|V^{\pi^M}_{h+1}(s';p^M,r) - V^{\widetilde{\pi}^M}_{h+1}(s';\widetilde{p}^M,\widehat{r})\Big| \\
&\overset{(3)}{=} \max\Big\{\max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\Big|V^{\pi^E}_{h+1}(s';p,r) - V^{\pi^E}_{h+1}(s';\widetilde{p}^M,\widehat{r})\Big|, \\
&\qquad \max_{s'\notin\mathcal{S}^{p,\pi^E}_{h+1}}\underbrace{\Big|V^{\pi^M}_{h+1}(s';p^M,r) - V^{\widetilde{\pi}^M}_{h+1}(s';\widetilde{p}^M,\widehat{r})\Big|}_{=0}\Big\} \\
&= \max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\Big|Q^{\pi^E}_{h+1}(s',a^E;p,r) - Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^M,\widehat{r})\Big|
\end{aligned}
$$

$$
\begin{aligned}
&=: \max_{s' \in \mathcal{S}^{p,\pi^E}_{h+1}} X_{h+1}(s') \\
&\overset{(4)}{\leqslant} 4MH \max_{s' \in \mathcal{S}^{p,\pi^E}_{h+1}} \Big(b_{h+1}(s', a^E) + \sum_{h' \in [\![h+2, H]\!]} \mathop{\mathbb{E}}_{s'' \sim \rho^{p,\pi^E}_{h'}(\cdot | s_{h+1} = s')} b_{h'}(s'', a^E) \Big),
\end{aligned} \tag{36}
$$

where at (1) we have used that $(s,a,h) \notin \mathcal{Z}^{p,\pi^b}$ and the definitions of p^M and $\widetilde{p}^M$, at (2) we have used that for any pair of real-valued functions f, g it holds that $|\max_x f(x) - \max_x g(x)| \leqslant \max_x |f(x) - g(x)|$, at (3) we use Eq. 29 to realize that in (s,h) outside $\mathcal{S}^{p,\pi^E}_{h+1}$ we have an equality of Q-functions, and thus the difference is 0, at (4) we have unfolded the recursion on the X terms by using Eq. 34.

By combining Eq. 32 with Eq. 33, Eq. 35, and Eq. 36, we get:

$$
\begin{aligned}
\mathcal{H}_d(\mathcal{R}^{\cap}_{p,\pi^E}, \widetilde{\mathcal{R}}^{\cap}) \leqslant & \sum_{h \in [\![H]\!]} \Bigg(\sum_{(s,a) \in \mathcal{Z}^{p,\pi^E}_h} \rho^{p,\pi^b}_h(s,a) 2H b_h(s,a) \\
& + \sum_{(s,a) \in \mathcal{Z}^{p,\pi^b}_h \setminus \mathcal{Z}^{p,\pi^E}_h} \rho^{p,\pi^b}_h(s,a) \Big(2H b_h(s,a) \\
& + 4H \sum_{s' \in \mathcal{S}^{p,\pi^E}_{h+1}} p_h(s'|s,a) \sum_{h' \in [\![h+1, H]\!]} \mathop{\mathbb{E}}_{s'' \sim \rho^{p,\pi^E}_{h'}(\cdot | s_{h+1} = s')} b_{h'}(s'', a^E) \Big) \\
& + \max_{(s,a) \notin \mathcal{Z}^{p,\pi^b}_h} 4H \max_{s' \in \mathcal{S}^{p,\pi^E}_{h+1}} \sum_{h' \in [\![h+1, H]\!]} \mathop{\mathbb{E}}_{s'' \sim \rho^{p,\pi^E}_{h'}(\cdot | s_{h+1} = s')} b_{h'}(s'', a^E) \Bigg) \\
\leqslant & \sum_{h \in [\![H]\!]} \Bigg(\sum_{(s,a) \in \mathcal{Z}^{p,\pi^E}_h} \rho^{p,\pi^b}_h(s,a) 2H b_h(s,a) \\
& + \sum_{(s,a) \in \mathcal{Z}^{p,\pi^b}_h \setminus \mathcal{Z}^{p,\pi^E}_h} \rho^{p,\pi^b}_h(s,a) \Big(2H b_h(s,a) + 4H^2 \max_{(s', a^E, h') \in \mathcal{Z}^{p,\pi^E}} b_{h'}(s', a^E) \Big) \\
& + \max_{(s,a) \notin \mathcal{Z}^{p,\pi^b}_h} 4H^2 \max_{(s', a^E, h') \in \mathcal{Z}^{p,\pi^E}} b_{h'}(s', a^E) \Bigg) \\
\leqslant & \sum_{h \in [\![H]\!]} \Bigg(\sum_{(s,a) \in \mathcal{Z}^{p,\pi^E}_h} \rho^{p,\pi^b}_h(s,a) 2H b_h(s,a) \\
& + \sum_{(s,a) \in \mathcal{Z}^{p,\pi^b}_h \setminus \mathcal{Z}^{p,\pi^E}_h} \rho^{p,\pi^b}_h(s,a) 2H b_h(s,a) + 8H^2 \max_{(s', a^E, h') \in \mathcal{Z}^{p,\pi^E}} b_{h'}(s', a^E) \Bigg) \\
\leqslant & 2H \sum_{h \in [\![H]\!]} \mathop{\mathbb{E}}_{(s,a) \sim \rho^{p,\pi^b}_h(\cdot,\cdot)} b_h(s,a) + 8H^3 \max_{(s,a,h) \in \mathcal{Z}^{p,\pi^E}} b_h(s,a).
\end{aligned}
$$

□

Lemma F.9 (Performance Decomposition Superset)**.** *Under good event $\mathcal{E}$, it holds that:*

$$
\mathcal{H}_d(\mathcal{R}^{\cup}_{p,\pi^E} \widetilde{\mathcal{R}}^{\cup}) \leqslant 2H \sum_{h \in [\![H]\!]} \mathop{\mathbb{E}}_{(s,a) \sim \rho^{p,\pi^b}_h(\cdot,\cdot)} b_h(s,a) + 8H^3 \max_{(s,a,h) \in \mathcal{Z}^{p,\pi^E}} b_h(s,a).
$$

Proof. Observe that:

$$
\begin{aligned}
\mathcal{H}_d(\mathcal{R}^{\cup}_{p,\pi^E}, \widetilde{\mathcal{R}}^{\cup}) &:= \max\{ \sup_{r \in \mathcal{R}^{\cup}_{p,\pi^E}} \inf_{\widetilde{r} \in \widetilde{\mathcal{R}}^{\cup}} d(r, \widetilde{r}), \sup_{\widetilde{r} \in \widetilde{\mathcal{R}}^{\cup}} \inf_{r \in \mathcal{R}^{\cup}_{p,\pi^E}} d(r, \widetilde{r}) \} \\
&\overset{(1)}{=} \sup_{\widetilde{r} \in \widetilde{\mathcal{R}}^{\cup}} \inf_{r \in \mathcal{R}^{\cup}_{p,\pi^E}} d(r, \widetilde{r})
\end{aligned}
$$

$$=: \sup_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cup}} \inf_{r\in\mathcal{R}^{\cup}_{p,\pi^E}} \frac{1}{M}\sum_{h\in[\![H]\!]}\Big(\mathbb{E}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}\big|r_h(s,a)-\widetilde{r}_h(s,a)\big| + \max_{(s,a)\notin\mathcal{Z}_h^{p,\pi^b}}\big|r_h(s,a)-\widetilde{r}_h(s,a)\big|\Big)$$
$$\overset{(2)}{\leqslant} \sup_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cup}} \frac{1}{M}\sum_{h\in[\![H]\!]}\Big(\mathbb{E}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}\big|r_h(s,a)-\widetilde{r}_h(s,a)\big| + \max_{(s,a)\notin\mathcal{Z}_h^{p,\pi^b}}\big|r_h(s,a)-\widetilde{r}_h(s,a)\big|\Big), \tag{37}$$

where at (1) we have used that, under event $\mathcal{E}$, we have $\mathcal{R}^{\cup}_{p,\pi^E}\subseteq\widetilde{\mathcal{R}}^{\cup}$, and at (2) we have applied Lemma F.7, denoting with r the reward chosen from $\mathcal{R}^{\cup}_{p,\pi^E}$.

Now, we consider the various triples $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]$ differently according to the definition of r in Lemma F.7. Let us begin with any $(s,a^E,h)\in\mathcal{Z}^{p,\pi^E}$. We can write:

$$\begin{aligned}
\big|r_h(s,a^E)-\widehat{r}_h(s,a^E)\big| &= \Big|\sum_{s'\in\mathcal{S}} p_h(s'|s,a^E)V^{\pi^E}_{h+1}(s';p,r) - \sum_{s'\in\mathcal{S}}\widetilde{p}^M_h(s'|s,a^E)V^{\pi^E}_{h+1}(s';\widetilde{p}^M,\widehat{r})\\
&= \Big|\sum_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} p_h(s'|s,a^E)V^{\pi^E}_{h+1}(s';p,r) - \sum_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\widetilde{p}^M_h(s'|s,a^E)V^{\pi^E}_{h+1}(s';\widetilde{p}^M,\widehat{r})\Big|\\
&= \Big|\sum_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} p_h(s'|s,a^E)V^{\pi^E}_{h+1}(s';p,r) - \sum_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\widetilde{p}^M_h(s'|s,a^E)V^{\pi^E}_{h+1}(s';p,r)\Big|\\
&\leqslant \big\|p_h(\cdot|s,a^E)-\widetilde{p}^M_h(\cdot|s,a^E)\big\|_1 MH\\
&\leqslant \big\|p_h(\cdot|s,a^E)-\hat{p}_h(\cdot|s,a^E)\big\|_1 MH + \big\|\hat{p}_h(\cdot|s,a^E)-\widetilde{p}^M_h(\cdot|s,a^E)\big\|_1 MH\\
&\leqslant 2MHb_h(s,a^E).
\end{aligned} \tag{38}$$

Now, let us consider any triple $(s,a,h)\in\mathcal{Z}^{p,\pi^b}\setminus\mathcal{Z}^{p,\pi^E}$. Thanks to Lemma F.7, we can write:

$$\begin{aligned}
\big|r_h(s,a)-\widehat{r}_h(s,a)\big| &= \Big|\sum_{s'\in\mathcal{S}} p^m_h(s'|s,a)V^{\pi^m}_{h+1}(s';p^m,r) - \sum_{s'\in\mathcal{S}}\widetilde{p}^m_h(s'|s,a)V^{\tilde{\pi}^m}_{h+1}(s';\widetilde{p}^m,\widehat{r})\Big|\\
&\overset{(1)}{=} \Big|\sum_{s'\in\mathcal{S}} p_h(s'|s,a)V^{\pi^m}_{h+1}(s';p^m,r) - \sum_{s'\in\mathcal{S}}\widetilde{p}^m_h(s'|s,a)V^{\tilde{\pi}^m}_{h+1}(s';\widetilde{p}^m,\widehat{r})\\
&\qquad \pm \sum_{s'\in\mathcal{S}} p_h(s'|s,a)V^{\tilde{\pi}^m}_{h+1}(s';\widetilde{p}^m,\widehat{r})\Big|\\
&\overset{(2)}{\leqslant} \Big|\sum_{s'\in\mathcal{S}}\big(p_h(s'|s,a)-\widetilde{p}^m_h(s'|s,a)\big)V^{\tilde{\pi}^m}_{h+1}(s';\widetilde{p}^m,\widehat{r})\Big|\\
&\qquad + \Big|\sum_{s'\in\mathcal{S}} p_h(s'|s,a)\big(V^{\pi^m}_{h+1}(s';p^m,r)-V^{\tilde{\pi}^m}_{h+1}(s';\widetilde{p}^m,\widehat{r})\big)\Big|\\
&\overset{(3)}{\leqslant} 2MHb_h(s,a) + \Big|\sum_{s'\in\mathcal{S}} p_h(s'|s,a)\big(V^{\pi^m}_{h+1}(s';p^m,r)-V^{\tilde{\pi}^m}_{h+1}(s';\widetilde{p}^m,\widehat{r})\big)\Big|\\
&\overset{(4)}{\leqslant} 2MHb_h(s,a) + \Big|\sum_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} p_h(s'|s,a)\big(Q^{\pi^E}_{h+1}(s',a^E;p,r)-Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^m,\widehat{r})\big)\Big|\\
&\qquad + \Big|\sum_{s'\notin\mathcal{S}^{p,\pi^E}_{h+1}} p_h(s'|s,a)\big(\max_{a'\in\mathcal{A}}Q^{\pi^m}_{h+1}(s',a';p^m,r)-\max_{a''\in\mathcal{A}}Q^{\tilde{\pi}^m}_{h+1}(s',a'';\widetilde{p}^m,\widehat{r})\big)\Big|\\
&\overset{(5)}{\leqslant} 2MHb_h(s,a) + \Big|\sum_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} p_h(s'|s,a)\big(Q^{\pi^E}_{h+1}(s',a^E;p,r)-Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^m,\widehat{r})\big)\Big|\\
&\qquad + \Big|\sum_{s'\notin\mathcal{S}^{p,\pi^E}_{h+1}} p_h(s'|s,a)\underbrace{\big(\max_{a'\in\mathcal{A}}Q^{\pi^m}_{h+1}(s',a';p^m,r)-\max_{a''\in\mathcal{A}}Q^{\tilde{\pi}^m}_{h+1}(s',a'';p^m,r)\big)}_{=0}\Big|\\
&\leqslant 2MHb_h(s,a) + \sum_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} p_h(s'|s,a)\underbrace{\big|Q^{\pi^E}_{h+1}(s',a^E;p,r)-Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^m,\widehat{r})\big|}_{=:Y_{h+1}(s')},
\end{aligned}$$

where at (1) we have used that, since $(s,a,h)\in\mathcal{Z}^{p,\pi^b}$, then $p_h^m(\cdot|s,a)=p_h(\cdot|s,a)$, at (2) we have applied triangle inequality, at (3) we upper bound the value function by H times the maximum reward and we use the event in Lemma F.2 twice, at (4) we use triangle inequality and the Bellman optimality equation and that the expert's action a^E is optimal in any $(s,h)\in\mathcal{S}^{p,\pi^E}$, at (5) we apply Eq. 31.

Now, having defined terms $\{Y_h(s)\}_h$ for all $s\in\mathcal{S}^{p,\pi^E}$ as above, we recursively bound term $Y_{h+1}(s')$. It should be remarked that $(s',h+1)\in\mathcal{S}^{p,\pi^E}$.

$$
\begin{aligned}
Y_{h+1}(s') &:= \left|Q_{h+1}^{\pi^E}(s',a^E;p,r)-Q_{h+1}^{\pi^E}(s',a^E;\widetilde{p}^m,\widehat{r})\right| \\
&\overset{(6)}{=} \left|r_{h+1}(s',a^E)-\widehat{r}_{h+1}(s',a^E)\right| \\
&\quad + \Big|\sum_{s''\in\mathcal{S}_{h+2}^{p,\pi^E}} p_{h+1}(s''|s',a^E)V_{h+2}^{\pi^E}(s'';p,r)-\sum_{s''\in\mathcal{S}_{h+2}^{p,\pi^E}}\widetilde{p}_{h+1}^m(s''|s',a^E)V_{h+2}^{\pi^E}(s'';\widetilde{p}^m,\widehat{r})\Big)\Big| \\
&\overset{(7)}{\leqslant} 2MHb_{h+1}(s',a^E)+\Big|\sum_{s''\in\mathcal{S}_{h+2}^{p,\pi^E}} p_{h+1}(s''|s',a^E)V_{h+2}^{\pi^E}(s'';p,r)-\sum_{s''\in\mathcal{S}_{h+2}^{p,\pi^E}}\widetilde{p}_{h+1}^m(s''|s',a^E)V_{h+2}^{\pi^E}(s'';\widetilde{p}^m,\widehat{r})\Big) \\
&\quad \pm \sum_{s''\in\mathcal{S}_{h+2}^{p,\pi^E}} p_{h+1}(s''|s',a^E)V_{h+2}^{\pi^E}(s'';\widetilde{p}^m,\widehat{r})\Big| \\
&\overset{(8)}{\leqslant} 2MHb_{h+1}(s',a^E)+\sum_{s''\in\mathcal{S}_{h+2}^{p,\pi^E}}\Big|\big(p_{h+1}(s''|s',a^E)-\widetilde{p}_{h+1}^m(s''|s',a^E)\big)V_{h+2}^{\pi^E}(s'';\widetilde{p}^m,\widehat{r})\Big| \\
&\quad + \sum_{s''\in\mathcal{S}_{h+2}^{p,\pi^E}} p_{h+1}(s''|s',a^E)\Big|V_{h+2}^{\pi^E}(s'';p,r)-V_{h+2}^{\pi^E}(s'';\widetilde{p}^m,\widehat{r})\Big| \\
&\overset{(9)}{\leqslant} 4MHb_{h+1}(s',a^E)+\sum_{s''\in\mathcal{S}_{h+2}^{p,\pi^E}} p_{h+1}(s''|s',a^E)\underbrace{\Big|Q_{h+2}^{\pi^E}(s'',a^E;p,r)-Q_{h+2}^{\pi^E}(s'',a^E;\widetilde{p}^m,\widehat{r})\Big|}_{=:Y_{h+2}(s'')},
\end{aligned}
$$

where at (6) we use the definition of Q function, and we apply triangle inequality; at (7) we use again triangle inequality and Eq. 38, at (8) we apply triangle inequality, and at (9) we use again the event in Lemma F.2 twice. The recursion on the Y terms tells us that:

$$
Y_h(s)\leqslant 4MHb_h(s,a^E)+\mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)}Y_{h+1}(s'). \tag{39}
$$

Therefore, we can upper bound the difference between rewards in $(s,a,h)\in\mathcal{Z}^{p,\pi^b}\setminus\mathcal{Z}^{p,\pi^E}$ as:

$$
\left|r_h(s,a)-\widehat{r}_h(s,a)\right|\leqslant 2MHb_h(s,a)+4MH\sum_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}}p_h(s'|s,a)\sum_{h'\in[\![h+1,H]\!]}\mathop{\mathbb{E}}_{s''\sim\rho_{h'}^{p,\pi^E}(\cdot|s_{h+1}=s')}b_{h'}(s'',a^E). \tag{40}
$$

Now, the only missing triples to consider are those $(s,a,h)\notin\mathcal{Z}^{p,\pi^b}$. Similarly to the triples just considered, we can write:

$$
\begin{aligned}
\left|r_h(s,a)-\widehat{r}_h(s,a)\right| &= \Big|\sum_{s'\in\mathcal{S}}p_h^m(s'|s,a)V_{h+1}^{\pi^m}(s';p^m,r)-\sum_{s'\in\mathcal{S}}\widetilde{p}_h^m(s'|s,a)V_{h+1}^{\widetilde{\pi}^m}(s';\widetilde{p}^m,\widehat{r})\Big| \\
&\overset{(1)}{=} \left|\min_{s'\in\mathcal{S}}V_{h+1}^{\pi^m}(s';p^m,r)-\min_{s''\in\mathcal{S}}V_{h+1}^{\widetilde{\pi}^m}(s'';\widetilde{p}^m,\widehat{r})\right| \\
&\overset{(2)}{\leqslant} \max_{s'\in\mathcal{S}}\left|V_{h+1}^{\pi^m}(s';p^m,r)-V_{h+1}^{\widetilde{\pi}^m}(s';\widetilde{p}^m,\widehat{r})\right| \\
&\overset{(3)}{=} \max\Big\{\max_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}}\left|V_{h+1}^{\pi^E}(s';p,r)-V_{h+1}^{\pi^E}(s';\widetilde{p}^m,\widehat{r})\right|,
\end{aligned}
$$

$$\max_{s'\notin\mathcal{S}^{p,\pi^E}_{h+1}}\Big\{\underbrace{\big|V^{\pi^m}_{h+1}(s';p^m,r)-V^{\widetilde{\pi}^m}_{h+1}(s';\widetilde{p}^m,\widehat{r})\big|}_{=0}\Big\}$$
$$=\max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\big|Q^{\pi^E}_{h+1}(s',a^E;p,r)-Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^m,\widehat{r})\big|$$
$$=:\max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}Y_{h+1}(s')$$
$$\overset{(4)}{\leqslant}4MH\max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\Big(b_{h+1}(s',a^E)+\sum_{h'\in[\![h+2,H]\!]}\mathbb{E}_{s''\sim\rho^{p,\pi^E}_{h'}(\cdot|s_{h+1}=s')}b_{h'}(s'',a^E)\Big),\tag{41}$$

where at (1) we have used that $(s,a,h)\notin\mathcal{Z}^{p,\pi^b}$ and the definitions of p^m and $\widetilde{p}^m$, at (2) we have used that for any pair of real-valued functions f,g it holds that $|\min_x f(x)-\min_x g(x)|\leqslant\max_x|f(x)-g(x)|$, at (3) we use Eq. 31 to realize that in (s,h) outside $\mathcal{S}^{p,\pi^E}_{h+1}$ we have an equality of Q-functions, and thus the difference is 0, and that in $\mathcal{S}^{p,\pi^E}$ the optimal action is always the expert's action, at (4) we have unfolded the recursion on the Y terms by using Eq. 39.

By combining Eq. 37 with Eq. 38, Eq. 40, and Eq. 41, we get:

$$\begin{aligned}
\mathcal{H}_d(\mathcal{R}^{\cup}_{p,\pi^E},\widetilde{\mathcal{R}}^{\cup})\leqslant&\sum_{h\in[\![H]\!]}\Bigg(\sum_{(s,a)\in\mathcal{Z}^{p,\pi^E}_h}\rho^{p,\pi^b}_h(s,a)2Hb_h(s,a)\\
&+\sum_{(s,a)\in\mathcal{Z}^{p,\pi^b}_h\setminus\mathcal{Z}^{p,\pi^E}_h}\rho^{p,\pi^b}_h(s,a)\Big(2Hb_h(s,a)\\
&+4H\sum_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}p_h(s'|s,a)\sum_{h'\in[\![h+1,H]\!]}\mathbb{E}_{s''\sim\rho^{p,\pi^E}_{h'}(\cdot|s_{h+1}=s')}b_{h'}(s'',a^E)\Big)\\
&+\max_{(s,a)\notin\mathcal{Z}^{p,\pi^b}_h}4H\max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\sum_{h'\in[\![h+1,H]\!]}\mathbb{E}_{s''\sim\rho^{p,\pi^E}_{h'}(\cdot|s_{h+1}=s')}b_{h'}(s'',a^E)\Bigg)\\
\leqslant&\sum_{h\in[\![H]\!]}\Bigg(\sum_{(s,a)\in\mathcal{Z}^{p,\pi^E}_h}\rho^{p,\pi^b}_h(s,a)2Hb_h(s,a)\\
&+\sum_{(s,a)\in\mathcal{Z}^{p,\pi^b}_h\setminus\mathcal{Z}^{p,\pi^E}_h}\rho^{p,\pi^b}_h(s,a)\Big(2Hb_h(s,a)+4H^2\max_{(s',a^E,h')\in\mathcal{Z}^{p,\pi^E}}b_{h'}(s',a^E)\Big)\\
&+\max_{(s,a)\notin\mathcal{Z}^{p,\pi^b}_h}4H^2\max_{(s',a^E,h')\in\mathcal{Z}^{p,\pi^E}}b_{h'}(s',a^E)\Bigg)\\
\leqslant&\sum_{h\in[\![H]\!]}\Bigg(\sum_{(s,a)\in\mathcal{Z}^{p,\pi^E}_h}\rho^{p,\pi^b}_h(s,a)2Hb_h(s,a)\\
&+\sum_{(s,a)\in\mathcal{Z}^{p,\pi^b}_h\setminus\mathcal{Z}^{p,\pi^E}_h}\rho^{p,\pi^b}_h(s,a)2Hb_h(s,a)+8H^2\max_{(s',a^E,h')\in\mathcal{Z}^{p,\pi^E}}b_{h'}(s',a^E)\Bigg)\\
\leqslant&2H\sum_{h\in[\![H]\!]}\mathbb{E}_{(s,a)\sim\rho^{p,\pi^b}_h(\cdot,\cdot)}b_h(s,a)+8H^3\max_{(s,a,h)\in\mathcal{Z}^{p,\pi^E}}b_h(s,a).
\end{aligned}$$

□

F.3.2. LEMMAS FOR THEOREM 6.2

Lemma F.10 (Performance Decomposition Subset)**.** *Under good event $\mathcal{E}$, it holds that:*

$$\mathcal{H}_\infty(\mathcal{R}^{\cap}_{p,\pi^E},\widetilde{\mathcal{R}}^{\cap})\leqslant2H^2\max_{(s,a,h)\in\mathcal{Z}^{p,\pi^b}}b_h(s,a)+4H^3\max_{(s,a,h)\in\mathcal{Z}^{p,\pi^E}}b_h(s,a).$$

Proof. Observe that:

$$\begin{aligned}\mathcal{H}_\infty(\mathcal{R}^{\cap}_{p,\pi^E},\widetilde{\mathcal{R}}^{\cap}) &:= \max\{\sup_{r\in\mathcal{R}^{\cap}_{p,\pi^E}}\inf_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cap}} d_\infty(r,\widetilde{r}),\ \sup_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cap}}\inf_{r\in\mathcal{R}^{\cap}_{p,\pi^E}} d_\infty(r,\widetilde{r})\}\\ &\overset{(1)}{=} \sup_{r\in\mathcal{R}^{\cap}_{p,\pi^E}}\inf_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cap}} d_\infty(r,\widetilde{r})\\ &=: \sup_{r\in\mathcal{R}^{\cap}_{p,\pi^E}}\inf_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cap}} \frac{1}{M}\sum_{h\in[\![H]\!]}\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|r_h(s,a)-\widetilde{r}_h(s,a)\right|\\ &\overset{(2)}{\leqslant} \sup_{r\in\mathcal{R}^{\cap}_{p,\pi^E}}\frac{1}{M}\sum_{h\in[\![H]\!]}\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|r_h(s,a)-\widehat{r}_h(s,a)\right|,\end{aligned} \tag{42}$$

where at (1) we have used that, under event $\mathcal{E}$, we have $\widetilde{\mathcal{R}}^{\cap}\subseteq\mathcal{R}^{\cap}_{p,\pi^E}$, and at (2) we have applied Lemma F.6, denoting with $\widehat{r}$ the reward chosen from $\widetilde{\mathcal{R}}^{\cap}$.

By combining Eq. 42 with Eq. 33, Eq. 35, and Eq. 36, we get:

$$\begin{aligned}\mathcal{H}_\infty(\mathcal{R}^{\cap}_{p,\pi^E},\widetilde{\mathcal{R}}^{\cap}) \leqslant \sum_{h\in[\![H]\!]}\max\Bigg\{&\max_{(s,a)\in\mathcal{Z}^{p,\pi^E}_h} 2Hb_h(s,a),\\ &\max_{(s,a)\in\mathcal{Z}^{p,\pi^b}_h\setminus\mathcal{Z}^{p,\pi^E}_h} 2Hb_h(s,a)+4H\sum_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}p_h(s'|s,a)\sum_{h'\in[\![h+1,H]\!]}\mathbb{E}_{s''\sim\rho^{p,\pi^E}_{h'}(\cdot|s_{h+1}=s')} b_{h'}(s'',a^E),\\ &\max_{(s,a)\notin\mathcal{Z}^{p,\pi^b}_h} 4H\max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\sum_{h'\in[\![h+1,H]\!]}\mathbb{E}_{s''\sim\rho^{p,\pi^E}_{h'}(\cdot|s_{h+1}=s')} b_{h'}(s'',a^E)\Bigg\}\\ \leqslant \sum_{h\in[\![H]\!]}\max\Bigg\{&\max_{(s,a)\in\mathcal{Z}^{p,\pi^E}_h} 2Hb_h(s,a),\\ &\max_{(s,a)\in\mathcal{Z}^{p,\pi^b}_h\setminus\mathcal{Z}^{p,\pi^E}_h} 2Hb_h(s,a)+4H^2\max_{(s',a^E,h')\in\mathcal{Z}^{p,\pi^E}} b_{h'}(s',a^E),\\ &\max_{(s,a)\notin\mathcal{Z}^{p,\pi^b}_h} 4H^2\max_{(s',a^E,h')\in\mathcal{Z}^{p,\pi^E}} b_{h'}(s',a^E)\Bigg\}\\ \leqslant \sum_{h\in[\![H]\!]}\max\Bigg\{&\max_{(s,a)\in\mathcal{Z}^{p,\pi^E}_h} 2Hb_h(s,a),\ \max_{(s,a)\in\mathcal{Z}^{p,\pi^b}_h\setminus\mathcal{Z}^{p,\pi^E}_h} 2Hb_h(s,a)+4H^2\max_{(s',a^E,h')\in\mathcal{Z}^{p,\pi^E}} b_{h'}(s',a^E)\Bigg\}\\ \leqslant \sum_{h\in[\![H]\!]}\Bigg\{&\max_{(s,a)\in\mathcal{Z}^{p,\pi^b}_h} 2Hb_h(s,a)+\sum_{h\in[\![H]\!]}4H^2\max_{(s',a^E,h')\in\mathcal{Z}^{p,\pi^E}} b_{h'}(s',a^E)\Bigg\}\\ \leqslant &2H^2\max_{(s,a,h)\in\mathcal{Z}^{p,\pi^b}} b_h(s,a)+4H^3\max_{(s,a,h)\in\mathcal{Z}^{p,\pi^E}} b_h(s,a).\end{aligned}$$

□

Lemma F.11 (Performance Decomposition Superset)**.** *Under good event $\mathcal{E}$, it holds that:*

$$\mathcal{H}_\infty(\mathcal{R}^{\cup}_{p,\pi^E},\widetilde{\mathcal{R}}^{\cup})\leqslant 2H^2\max_{(s,a,h)\in\mathcal{Z}^{p,\pi^b}} b_h(s,a)+4H^3\max_{(s,a,h)\in\mathcal{Z}^{p,\pi^E}} b_h(s,a).$$

Proof. Observe that:

$$\begin{aligned}\mathcal{H}_\infty(\mathcal{R}^{\cup}_{p,\pi^E},\widetilde{\mathcal{R}}^{\cup}) &:= \max\{\sup_{r\in\mathcal{R}^{\cup}_{p,\pi^E}}\inf_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cup}} d_\infty(r,\widetilde{r}),\ \sup_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cup}}\inf_{r\in\mathcal{R}^{\cup}_{p,\pi^E}} d_\infty(r,\widetilde{r})\}\\ &\overset{(1)}{=} \sup_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cup}}\inf_{r\in\mathcal{R}^{\cup}_{p,\pi^E}} d_\infty(r,\widetilde{r})\\ &=: \sup_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cup}}\inf_{r\in\mathcal{R}^{\cup}_{p,\pi^E}} \frac{1}{M}\sum_{h\in[\![H]\!]}\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|r_h(s,a)-\widetilde{r}_h(s,a)\right|\end{aligned}$$

$$\overset{(2)}{\leqslant} \sup_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cup}} \frac{1}{M}\sum_{h\in[\![H]\!]}\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|r_h(s,a)-\widetilde{r}_h(s,a)\right|, \tag{43}$$

where at (1) we have used that, under event $\mathcal{E}$, we have $\mathcal{R}^{\cup}_{p,\pi^E}\subseteq\widetilde{\mathcal{R}}^{\cup}$, and at (2) we have applied Lemma F.7, denoting with r the reward chosen from $\mathcal{R}^{\cup}_{p,\pi^E}$.

By combining Eq. 43 with Eq. 38, Eq. 40, and Eq. 41, we get:

$$\begin{aligned}
\mathcal{H}_\infty(\mathcal{R}^{\cup}_{p,\pi^E},\widetilde{\mathcal{R}}^{\cup}) \leqslant \sum_{h\in[\![H]\!]}\max\Bigg\{ & \max_{(s,a)\in\mathcal{Z}^{p,\pi^E}_h} 2Hb_h(s,a), \\
& \max_{(s,a)\in\mathcal{Z}^{p,\pi^b}_h\setminus\mathcal{Z}^{p,\pi^E}_h} 2Hb_h(s,a)+4H\sum_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}p_h(s'|s,a)\sum_{h'\in[\![h+1,H]\!]}\mathbb{E}_{s''\sim\rho^{p,\pi^E}_{h'}(\cdot|s_{h+1}=s')} b_{h'}(s'',a^E), \\
& \max_{(s,a)\notin\mathcal{Z}^{p,\pi^b}_h} 4H\max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\sum_{h'\in[\![h+1,H]\!]}\mathbb{E}_{s''\sim\rho^{p,\pi^E}_{h'}(\cdot|s_{h+1}=s')} b_{h'}(s'',a^E)\Bigg\} \\
\leqslant \sum_{h\in[\![H]\!]}\max\Bigg\{ & \max_{(s,a)\in\mathcal{Z}^{p,\pi^E}_h} 2Hb_h(s,a), \\
& \max_{(s,a)\in\mathcal{Z}^{p,\pi^b}_h\setminus\mathcal{Z}^{p,\pi^E}_h} 2Hb_h(s,a)+4H^2\max_{(s',a^E,h')\in\mathcal{Z}^{p,\pi^E}} b_{h'}(s',a^E), \\
& \max_{(s,a)\notin\mathcal{Z}^{p,\pi^b}_h} 4H^2\max_{(s',a^E,h')\in\mathcal{Z}^{p,\pi^E}} b_{h'}(s',a^E)\Bigg\} \\
\leqslant \sum_{h\in[\![H]\!]}\max\Bigg\{ & \max_{(s,a)\in\mathcal{Z}^{p,\pi^E}_h} 2Hb_h(s,a), \max_{(s,a)\in\mathcal{Z}^{p,\pi^b}_h\setminus\mathcal{Z}^{p,\pi^E}_h} 2Hb_h(s,a)+4H^2\max_{(s',a^E,h')\in\mathcal{Z}^{p,\pi^E}} b_{h'}(s',a^E)\Bigg\} \\
\leqslant \sum_{h\in[\![H]\!]}\max\Bigg\{ & \max_{(s,a)\in\mathcal{Z}^{p,\pi^b}_h} 2Hb_h(s,a)+\sum_{h\in[\![H]\!]}4H^2\max_{(s',a^E,h')\in\mathcal{Z}^{p,\pi^E}} b_{h'}(s',a^E)\Bigg\} \\
\leqslant & 2H^2\max_{(s,a,h)\in\mathcal{Z}^{p,\pi^b}} b_h(s,a)+4H^3\max_{(s,a,h)\in\mathcal{Z}^{p,\pi^E}} b_h(s,a).
\end{aligned}$$

□

F.3.3. Proofs of the main theorems

Thanks to Lemma F.8 and Lemma F.9, we can conclude the proof of Theorem 6.1.

Theorem 6.1. *Let $\mathcal{M}$ be an MDP without reward and let π^E be the expert's policy. Let $\mathcal{D}^E$ and $\mathcal{D}^b$ be two datasets of τ^E and τ^b trajectories collected by executing policies π^E and π^b in $\mathcal{M}$. Under Assumption 2.1, **PIRLO** is (ϵ,δ)-PAC for d-IRL with a sample complexity at most:*

$$\begin{aligned}
\tau^b \leqslant \widetilde{\mathcal{O}}\Bigg(& \frac{H^3 Z^{p,\pi^b}\ln\frac{1}{\delta}}{\epsilon^2}\left(\ln\frac{1}{\delta}+S^{p,\pi^b}_{\max}\right) \\
& +\frac{H^6\ln\frac{1}{\delta}}{\rho^{\pi^b,\mathcal{Z}^{p,\pi^E}}_{\min}\epsilon^2}\left(\ln\frac{1}{\delta}+S^{p,\pi^b}_{\max}\right)+\frac{\ln\frac{1}{\delta}}{\ln\frac{1}{1-\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min}}}\Bigg),
\end{aligned}$$

*and τ^E is bounded as in Theorem 5.1. Furthermore, **PIRLO** is inclusion monotonic.*

Proof. The proof for the subset and superset is completely analogous. Under good event $\mathcal{E}$, the performance decomposition lemma for the subset (Lemma F.8) tells us that:

$$\mathcal{H}_d(\mathcal{R}^{\cap}_{p,\pi^E},\widetilde{\mathcal{R}}^{\cap}) \leqslant 2H\sum_{h\in[\![H]\!]}\mathbb{E}_{(s,a)\sim\rho^{p,\pi^b}_h(\cdot,\cdot)} b_h(s,a)+8H^3\max_{(s,a,h)\in\mathcal{Z}^{p,\pi^E}} b_h(s,a)\leqslant\epsilon.$$

We upper bound both the terms of the sum by $\epsilon/2$. The bound of the first term is analogous to the bound of the term provided in the proof of Theorem 5.1, so we will not rewrite it here; with regards to the other term, we have:

$$
\begin{aligned}
8H^3 \max_{(s,a^E,h)\in\mathcal{Z}^{p,\pi^E}} b_h(s,a^E)) &\overset{(1)}{=} 8H^3 \max_{(s,a^E,h)\in\mathcal{Z}^{p,\pi^E}} \sqrt{2\frac{\beta(N_h^b(s,a^E),\delta)}{N_h^b(s,a^E)}} \\
&\overset{(2)}{\leqslant} 8\sqrt{2}H^3\sqrt{\beta(\tau^b,\delta)} \max_{(s,a^E,h)\in\mathcal{Z}^{p,\pi^E}} \sqrt{\frac{1}{N_h^b(s,a^E)}} \\
&\overset{(3)}{\leqslant} 8\sqrt{2}H^3\sqrt{\beta(\tau^b,\delta)} \max_{(s,a^E,h)\in\mathcal{Z}^{p,\pi^E}} \sqrt{c_4\frac{\ln\frac{|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\tau^b\rho_h^{p,\pi^b}(s,a^E)}} \\
&\overset{(4)}{=} c_5 H^3 \sqrt{\frac{\beta(\tau^b,\delta)\ln\frac{|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\tau^b\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^E}}}} \leqslant \epsilon/2,
\end{aligned}
$$

where at (1) we have used the definition of the b terms, at (2) we have upper bounded $\beta(N_{\bar{h}}^b(\bar{s},a^E),\delta)\leqslant\beta(\tau^b,\delta)$ for all $(\bar{s},\bar{h})\in\mathcal{S}^{p,\pi^E}$, at (3) we use event $\mathcal{E}_4$, at (4) we define $c_5 := 8\sqrt{2c_4}$ and we use the definition of $\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^E}}$.

Similarly to the proof of Theorem 5.1, we apply Lemma J.3 to ($c_6 := 4c_5^2$):

$$
\tau^b \geqslant c_6\frac{H^6\ln\frac{|\mathcal{Z}^{p,\pi^b}|}{\delta}}{\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^E}}\epsilon^2}\Big(\ln\frac{4|\mathcal{Z}^{p,\pi^b}|}{\delta}+(|\mathcal{S}_{\max}^{p,\pi^b}|-1)\ln\big(e(1+\tau^b/(|\mathcal{S}_{\max}^{p,\pi^b}|-1))\big)\Big),
$$

to obtain:

$$
\tau^b \leqslant \widetilde{\mathcal{O}}\left(\frac{H^6\ln\frac{1}{\delta}}{\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^E}}\epsilon^2}\left(\ln\frac{1}{\delta}+|\mathcal{S}_{\max}^{p,\pi^b}|\right)\right).
$$

We can do the same for the superset through Lemma F.9. By combining the various bounds, we get the result. □

Thanks to Lemma F.10 and Lemma F.11, we can conclude the proof of Theorem 6.2.

Theorem 6.2. *Under the conditions of Theorem 6.1,* **PIRLO** *is (ϵ,δ)-PAC for d_∞-IRL with a sample complexity at most:*

$$
\begin{aligned}
\tau^b \leqslant \widetilde{\mathcal{O}}\Bigg(&\frac{H^4\ln\frac{1}{\delta}}{\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}}\epsilon^2}\left(\ln\frac{1}{\delta}+S_{\max}^{p,\pi^b}\right) \\
&+\frac{H^6\ln\frac{1}{\delta}}{\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^E}}\epsilon^2}\left(\ln\frac{1}{\delta}+S_{\max}^{p,\pi^b}\right)+\frac{\ln\frac{1}{\delta}}{\ln\frac{1}{1-\rho_{\min}^{\pi^b,\mathcal{Z}^{p,\pi^b}}}}\Bigg),
\end{aligned}
$$

and τ^E is bounded as in Theorem 5.1. Furthermore, **PIRLO** *is inclusion monotonic.*

Proof Sketch. The proof is analogous to that of Theorem 6.1. The only difference is that we use Lemma F.10 and Lemma F.10, and that we follow the proof of Theorem 5.2 instead of that of Theorem 5.1 to bound the first term of:

$$
2H^2 \max_{(s,a,h)\in\mathcal{Z}^{p,\pi^b}} b_h(s,a)+4H^3 \max_{(s,a,h)\in\mathcal{Z}^{p,\pi^E}} b_h(s,a)\leqslant\epsilon.
$$

□

F.4. Sample complexity for PIRLO with additional requirements

In the proofs of Theorem 6.1 and Theorem 6.2, we have used reward choice lemmas that set $\widehat{r}_h(s,a)\neq r_h(s,a)$ in $(s,a,h)\notin\mathcal{Z}^{p,\pi^b}$. However, it might be interesting to relate directly the error in the estimation of the transition model with the difference

in the reward functions, so that where we do not have samples we have zero error. We would like to have $\widehat{r}_h(s,a)=r_h(s,a)$ in $(s,a,h)\notin\mathcal{Z}^{p,\pi^b}$. Notice that this property is satisfied in the proofs of Theorem 5.1 and Theorem 5.2. Moreover, for a notion of distance other than d or d_∞, the condition $\widehat{r}_h(s,a)=r_h(s,a)$ in $(s,a,h)\notin\mathcal{Z}^{p,\pi^b}$ might even be needed. Therefore, in this section, we provide reward choice lemmas that satisfy this property, and we show that this selection ends up in a H^8 dependence in the sample complexity instead of H^6.

Lemma F.12 (Reward Choice Subset)**.** *Under good event $\mathcal{E}$, for any $r\in\mathcal{R}^{\cap}_{p,\pi^E}$, the reward $\widehat{r}$ constructed (recursively) as:*

$$\begin{cases}\widehat{r}_h(s,a^E)=r_h(s,a^E)+\sum\limits_{s'\in\mathcal{S}}p_h(s'|s,a^E)V^{\pi^E}_{h+1}(s';p,r)-\sum\limits_{s'\in\mathcal{S}}\widetilde{p}^m_h(s'|s,a^E)V^{\pi^E}_{h+1}(s';\widetilde{p}^m,\widehat{r})\\ \qquad\qquad+\max\limits_{s'\in\mathcal{S}}V^{\widetilde{\pi}^M}_{h+1}(s';\widetilde{p}^M,\widehat{r})-\max\limits_{s'\in\mathcal{S}}V^{\pi^M}_{h+1}(s';p^M,r), \qquad \forall(s,h)\in\mathcal{S}^{p,\pi^E}\\ \widehat{r}_h(s,a)=r_h(s,a), \quad \forall(s,a,h)\notin\mathcal{Z}^{p,\pi^b}\\ \widehat{r}_h(s,a)=r_h(s,a)+\sum\limits_{s'\in\mathcal{S}}p^M_h(s'|s,a)V^{\pi^M}_{h+1}(s';p^M,r)-\sum\limits_{s'\in\mathcal{S}}\widetilde{p}^M_h(s'|s,a)V^{\widetilde{\pi}^M}_{h+1}(s';\widetilde{p}^M,\widehat{r}), \quad \textit{otherwise}\end{cases},$$

belongs to $\widetilde{\mathcal{R}}^{\cap}$.

Proof. By definition of $\widetilde{\mathcal{R}}^{\cap}$, the reward $\widehat{r}$ belongs to $\widetilde{\mathcal{R}}^{\cap}$ if and only if:

$$\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\backslash\{a^E\}:Q^{\pi^E}_h(s,a^E;\widetilde{p}^m,\widehat{r})\geqslant Q^{\widetilde{\pi}^M}_h(s,a;\widetilde{p}^M,\widehat{r}).$$

By hypothesis, $r\in\mathcal{R}^{\cap}_{p,\pi^E}$, therefore:

$$\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\backslash\{a^E\}:Q^{\pi^E}_h(s,a^E;p,r)\geqslant Q^{\pi^M}_h(s,a;p^M,r),$$

thus, if we show that $\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\backslash\{a^E\}$, it holds that:

$$Q^{\pi^E}_h(s,a^E;\widetilde{p}^m,\widehat{r})-Q^{\widetilde{\pi}^M}_h(s,a;\widetilde{p}^M,\widehat{r})\geqslant Q^{\pi^E}_h(s,a^E;p,r)-Q^{\pi^M}_h(s,a;p^M,r),$$

then we are done.

Let us begin with triples $(s,a,h)\notin\mathcal{Z}^{p,\pi^b}$ such that $(s,h)\in\mathcal{S}^{p,\pi^E}$. By rearranging the terms in the definition of $\widehat{r}$, we observe that:

$$\begin{aligned}&\widehat{r}_h(s,a^E)+\sum_{s'\in\mathcal{S}}\widetilde{p}^m_h(s'|s,a^E)V^{\pi^E}_{h+1}(s';\widetilde{p}^m,\widehat{r})=r_h(s,a^E)+\sum_{s'\in\mathcal{S}}p_h(s'|s,a^E)V^{\pi^E}_{h+1}(s';p,r)\\ &\qquad+\max_{s'\in\mathcal{S}}V^{\widetilde{\pi}^M}_{h+1}(s';\widetilde{p}^M,\widehat{r})-\max_{s'\in\mathcal{S}}V^{\pi^M}_{h+1}(s';p^M,r)\\ &\iff Q^{\pi^E}_h(s,a^E;\widetilde{p}^m,\widehat{r})=Q^{\pi^E}_h(s,a^E;p,r)+\max_{s'\in\mathcal{S}}V^{\widetilde{\pi}^M}_{h+1}(s';\widetilde{p}^M,\widehat{r})-\max_{s'\in\mathcal{S}}V^{\pi^M}_{h+1}(s';p^M,r)\pm\widehat{r}_h(s,a)\\ &\overset{(1)}{\iff} Q^{\pi^E}_h(s,a^E;\widetilde{p}^m,\widehat{r})=Q^{\pi^E}_h(s,a^E;p,r)+\widehat{r}_h(s,a)+\max_{s'\in\mathcal{S}}V^{\widetilde{\pi}^M}_{h+1}(s';\widetilde{p}^M,\widehat{r})-\left(r_h(s,a)+\max_{s'\in\mathcal{S}}V^{\pi^M}_{h+1}(s';p^M,r)\right)\\ &\iff Q^{\pi^E}_h(s,a^E;\widetilde{p}^m,\widehat{r})=Q^{\pi^E}_h(s,a^E;p,r)+Q^{\widetilde{\pi}^M}_h(s,a;\widetilde{p}^M,\widehat{r})-Q^{\pi^M}_h(s,a;p^M,r)\\ &\implies Q^{\pi^E}_h(s,a^E;\widetilde{p}^m,\widehat{r})-Q^{\widetilde{\pi}^M}_h(s,a;\widetilde{p}^M,\widehat{r})\geqslant Q^{\pi^E}_h(s,a^E;p,r)-Q^{\pi^M}_h(s,a;p^M,r),\end{aligned}$$

where at (1) we have used that $\widehat{r}_h(s,a)=r_h(s,a)$ by definition.

Now, consider any other triple $(s,a,h)\in\mathcal{Z}^{p,\pi^b}\backslash\mathcal{Z}^{p,\pi^E}$ such that $(s,h)\in\mathcal{S}^{p,\pi^E}$. By rearranging the terms, we obtain:

$$Q^{\widetilde{\pi}^M}_h(s,a;\widetilde{p}^M,\widehat{r})=Q^{\pi^M}_h(s,a;p^M,r), \tag{44}$$

therefore, it suffices to show that

$$Q^{\pi^E}_h(s,a^E;\widetilde{p}^m,\widehat{r})\geqslant Q^{\pi^E}_h(s,a^E;p,r).$$

By using again the definition of $\widehat{r}$ for $(s,a,h)\in\mathcal{Z}^{p,\pi^E}$, we know that:

$$Q^{\pi^E}_h(s,a^E;\widetilde{p}^m,\widehat{r})=Q^{\pi^E}_h(s,a^E;p,r)+\max_{s'\in\mathcal{S}}V^{\widetilde{\pi}^M}_{h+1}(s';\widetilde{p}^M,\widehat{r})-\max_{s'\in\mathcal{S}}V^{\pi^M}_{h+1}(s';p^M,r), \tag{45}$$

therefore, if we show that

$$\max_{s'\in\mathcal{S}} V_{h+1}^{\widetilde{\pi}^M}(s';\widetilde{p}^M,\widehat{r}) \geqslant \max_{s'\in\mathcal{S}} V_{h+1}^{\pi^M}(s';p^M,r),$$

then we are done. We do it by induction. At stage $H-1$, we have that:

$$\begin{aligned}
\max_{s'\in\mathcal{S}} V_H^{\widetilde{\pi}^M}(s';\widetilde{p}^M,\widehat{r}) &= \max_{s'\in\mathcal{S}} \mathop{\mathbb{E}}_{a'\sim\widetilde{\pi}_H^M(\cdot|s')} \widehat{r}_H(s',a') \\
&\overset{(1)}{=} \max_{s'\in\mathcal{S}} \mathop{\mathbb{E}}_{a'\sim\pi_H^M(\cdot|s')} r_H(s',a') \\
&= \max_{s'\in\mathcal{S}} V_H^{\pi^M}(s';p^M,r),
\end{aligned}$$

where at (1) we have used the definition of $\widehat{r}$ at stage H, and the definitions of $\widetilde{\pi}^M$ and π^M. We make the inductive hypothesis that, at stage $h+1$, it holds that $\max_{s'\in\mathcal{S}} V_{h+2}^{\widetilde{\pi}^M}(s';\widetilde{p}^M,\widehat{r}) \geqslant \max_{s'\in\mathcal{S}} V_{h+2}^{\pi^M}(s';p^M,r)$, and we consider stage h:

$$\begin{aligned}
\max_{s'\in\mathcal{S}} V_{h+1}^{\widetilde{\pi}^M}(s';\widetilde{p}^M,\widehat{r}) &\overset{(1)}{=} \max\Big\{ \max_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}} Q_{h+1}^{\pi^E}(s',a^E;\widetilde{p}^M,\widehat{r}),\ \max_{s'\notin\mathcal{S}_{h+1}^{p,\pi^E}} \max_{a'\in\mathcal{A}} Q_{h+1}^{\widetilde{\pi}^M}(s',a';\widetilde{p}^M,\widehat{r}) \Big\} \\
&\overset{(2)}{\geqslant} \max\Big\{ \max_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}} Q_{h+1}^{\pi^E}(s',a^E;\widetilde{p}^m,\widehat{r}),\ \max_{s'\notin\mathcal{S}_{h+1}^{p,\pi^E}} \max_{a'\in\mathcal{A}} Q_{h+1}^{\widetilde{\pi}^M}(s',a';\widetilde{p}^M,\widehat{r}) \Big\} \\
&\overset{(3)}{\geqslant} \max\Big\{ \max_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}} Q_{h+1}^{\pi^E}(s',a^E;p,r),\ \max_{s'\notin\mathcal{S}_{h+1}^{p,\pi^E}} \max_{a'\in\mathcal{A}} Q_{h+1}^{\widetilde{\pi}^M}(s',a';\widetilde{p}^M,\widehat{r}) \Big\} \\
&= \max\Big\{ \max_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}} Q_{h+1}^{\pi^E}(s',a^E;p,r), \\
&\qquad \max_{s'\notin\mathcal{S}_{h+1}^{p,\pi^E}} \max\Big\{ \max_{a'\in\mathcal{A}:(s',a',h+1)\in\mathcal{Z}^{p,\pi^b}} Q_{h+1}^{\widetilde{\pi}^M}(s',a';\widetilde{p}^M,\widehat{r}),\ \max_{a'\in\mathcal{A}:(s',a',h+1)\notin\mathcal{Z}^{p,\pi^b}} Q_{h+1}^{\widetilde{\pi}^M}(s',a';\widetilde{p}^M,\widehat{r}) \Big\} \Big\} \\
&\overset{(4)}{=} \max\Big\{ \max_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}} Q_{h+1}^{\pi^E}(s',a^E;p,r), \\
&\qquad \max_{s'\notin\mathcal{S}_{h+1}^{p,\pi^E}} \max\Big\{ \max_{a'\in\mathcal{A}:(s',a',h+1)\in\mathcal{Z}^{p,\pi^b}} Q_{h+1}^{\pi^M}(s',a';p^M,r), \\
&\qquad\qquad \max_{a'\in\mathcal{A}:(s',a',h+1)\notin\mathcal{Z}^{p,\pi^b}} \widehat{r}_{h+1}(s',a') + \max_{s''\in\mathcal{S}} V_{h+2}^{\widetilde{\pi}^M}(s'';\widetilde{p}^M,\widehat{r}) \Big\} \Big\} \\
&\overset{(5)}{\geqslant} \max\Big\{ \max_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}} Q_{h+1}^{\pi^E}(s',a^E;p,r), \\
&\qquad \max_{s'\notin\mathcal{S}_{h+1}^{p,\pi^E}} \max\Big\{ \max_{a'\in\mathcal{A}:(s',a',h+1)\in\mathcal{Z}^{p,\pi^b}} Q_{h+1}^{\pi^M}(s',a';p^M,r), \\
&\qquad\qquad \max_{a'\in\mathcal{A}:(s',a',h+1)\notin\mathcal{Z}^{p,\pi^b}} r_{h+1}(s',a') + \max_{s''\in\mathcal{S}} V_{h+2}^{\pi^M}(s'';p^M,r) \Big\} \Big\} \\
&= \max\Big\{ \max_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}} Q_{h+1}^{\pi^E}(s',a^E;p,r), \\
&\qquad \max_{s'\notin\mathcal{S}_{h+1}^{p,\pi^E}} \max\Big\{ \max_{a'\in\mathcal{A}:(s',a',h+1)\in\mathcal{Z}^{p,\pi^b}} Q_{h+1}^{\pi^M}(s',a';p^M,r),\ \max_{a'\in\mathcal{A}:(s',a',h+1)\notin\mathcal{Z}^{p,\pi^b}} Q_{h+1}^{\pi^M}(s',a';p^M,r) \Big\} \Big\} \\
&= \max\Big\{ \max_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}} Q_{h+1}^{\pi^E}(s',a^E;p,r),\ \max_{s'\notin\mathcal{S}_{h+1}^{p,\pi^E}} \max_{a'\in\mathcal{A}} Q_{h+1}^{\pi^M}(s',a';p^M,r) \Big\} \\
&= \max_{s'\in\mathcal{S}} V_{h+1}^{\pi^M}(s';p^M,r),
\end{aligned}$$

where at (1) we use the definition of $\widetilde{\pi}^M$, at (2) we use the definition of $\widetilde{p}^m$ and $\widetilde{p}^M$, at (3) we use the inductive hypothesis along with Eq. 45, at (4) we use Eq. 44 and the definition of Q-function, at (5) we use the definition of $\widehat{r}$ and the inductive hypothesis.

This concludes the proof. □

Lemma F.13 (Reward Choice Superset)**.** *Under good event $\mathcal{E}$, for any $\widehat{r}\in\widetilde{\mathcal{R}}^{\cup}$, the reward r constructed (recursively) as:*

$$\begin{cases} r_h(s,a^E)=\widehat{r}_h(s,a^E)+\sum\limits_{s'\in\mathcal{S}}\widetilde{p}_h^M(s'|s,a^E)V_{h+1}^{\pi^E}(s';\widetilde{p}^M,\widehat{r})-\sum\limits_{s'\in\mathcal{S}}p_h(s'|s,a^E)V_{h+1}^{\pi^E}(s';p,r) \\ \qquad\qquad +\min\limits_{s'\in\mathcal{S}}V_{h+1}^{\pi^m}(s';p^m,r)-\min\limits_{s'\in\mathcal{S}}V_{h+1}^{\widetilde{\pi}^m}(s';\widetilde{p}^m,\widehat{r}), \qquad \forall(s,h)\in\mathcal{S}^{p,\pi^E} \\ r_h(s,a)=\widehat{r}_h(s,a), \quad \forall(s,a,h)\notin\mathcal{Z}^{p,\pi^b} \\ r_h(s,a)=\widehat{r}_h(s,a)+\sum\limits_{s'\in\mathcal{S}}\widetilde{p}_h^m(s'|s,a)V_{h+1}^{\widetilde{\pi}^m}(s';\widetilde{p}^m,\widehat{r})-\sum\limits_{s'\in\mathcal{S}}p_h^m(s'|s,a)V_{h+1}^{\pi^m}(s';p^m,r), \quad \textit{otherwise} \end{cases},$$

belongs to $\mathcal{R}^{\cup}_{p,\pi^E}$.

Proof. By definition of $\mathcal{R}^{\cup}_{p,\pi^E}$, a sufficient condition for having the reward r belong to $\mathcal{R}^{\cup}_{p,\pi^E}$ is:

$$\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\backslash\{a^E\}:Q_h^{\pi^E}(s,a^E;p,r)\geqslant Q_h^{\pi^m}(s,a;p^m,r).$$

By hypothesis, $r\in\widetilde{\mathcal{R}}^{\cup}$, therefore:

$$\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\backslash\{a^E\}:Q_h^{\pi^E}(s,a^E;\widetilde{p}^M,\widehat{r})\geqslant Q_h^{\widetilde{\pi}^m}(s,a;\widetilde{p}^m,\widehat{r}),$$

thus, if we show that $\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\backslash\{a^E\}$, it holds that:

$$Q_h^{\pi^E}(s,a^E;p,r)-Q_h^{\pi^m}(s,a;p^m,r)\geqslant Q_h^{\pi^E}(s,a^E;\widetilde{p}^M,\widehat{r})-Q_h^{\widetilde{\pi}^m}(s,a;\widetilde{p}^m,\widehat{r}),$$

then we are done.

Let us begin with triples $(s,a,h)\notin\mathcal{Z}^{p,\pi^b}$ such that $(s,h)\in\mathcal{S}^{p,\pi^E}$. By rearranging the terms in the definition of r, we observe that:

$$\begin{aligned}
&\widehat{r}_h(s,a^E)+\sum_{s'\in\mathcal{S}}\widetilde{p}_h^M(s'|s,a^E)V_{h+1}^{\pi^E}(s';\widetilde{p}^M,\widehat{r})=r_h(s,a^E)+\sum_{s'\in\mathcal{S}}p_h(s'|s,a^E)V_{h+1}^{\pi^E}(s';p,r)\\
&\qquad+\min_{s'\in\mathcal{S}}V_{h+1}^{\widetilde{\pi}^m}(s';\widetilde{p}^m,\widehat{r})-\min_{s'\in\mathcal{S}}V_{h+1}^{\pi^m}(s';p^m,r)\\
&\iff Q_h^{\pi^E}(s,a^E;\widetilde{p}^M,\widehat{r})=Q_h^{\pi^E}(s,a^E;p,r)+\min_{s'\in\mathcal{S}}V_{h+1}^{\widetilde{\pi}^m}(s';\widetilde{p}^m,\widehat{r})-\min_{s'\in\mathcal{S}}V_{h+1}^{\pi^m}(s';p^m,r)\pm r_h(s,a)\\
&\overset{(1)}{\iff} Q_h^{\pi^E}(s,a^E;\widetilde{p}^M,\widehat{r})=Q_h^{\pi^E}(s,a^E;p,r)+\widehat{r}_h(s,a)+\min_{s'\in\mathcal{S}}V_{h+1}^{\widetilde{\pi}^m}(s';\widetilde{p}^m,\widehat{r})-\big(r_h(s,a)+\min_{s'\in\mathcal{S}}V_{h+1}^{\pi^m}(s';p^m,r)\big)\\
&\iff Q_h^{\pi^E}(s,a^E;\widetilde{p}^M,\widehat{r})=Q_h^{\pi^E}(s,a^E;p,r)+Q_h^{\widetilde{\pi}^m}(s,a;\widetilde{p}^m,\widehat{r})-Q_h^{\pi^m}(s,a;p^m,r)\\
&\implies Q_h^{\pi^E}(s,a^E;\widetilde{p}^M,\widehat{r})-Q_h^{\widetilde{\pi}^m}(s,a;\widetilde{p}^m,\widehat{r})\geqslant Q_h^{\pi^E}(s,a^E;p,r)-Q_h^{\pi^m}(s,a;p^m,r),
\end{aligned}$$

where at (1) we have used that $r_h(s,a)=\widehat{r}_h(s,a)$ by definition.

Now, consider any other triple $(s,a,h)\in\mathcal{Z}^{p,\pi^b}\backslash\mathcal{Z}^{p,\pi^E}$ such that $(s,h)\in\mathcal{S}^{p,\pi^E}$. By rearranging the terms, we obtain:

$$Q_h^{\widetilde{\pi}^m}(s,a;\widetilde{p}^m,\widehat{r})=Q_h^{\pi^m}(s,a;p^m,r), \tag{46}$$

therefore, it suffices to show that

$$Q_h^{\pi^E}(s,a^E;p,r)\geqslant Q_h^{\pi^E}(s,a^E;\widetilde{p}^M,\widehat{r}).$$

By using again the definition of $\widehat{r}$ for $(s,a,h)\in\mathcal{Z}^{p,\pi^E}$, we know that:

$$Q_h^{\pi^E}(s,a^E;\widetilde{p}^M,\widehat{r})=Q_h^{\pi^E}(s,a^E;p,r)+\min_{s'\in\mathcal{S}}V_{h+1}^{\widetilde{\pi}^m}(s';\widetilde{p}^m,\widehat{r})-\min_{s'\in\mathcal{S}}V_{h+1}^{\pi^m}(s';p^m,r), \tag{47}$$

therefore, if we show that

$$\min_{s'\in\mathcal{S}}V_{h+1}^{\widetilde{\pi}^m}(s';\widetilde{p}^m,\widehat{r})\leqslant\min_{s'\in\mathcal{S}}V_{h+1}^{\pi^m}(s';p^m,r),$$

then we are done. We do it by induction. At stage $H-1$, we have that:

$$\begin{aligned}\min_{s'\in\mathcal{S}} V_H^{\widetilde{\pi}^m}(s';\widetilde{p}^m,\widehat{r}) &= \min_{s'\in\mathcal{S}} \mathop{\mathbb{E}}_{a'\sim\widetilde{\pi}_H^m(\cdot|s')} \widehat{r}_H(s',a') \\ &\overset{(1)}{=} \min_{s'\in\mathcal{S}} \mathop{\mathbb{E}}_{a'\sim\pi_H^m(\cdot|s')} r_H(s',a') \\ &= \min_{s'\in\mathcal{S}} V_H^{\pi^m}(s';p^m,r),\end{aligned}$$

where at (1) we have used the definition of $\widehat{r}$ at stage H, and also the definitions of $\widetilde{\pi}^m$ and π^m. We make the inductive hypothesis that, at stage $h+1$, it holds that $\min_{s'\in\mathcal{S}} V_{h+2}^{\widetilde{\pi}^m}(s';\widetilde{p}^m,\widehat{r}) \leqslant \min_{s'\in\mathcal{S}} V_{h+2}^{\pi^m}(s';p^m,r)$, and we consider stage h:

$$\begin{aligned}\min_{s'\in\mathcal{S}} V_{h+1}^{\widetilde{\pi}^m}(s';\widetilde{p}^m,\widehat{r}) &\overset{(1)}{=} \min\Big\{ \min_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}} Q_{h+1}^{\pi^E}(s',a^E;\widetilde{p}^m,\widehat{r}),\ \min_{s'\notin\mathcal{S}_{h+1}^{p,\pi^E}} \max_{a'\in\mathcal{A}} Q_{h+1}^{\widetilde{\pi}^m}(s',a';\widetilde{p}^m,\widehat{r}) \Big\} \\ &\overset{(2)}{\leqslant} \min\Big\{ \min_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}} Q_{h+1}^{\pi^E}(s',a^E;\widetilde{p}^M,\widehat{r}),\ \min_{s'\notin\mathcal{S}_{h+1}^{p,\pi^E}} \max_{a'\in\mathcal{A}} Q_{h+1}^{\widetilde{\pi}^m}(s',a';\widetilde{p}^m,\widehat{r}) \Big\} \\ &\overset{(3)}{\leqslant} \min\Big\{ \min_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}} Q_{h+1}^{\pi^E}(s',a^E;p,r),\ \min_{s'\notin\mathcal{S}_{h+1}^{p,\pi^E}} \max_{a'\in\mathcal{A}} Q_{h+1}^{\widetilde{\pi}^m}(s',a';\widetilde{p}^m,\widehat{r}) \Big\} \\ &= \min\Big\{ \min_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}} Q_{h+1}^{\pi^E}(s',a^E;p,r), \\ &\qquad \min_{s'\notin\mathcal{S}_{h+1}^{p,\pi^E}} \max\Big\{ \max_{a'\in\mathcal{A}:(s',a',h+1)\in\mathcal{Z}^{p,\pi^b}} Q_{h+1}^{\widetilde{\pi}^m}(s',a';\widetilde{p}^m,\widehat{r}),\ \max_{a'\in\mathcal{A}:(s',a',h+1)\notin\mathcal{Z}^{p,\pi^b}} Q_{h+1}^{\widetilde{\pi}^m}(s',a';\widetilde{p}^m,\widehat{r}) \Big\} \Big\} \\ &\overset{(4)}{=} \min\Big\{ \min_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}} Q_{h+1}^{\pi^E}(s',a^E;p,r), \\ &\qquad \min_{s'\notin\mathcal{S}_{h+1}^{p,\pi^E}} \max\Big\{ \max_{a'\in\mathcal{A}:(s',a',h+1)\in\mathcal{Z}^{p,\pi^b}} Q_{h+1}^{\pi^m}(s',a';p^m,r), \\ &\qquad\qquad \max_{a'\in\mathcal{A}:(s',a',h+1)\notin\mathcal{Z}^{p,\pi^b}} \widehat{r}_{h+1}(s',a') + \min_{s''\in\mathcal{S}} V_{h+2}^{\widetilde{\pi}^m}(s'';\widetilde{p}^m,\widehat{r}) \Big\} \Big\} \\ &\overset{(5)}{\leqslant} \min\Big\{ \min_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}} Q_{h+1}^{\pi^E}(s',a^E;p,r), \\ &\qquad \min_{s'\notin\mathcal{S}_{h+1}^{p,\pi^E}} \max\Big\{ \max_{a'\in\mathcal{A}:(s',a',h+1)\in\mathcal{Z}^{p,\pi^b}} Q_{h+1}^{\pi^m}(s',a';p^m,r), \\ &\qquad\qquad \max_{a'\in\mathcal{A}:(s',a',h+1)\notin\mathcal{Z}^{p,\pi^b}} r_{h+1}(s',a') + \max_{s''\in\mathcal{S}} V_{h+2}^{\pi^m}(s'';p^m,r) \Big\} \Big\} \\ &= \min\Big\{ \min_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}} Q_{h+1}^{\pi^E}(s',a^E;p,r), \\ &\qquad \min_{s'\notin\mathcal{S}_{h+1}^{p,\pi^E}} \max\Big\{ \max_{a'\in\mathcal{A}:(s',a',h+1)\in\mathcal{Z}^{p,\pi^b}} Q_{h+1}^{\pi^m}(s',a';p^m,r),\ \max_{a'\in\mathcal{A}:(s',a',h+1)\notin\mathcal{Z}^{p,\pi^b}} Q_{h+1}^{\pi^m}(s',a';p^m,r) \Big\} \Big\} \\ &= \min\Big\{ \min_{s'\in\mathcal{S}_{h+1}^{p,\pi^E}} Q_{h+1}^{\pi^E}(s',a^E;p,r),\ \min_{s'\notin\mathcal{S}_{h+1}^{p,\pi^E}} \max_{a'\in\mathcal{A}} Q_{h+1}^{\pi^m}(s',a';p^m,r) \Big\} \\ &= \min_{s'\in\mathcal{S}} V_{h+1}^{\pi^m}(s';p^m,r),\end{aligned}$$

where at (1) we use the definition of $\widetilde{\pi}^m$, at (2) we use the definition of $\widetilde{p}^m$ and $\widetilde{p}^M$, at (3) we use the inductive hypothesis along with Eq. 47, at (4) we use Eq. 46 and the Bellman's equation, at (5) we use the definition of r and the inductive hypothesis.

This concludes the proof. □

F.4.1. Lemmas for Theorem F.18

We can exploit Lemma F.12 to bound the error for the subset.

Lemma F.14 (Performance Decomposition Subset)**.** *Under good event $\mathcal{E}$, it holds that:*

$$\mathcal{H}_d(\mathcal{R}^{\cap}_{p,\pi^E}, \widetilde{\mathcal{R}}^{\cap}) \leqslant 2H \sum_{h\in[\![H]\!]} \mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} b_h(s,a) + 4H^4 \max_{(s,h)\in\mathcal{S}^{p,\pi^E}} b_h(s,a^E).$$

Proof. We can write:

$$\begin{aligned}
\mathcal{H}_d(\mathcal{R}^{\cap}_{p,\pi^E}, \widetilde{\mathcal{R}}^{\cap}) &:= \max\{\sup_{r\in\mathcal{R}^{\cap}_{p,\pi^E}} \inf_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cap}} d(r,\widetilde{r}), \sup_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cap}} \inf_{r\in\mathcal{R}^{\cap}_{p,\pi^E}} d(r,\widetilde{r})\} \\
&\overset{(1)}{=} \sup_{r\in\mathcal{R}^{\cap}_{p,\pi^E}} \inf_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cap}} d(r,\widetilde{r}) \\
&=: \sup_{r\in\mathcal{R}^{\cap}_{p,\pi^E}} \inf_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cap}} \frac{1}{M} \sum_{h\in[\![H]\!]} \left(\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} \left|r_h(s,a) - \widetilde{r}_h(s,a)\right| + \max_{(s,a)\notin\mathcal{Z}_h^{p,\pi^b}} \left|r_h(s,a) - \widetilde{r}_h(s,a)\right| \right) \\
&\overset{(2)}{\leqslant} \sup_{r\in\mathcal{R}^{\cap}_{p,\pi^E}} \frac{1}{M} \sum_{h\in[\![H]\!]} \left(\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} \left|r_h(s,a) - \widehat{r}_h(s,a)\right| + \underbrace{\max_{(s,a)\notin\mathcal{Z}_h^{p,\pi^b}} \left|r_h(s,a) - \widehat{r}_h(s,a)\right|}_{=0} \right) \\
&= \sup_{r\in\mathcal{R}^{\cap}_{p,\pi^E}} \frac{1}{M} \sum_{h\in[\![H]\!]} \mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} \left|r_h(s,a) - \widehat{r}_h(s,a)\right|
\end{aligned} \tag{48}$$

where at (1) we have used that, under good event $\mathcal{E}$, $\widetilde{\mathcal{R}}^{\cap} \subseteq \mathcal{R}^{\cap}_{p,\pi^E}$, and at (2) we apply Lemma F.12, by denoting with $\widehat{r}$ the chosen reward from $\widetilde{\mathcal{R}}^{\cap}$.

In the following, it is useful to denote, for any $h\in[\![H]\!]$:

$$X_h := \max_{s'\in\mathcal{S}} \left|V^{\widetilde{\pi}^M}_{h+1}(s';\widetilde{p}^M,\widehat{r}) - V^{\pi^M}_{h+1}(s';p^M,r)\right|.$$

Let us consider any $(s,h)\in\mathcal{S}^{p,\pi^E}$. The difference between the rewards of expert's action can be bounded by:

$$\begin{aligned}
\left|\widehat{r}_h(s,a^E) - r_h(s,a^E)\right| &\overset{(1)}{\leqslant} \Big| \mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)} V^{\pi^E}_{h+1}(s';p,r) - \mathop{\mathbb{E}}_{s'\sim\widetilde{p}^m_h(\cdot|s,a^E)} V^{\pi^E}_{h+1}(s';\widetilde{p}^m,\widehat{r})\Big| \\
&\quad + \left|\max_{s'\in\mathcal{S}} V^{\widetilde{\pi}^M}_{h+1}(s';\widetilde{p}^M,\widehat{r}) - \max_{s'\in\mathcal{S}} V^{\pi^M}_{h+1}(s';p^M,r)\right| \\
&\overset{(2)}{\leqslant} \Big| \mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)} V^{\pi^E}_{h+1}(s';p,r) - \mathop{\mathbb{E}}_{s'\sim\widetilde{p}^m_h(\cdot|s,a^E)} V^{\pi^E}_{h+1}(s';\widetilde{p}^m,\widehat{r}) \pm \mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)} V^{\pi^E}_{h+1}(s';\widetilde{p}^m,\widehat{r})\Big| \\
&\quad + \max_{s'\in\mathcal{S}} \left|V^{\widetilde{\pi}^M}_{h+1}(s';\widetilde{p}^M,\widehat{r}) - V^{\pi^M}_{h+1}(s';p^M,r)\right| \\
&\overset{(3)}{\leqslant} MH \left|\left|p_h(\cdot|s,a^E) - \widetilde{p}^m_h(\cdot|s,a^E)\right|\right|_1 + \mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)} \left|V^{\pi^E}_{h+1}(s';p,r) - V^{\pi^E}_{h+1}(s';\widetilde{p}^m,\widehat{r})\right| + X_h \\
&\overset{(4)}{\leqslant} 2MHb_h(s,a^E) + \mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)} \left|Q^{\pi^E}_{h+1}(s',a^E;p,r) - Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^m,\widehat{r})\right| + X_h \\
&\overset{(5)}{=} 2MHb_h(s,a^E) + \mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)} \left|\max_{s''\in\mathcal{S}} V^{\pi^M}_{h+2}(s'';p^M,r) - \max_{s''\in\mathcal{S}} V^{\widetilde{\pi}^M}_{h+2}(s'';\widetilde{p}^M,\widehat{r})\right| + X_h \\
&\leqslant 2MHb_h(s,a^E) + \mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)} \left[X_{h+1}\right] + X_h \\
&\overset{(6)}{=} 2MHb_h(s,a^E) + X_h + X_{h+1},
\end{aligned}$$

where at (1) we use the definition of $\widehat{r}$ in Lemma F.12 and triangle inequality, at (2) we use that, for any pair f,g of real-valued functions, it holds that $|\max_x f(x) - \max_x g(x)| \leqslant \max_x |f(x) - g(x)|$, at (3) we apply triangle inequality twice, we recognize the definition of X_h, we upper bound the value function by MH, and we recognize the definition of ℓ_1 norm, at (4) we first use triangle inequality $\|p_h(\cdot|s,a^E) - \widetilde{p}^m_h(\cdot|s,a^E)\|_1 \leqslant \|p_h(\cdot|s,a^E) - \widehat{p}_h(\cdot|s,a^E)\|_1 + \|\widetilde{p}^m_h(\cdot|s,a^E) - \widehat{p}_h(\cdot|s,a^E)\|_1$, then we use Pinsker's inequality, event $\mathcal{E}_3$ from Lemma F.2, and the definition of $b_h(s,a^E)$; at (5) we use the definition of $\widehat{r}$

in Lemma F.12 in the form of Eq. 45, by noticing that the support of $p_h(\cdot|s,a^E)$ is contained in $\mathcal{S}^{p,\pi^E}$, and at (6) we realize that X_{h+1} depends only on h and not on s'.

In order to upper bound the term X_h, we write:

$$
\begin{aligned}
X_h &:= \max_{s'\in\mathcal{S}} \left| V^{\widetilde{\pi}^M}_{h+1}(s';\widetilde{p}^M,\widehat{r}) - V^{\pi^M}_{h+1}(s';p^M,r) \right| \\
&\overset{(1)}{=} \max\Bigg\{ \max_{s'\notin\mathcal{S}^{p,\pi^E}_{h+1}} \left| \max_{a'\in\mathcal{A}} Q^{\widetilde{\pi}^M}_{h+1}(s',a';\widetilde{p}^M,\widehat{r}) - \max_{a'\in\mathcal{A}} Q^{\pi^M}_{h+1}(s',a';p^M,r) \right|, \\
&\qquad \max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} \left| Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^M,\widehat{r}) - Q^{\pi^E}_{h+1}(s',a^E;p,r) \right| \Bigg\} \\
&\overset{(2)}{\leqslant} \max\Bigg\{ \max_{s'\notin\mathcal{S}^{p,\pi^E}_{h+1}} \max_{a'\in\mathcal{A}} \left| Q^{\widetilde{\pi}^M}_{h+1}(s',a';\widetilde{p}^M,\widehat{r}) - Q^{\pi^M}_{h+1}(s',a';p^M,r) \right|, \\
&\qquad \max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} \left| Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^M,\widehat{r}) - Q^{\pi^E}_{h+1}(s',a^E;p,r) \pm Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^m,\widehat{r}) \right| \Bigg\} \\
&\overset{(3)}{=} \max\Bigg\{ \max_{s'\notin\mathcal{S}^{p,\pi^E}_{h+1}} \max\Bigg\{ \max_{a'\in\mathcal{A}:(s',a',h+1)\in\mathcal{Z}^{p,\pi^b}} \Big| \underbrace{Q^{\widetilde{\pi}^M}_{h+1}(s',a';\widetilde{p}^M,\widehat{r}) - Q^{\pi^M}_{h+1}(s',a';p^M,r)}_{=0} \Big|, \\
&\qquad \max_{a'\in\mathcal{A}:(s',a',h+1)\notin\mathcal{Z}^{p,\pi^b}} \left| Q^{\widetilde{\pi}^M}_{h+1}(s',a';\widetilde{p}^M,\widehat{r}) - Q^{\pi^M}_{h+1}(s',a';p^M,r) \right| \Bigg\}, \\
&\qquad \max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} \left| Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^M,\widehat{r}) - Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^m,\widehat{r}) + \max_{s''\in\mathcal{S}} V^{\widetilde{\pi}^M}_{h+2}(s'';\widetilde{p}^M,\widehat{r}) - \max_{s''\in\mathcal{S}} V^{\pi^M}_{h+2}(s';p^M,r) \right| \Bigg\} \\
&\overset{(4)}{\leqslant} \max\Bigg\{ \max_{s'\notin\mathcal{S}^{p,\pi^E}_{h+1}} \max_{a'\in\mathcal{A}:(s',a',h+1)\notin\mathcal{Z}^{p,\pi^b}} \Big| \underbrace{\widehat{r}_{h+1}(s',a') - r_{h+1}(s',a')}_{=0} \Big| + \max_{s''\in\mathcal{S}} V^{\widetilde{\pi}^M}_{h+2}(s'';\widetilde{p}^M,\widehat{r}) - \max_{s''\in\mathcal{S}} V^{\pi^M}_{h+2}(s'';p^M,r) \Big|, \\
&\qquad \max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} \left| Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^M,\widehat{r}) - Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^m,\widehat{r}) \right| + X_{h+1} \Bigg\} \\
&\overset{(5)}{\leqslant} \max\left\{ X_{h+1}, \max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} \left(\left| \mathop{\mathbb{E}}_{s''\sim\widetilde{p}^M_{h+1}(\cdot|s',a^E)} V^{\pi^E}_{h+2}(s'';\widetilde{p}^M,\widehat{r}) - \mathop{\mathbb{E}}_{s''\sim\widetilde{p}^m_{h+1}(\cdot|s',a^E)} V^{\pi^E}_{h+2}(s'';\widetilde{p}^m,\widehat{r}) \right| + X_{h+1} \right) \right\} \\
&= X_{h+1} + \max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} \left| \mathop{\mathbb{E}}_{s''\sim\widetilde{p}^M_{h+1}(\cdot|s',a^E)} V^{\pi^E}_{h+2}(s'';\widetilde{p}^M,\widehat{r}) - \mathop{\mathbb{E}}_{s''\sim\widetilde{p}^m_{h+1}(\cdot|s',a^E)} V^{\pi^E}_{h+2}(s'';\widetilde{p}^m,\widehat{r}) \pm \mathop{\mathbb{E}}_{s''\sim\widetilde{p}^m_{h+1}(\cdot|s',a^E)} V^{\pi^E}_{h+2}(s'';\widetilde{p}^M,\widehat{r}) \right| \\
&\overset{(6)}{\leqslant} X_{h+1} + \max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} \left(MH \left\| \widetilde{p}^M_{h+1}(\cdot|s',a^E) - \widetilde{p}^m_{h+1}(\cdot|s',a^E) \right\|_1 + \mathop{\mathbb{E}}_{s''\sim\widetilde{p}^m_{h+1}(\cdot|s',a^E)} \left| V^{\pi^E}_{h+2}(s'';\widetilde{p}^M,\widehat{r}) - V^{\pi^E}_{h+2}(s'';\widetilde{p}^m,\widehat{r}) \right| \right) \\
&\overset{(7)}{\leqslant} X_{h+1} + \max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} \Bigg(2MHb_{h+1}(s',a^E) \\
&\qquad + \mathop{\mathbb{E}}_{s''\sim\widetilde{p}^m_{h+1}(\cdot|s',a^E)} \left| \mathop{\mathbb{E}}_{s'''\sim\widetilde{p}^M_{h+2}(\cdot|s'',a^E)} V^{\pi^E}_{h+3}(s''';\widetilde{p}^M,\widehat{r}) - \mathop{\mathbb{E}}_{s'''\sim\widetilde{p}^m_{h+2}(\cdot|s'',a^E)} V^{\pi^E}_{h+3}(s''';\widetilde{p}^m,\widehat{r}) \right| \Bigg) \\
&\overset{(8)}{\leqslant} X_{h+1} + 2MH \max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}} \sum_{h'\in[\![h+1,H-1]\!]} \mathop{\mathbb{E}}_{s''\sim\rho^{\widetilde{p}^m,\pi^E}_{h'}(\cdot|s_{h+1}=s')} b_{h'}(s'',a^E) \\
&\overset{(9)}{\leqslant} 2MH \sum_{h'\in[\![h,H-1]\!]} \max_{s'\in\mathcal{S}^{p,\pi^E}_{h'+1}} \sum_{h''\in[\![h'+1,H-1]\!]} \mathop{\mathbb{E}}_{s''\sim\rho^{\widetilde{p}^m,\pi^E}_{h''}(\cdot|s_{h'+1}=s')} b_{h''}(s'',a^E) \\
&\leqslant 2MH^3 \max_{(s',h')\in\mathcal{S}^{p,\pi^E}} b_{h'}(s',a^E),
\end{aligned}
$$

where at (1) we apply the Bellman's equation and the definition of $\widetilde{\pi}^M$ and π^M, at (2) we use that $|\max_x f(x) - \max_x g(x)| \leqslant \max_x |f(x)-g(x)|$, at (3) we use the definition of $\widehat{r}$ in the form of Eq. 44 and Eq.45, at (4) we apply the Bellman's equation and the definition of $\widehat{r}$ to recognize that $\widehat{r}_{h+1}(s',a') - r_{h+1}(s',a') = 0$; moreover, we apply triangle

inequality along with the usual bound $|\max_x f(x) - \max_x g(x)| \leqslant \max_x |f(x) - g(x)|$, and we recognize the definition of X_{h+1}. At (5) we proceed similarly as (4) and we use the Bellman optimality equation, and we observe that X_{h+1} does not depend on s'; at (6) we upper bound the value function by HM and recognize the ℓ_1-norm, at (7) we use the concentration bound of event $\mathcal{E}_3$ in Lemma F.2 (both $\widetilde{p}^M$ and $\widetilde{p}^m$ lie at a "distance" of b from $\widehat{p}$). At (8) we have unfolded the recursion to bound the difference of value functions between transition models $\widetilde{p}^m$ and $\widetilde{p}^M$, at (9) we have unfolded the recursion on the X terms.

Thanks to this expression, we can upper bound the difference of rewards in expert's action as:

$$\left|\widehat{r}_h(s,a^E) - r_h(s,a^E)\right| \leqslant 2MHb_h(s,a^E) + 4MH^3 \max_{(s',h')\in\mathcal{S}^{p,\pi^E}} b_{h'}(s',a^E).$$

With regards to visited $(s,a,h) \in \mathcal{Z}^{p,\pi^b} \setminus \mathcal{Z}^{p,\pi^E}$, we can write:

$$\begin{aligned}
\left|\widehat{r}_h(s,a) - r_h(s,a)\right| &= \Big|\mathop{\mathbb{E}}_{s'\sim\widetilde{p}_h^M(\cdot|s,a)} V_{h+1}^{\widetilde{\pi}^M}(s';\widetilde{p}^M,\widehat{r}) - \mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a)} V_{h+1}^{\pi^M}(s';p^M,r)\Big| \\
&\leqslant 2MHb_h(s,a) + \mathop{\mathbb{E}}_{s'\sim\widetilde{p}_h^M(\cdot|s,a)} \left|V_{h+1}^{\widetilde{\pi}^M}(s';\widetilde{p}^M,\widehat{r}) - V_{h+1}^{\pi^M}(s';p^M,r)\right| \\
&\leqslant 2MHb_h(s,a) + \max_{s'\in\mathcal{S}} \left|V_{h+1}^{\widetilde{\pi}^M}(s';\widetilde{p}^M,\widehat{r}) - V_{h+1}^{\pi^M}(s';p^M,r)\right| \\
&= 2MHb_h(s,a) + X_h \\
&\leqslant 2MHb_h(s,a) + 2MH^3 \max_{(s',h')\in\mathcal{S}^{p,\pi^E}} b_{h'}(s',a^E) \\
&\leqslant 2MHb_h(s,a) + 4MH^3 \max_{(s',h')\in\mathcal{S}^{p,\pi^E}} b_{h'}(s',a^E).
\end{aligned}$$

Obviously, for $(s,a,h) \notin \mathcal{Z}^{p,\pi^b}$, we have:

$$\left|\widehat{r}_h(s,a) - r_h(s,a)\right| = 0.$$

Therefore, by Eq. 48, we can write:

$$\begin{aligned}
\mathcal{H}_d(\mathcal{R}^{\cap}_{p,\pi^E}, \widetilde{\mathcal{R}}^{\cap}) &\leqslant \sup_{r\in\mathcal{R}^{\cap}_{p,\pi^E}} \frac{1}{M} \sum_{h\in[\![H]\!]} \mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} \left|r_h(s,a) - \widehat{r}_h(s,a)\right| \\
&\leqslant \sum_{h\in[\![H]\!]} \mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} \left(2Hb_h(s,a) + 4H^3 \max_{(s',h')\in\mathcal{S}^{p,\pi^E}} b_{h'}(s',a^E)\right) \\
&= 2H \sum_{h\in[\![H]\!]} \mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} b_h(s,a) + 4H^3 \sum_{h\in[\![H]\!]} \mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} \max_{(s',h')\in\mathcal{S}^{p,\pi^E}} b_{h'}(s',a^E) \\
&= 2H \sum_{h\in[\![H]\!]} \mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} b_h(s,a) + 4H^3 \sum_{h\in[\![H]\!]} \max_{(s',h')\in\mathcal{S}^{p,\pi^E}} b_{h'}(s',a^E) \\
&= 2H \sum_{h\in[\![H]\!]} \mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} b_h(s,a) + 4H^4 \max_{(s',h')\in\mathcal{S}^{p,\pi^E}} b_{h'}(s',a^E).
\end{aligned}$$

This concludes the proof w.r.t. the subset. □

Now we can exploit Lemma F.13 to bound the error for the superset.

Lemma F.15 (Performance Decomposition Superset)**.** *Under good event $\mathcal{E}$, it holds that:*

$$\mathcal{H}_d(\mathcal{R}^{\cup}_{p,\pi^E}, \widetilde{\mathcal{R}}^{\cup}) \leqslant 2H \sum_{h\in[\![H]\!]} \mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} b_h(s,a) + 4H^4 \max_{(s,h)\in\mathcal{S}^{p,\pi^E}} b_h(s,a^E).$$

Proof. We can write:

$$\begin{aligned}
\mathcal{H}_d(\mathcal{R}^{\cup}_{p,\pi^E}, \widetilde{\mathcal{R}}^{\cup}) &:= \max\{\sup_{r\in\mathcal{R}^{\cup}_{p,\pi^E}} \inf_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cup}} d(r,\widetilde{r}),\ \sup_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cup}} \inf_{r\in\mathcal{R}^{\cup}_{p,\pi^E}} d(r,\widetilde{r})\} \\
&\overset{(1)}{=} \sup_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cup}} \inf_{r\in\mathcal{R}^{\cup}_{p,\pi^E}} d(r,\widetilde{r}) \\
&=: \sup_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cup}} \inf_{r\in\mathcal{R}^{\cup}_{p,\pi^E}} \frac{1}{M}\sum_{h\in[\![H]\!]} \left(\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} \left|r_h(s,a)-\widetilde{r}_h(s,a)\right| + \max_{(s,a)\notin\mathcal{Z}_h^{p,\pi^b}} \left|r_h(s,a)-\widetilde{r}_h(s,a)\right| \right) \\
&\overset{(2)}{\leqslant} \sup_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cup}} \frac{1}{M}\sum_{h\in[\![H]\!]} \left(\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} \left|r_h(s,a)-\widetilde{r}_h(s,a)\right| + \underbrace{\max_{(s,a)\notin\mathcal{Z}_h^{p,\pi^b}} \left|r_h(s,a)-\widetilde{r}_h(s,a)\right|}_{=0} \right) \\
&= \sup_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cup}} \frac{1}{M}\sum_{h\in[\![H]\!]} \mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} \left|r_h(s,a)-\widetilde{r}_h(s,a)\right|
\end{aligned} \tag{49}$$

where at (1) we have used that, under good event $\mathcal{E}$, $\mathcal{R}^{\cup}_{p,\pi^E} \subseteq \widetilde{\mathcal{R}}^{\cup}$, and at (2) we apply Lemma F.13, by denoting with r the chosen reward from $\mathcal{R}^{\cup}_{p,\pi^E}$.

In the following, it is useful to denote, for any $h\in[\![H]\!]$:

$$Y_h := \max_{s'\in\mathcal{S}} \left|V^{\widetilde{\pi}^m}_{h+1}(s';\widetilde{p}^m,\widehat{r}) - V^{\pi^m}_{h+1}(s';p^m,r)\right|.$$

Let us consider any $(s,h)\in\mathcal{S}^{p,\pi^E}$. The difference between the rewards of expert's action can be bounded by:

$$\begin{aligned}
\left|\widehat{r}_h(s,a^E)-r_h(s,a^E)\right| &\overset{(1)}{\leqslant} \Big| \mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)} V^{\pi^E}_{h+1}(s';p,r) - \mathop{\mathbb{E}}_{s'\sim\widetilde{p}_h^M(\cdot|s,a^E)} V^{\pi^E}_{h+1}(s';\widetilde{p}^M,\widehat{r})\Big| \\
&\quad + \left|\min_{s'\in\mathcal{S}} V^{\widetilde{\pi}^m}_{h+1}(s';\widetilde{p}^m,\widehat{r}) - \min_{s'\in\mathcal{S}} V^{\pi^m}_{h+1}(s';p^m,r)\right| \\
&\overset{(2)}{\leqslant} \Big| \mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)} V^{\pi^E}_{h+1}(s';p,r) - \mathop{\mathbb{E}}_{s'\sim\widetilde{p}_h^M(\cdot|s,a^E)} V^{\pi^E}_{h+1}(s';\widetilde{p}^M,\widehat{r}) \pm \mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)} V^{\pi^E}_{h+1}(s';\widetilde{p}^M,\widehat{r})\Big| \\
&\quad + \max_{s'\in\mathcal{S}} \left|V^{\widetilde{\pi}^m}_{h+1}(s';\widetilde{p}^m,\widehat{r}) - V^{\pi^m}_{h+1}(s';p^m,r)\right| \\
&\overset{(3)}{\leqslant} MH\left\|p_h(\cdot|s,a^E)-\widetilde{p}_h^M(\cdot|s,a^E)\right\|_1 + \mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)} \left|V^{\pi^E}_{h+1}(s';p,r) - V^{\pi^E}_{h+1}(s';\widetilde{p}^M,\widehat{r})\right| + Y_h \\
&\overset{(4)}{\leqslant} 2MHb_h(s,a^E) + \mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)} \left|Q^{\pi^E}_{h+1}(s',a^E;p,r) - Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^M,\widehat{r})\right| + Y_h \\
&\overset{(5)}{=} 2MHb_h(s,a^E) + \mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)} \left|\min_{s''\in\mathcal{S}} V^{\pi^m}_{h+2}(s'';p^m,r) - \min_{s''\in\mathcal{S}} V^{\widetilde{\pi}^m}_{h+2}(s'';\widetilde{p}^m,\widehat{r})\right| + Y_h \\
&\leqslant 2MHb_h(s,a^E) + \mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)} \left[Y_{h+1}\right] + Y_h \\
&\overset{(6)}{=} 2MHb_h(s,a^E) + Y_h + Y_{h+1},
\end{aligned}$$

where at (1) we use the definition of r in Lemma F.13 and triangle inequality, at (2) we use that, for any pair f,g of real-valued functions, it holds that $|\min_x f(x) - \min_x g(x)| \leqslant \max_x |f(x)-g(x)|$, at (3) we apply triangle inequality twice, we recognize the definition of X_h, we upper bound the value function by MH, and we recognize the definition of ℓ_1 norm, at (4) we first use triangle inequality $\|p_h(\cdot|s,a^E)-\widetilde{p}_h^M(\cdot|s,a^E)\|_1 \leqslant \|p_h(\cdot|s,a^E)-\widehat{p}_h(\cdot|s,a^E)\|_1 + \|\widetilde{p}_h^M(\cdot|s,a^E)-\widehat{p}_h(\cdot|s,a^E)\|_1$, then we use Pinsker's inequality, event $\mathcal{E}_3$ from Lemma F.2, and the definition of $b_h(s,a^E)$; at (5) we use the definition of $\widehat{r}$ in Lemma F.13 in the form of Eq. 47, by noticing that the support of $p_h(\cdot|s,a^E)$ is contained in $\mathcal{S}^{p,\pi^E}$, and at (6) we realize that Y_{h+1} depends only on h and not on s'.

In order to upper bound the term Y_h, we write:

$$Y_h := \max_{s'\in\mathcal{S}} \left|V^{\widetilde{\pi}^m}_{h+1}(s';\widetilde{p}^m,\widehat{r}) - V^{\pi^m}_{h+1}(s';p^m,r)\right|$$

$$
\begin{aligned}
&\overset{(1)}{=}\max\Bigg\{\max_{s'\notin\mathcal{S}^{p,\pi^E}_{h+1}}\Big|\max_{a'\in\mathcal{A}}Q^{\widetilde{\pi}^m}_{h+1}(s',a';\widetilde{p}^m,\widehat{r})-\max_{a'\in\mathcal{A}}Q^{\pi^m}_{h+1}(s',a';p^m,r)\Big|,\\
&\qquad\max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\Big|Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^m,\widehat{r})-Q^{\pi^E}_{h+1}(s',a^E;p,r)\Big|\Bigg\}\\
&\overset{(2)}{\leqslant}\max\Bigg\{\max_{s'\notin\mathcal{S}^{p,\pi^E}_{h+1}}\max_{a'\in\mathcal{A}}\Big|Q^{\widetilde{\pi}^m}_{h+1}(s',a';\widetilde{p}^m,\widehat{r})-Q^{\pi^m}_{h+1}(s',a';p^m,r)\Big|,\\
&\qquad\max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\Big|Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^m,\widehat{r})-Q^{\pi^E}_{h+1}(s',a^E;p,r)\pm Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^M,\widehat{r})\Big|\Bigg\}\\
&\overset{(3)}{=}\max\Bigg\{\max_{s'\notin\mathcal{S}^{p,\pi^E}_{h+1}}\max\Bigg\{\max_{a'\in\mathcal{A}:(s',a',h+1)\in\mathcal{Z}^{p,\pi^b}}\Big|\underbrace{Q^{\widetilde{\pi}^m}_{h+1}(s',a';\widetilde{p}^m,\widehat{r})-Q^{\pi^m}_{h+1}(s',a';p^m,r)}_{=0}\Big|,\\
&\qquad\max_{a'\in\mathcal{A}:(s',a',h+1)\notin\mathcal{Z}^{p,\pi^b}}\Big|Q^{\widetilde{\pi}^m}_{h+1}(s',a';\widetilde{p}^m,\widehat{r})-Q^{\pi^m}_{h+1}(s',a';p^m,r)\Big|\Bigg\},\\
&\qquad\max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\Big|Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^m,\widehat{r})-Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^M,\widehat{r})+\min_{s''\in\mathcal{S}}V^{\widetilde{\pi}^m}_{h+2}(s'';\widetilde{p}^m,\widehat{r})-\min_{s''\in\mathcal{S}}V^{\pi^m}_{h+2}(s';p^m,r)\Big|\Bigg\}\\
&\overset{(4)}{\leqslant}\max\Bigg\{\max_{s'\notin\mathcal{S}^{p,\pi^E}_{h+1}}\max_{a'\in\mathcal{A}:(s',a',h+1)\notin\mathcal{Z}^{p,\pi^b}}\Big|\underbrace{\widehat{r}_{h+1}(s',a')-r_{h+1}(s',a')}_{=0}+\min_{s''\in\mathcal{S}}V^{\widetilde{\pi}^m}_{h+2}(s'';\widetilde{p}^m,\widehat{r})-\min_{s''\in\mathcal{S}}V^{\pi^m}_{h+2}(s'';p^m,r)\Big|,\\
&\qquad\max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\Big|Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^M,\widehat{r})-Q^{\pi^E}_{h+1}(s',a^E;\widetilde{p}^m,\widehat{r})\Big|+Y_{h+1}\Bigg\}\\
&\overset{(5)}{\leqslant}\max\Bigg\{Y_{h+1},\max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\Big(\Big|\mathop{\mathbb{E}}_{s''\sim\widetilde{p}^M_{h+1}(\cdot|s',a^E)}V^{\pi^E}_{h+2}(s'';\widetilde{p}^M,\widehat{r})-\mathop{\mathbb{E}}_{s''\sim\widetilde{p}^m_{h+1}(\cdot|s',a^E)}V^{\pi^E}_{h+2}(s'';\widetilde{p}^m,\widehat{r})\Big|+Y_{h+1}\Big)\Bigg\}\\
&=Y_{h+1}+\max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\Big|\mathop{\mathbb{E}}_{s''\sim\widetilde{p}^M_{h+1}(\cdot|s',a^E)}V^{\pi^E}_{h+2}(s'';\widetilde{p}^M,\widehat{r})-\mathop{\mathbb{E}}_{s''\sim\widetilde{p}^m_{h+1}(\cdot|s',a^E)}V^{\pi^E}_{h+2}(s'';\widetilde{p}^m,\widehat{r})\pm\mathop{\mathbb{E}}_{s''\sim\widetilde{p}^m_{h+1}(\cdot|s',a^E)}V^{\pi^E}_{h+2}(s'';\widetilde{p}^M,\widehat{r})\Big|\\
&\overset{(6)}{\leqslant}Y_{h+1}+\max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\Big(MH\Big\|\widetilde{p}^M_{h+1}(\cdot|s',a^E)-\widetilde{p}^m_{h+1}(\cdot|s',a^E)\Big\|_1+\mathop{\mathbb{E}}_{s''\sim\widetilde{p}^m_{h+1}(\cdot|s',a^E)}\Big|V^{\pi^E}_{h+2}(s'';\widetilde{p}^M,\widehat{r})-V^{\pi^E}_{h+2}(s'';\widetilde{p}^m,\widehat{r})\Big|\Big)\\
&\overset{(7)}{\leqslant}Y_{h+1}+\max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\Big(2MHb_{h+1}(s',a^E)\\
&\qquad+\mathop{\mathbb{E}}_{s''\sim\widetilde{p}^m_{h+1}(\cdot|s',a^E)}\Big|\mathop{\mathbb{E}}_{s'''\sim\widetilde{p}^M_{h+2}(\cdot|s'',a^E)}V^{\pi^E}_{h+3}(s''';\widetilde{p}^M,\widehat{r})-\mathop{\mathbb{E}}_{s'''\sim\widetilde{p}^m_{h+2}(\cdot|s'',a^E)}V^{\pi^E}_{h+3}(s''';\widetilde{p}^m,\widehat{r})\Big|\Big)\\
&\overset{(8)}{\leqslant}Y_{h+1}+2MH\max_{s'\in\mathcal{S}^{p,\pi^E}_{h+1}}\sum_{h'\in[\![h+1,H-1]\!]}\mathop{\mathbb{E}}_{s''\sim\rho^{\widetilde{p}^m,\pi^E}_{h'}(\cdot|s_{h+1}=s')}b_{h'}(s'',a^E)\\
&\overset{(9)}{\leqslant}2MH\sum_{h'\in[\![h,H-1]\!]}\max_{s'\in\mathcal{S}^{p,\pi^E}_{h'+1}}\sum_{h''\in[\![h'+1,H-1]\!]}\mathop{\mathbb{E}}_{s''\sim\rho^{\widetilde{p}^m,\pi^E}_{h''}(\cdot|s_{h'+1}=s')}b_{h''}(s'',a^E)\\
&\leqslant 2MH^3\max_{(s',h')\in\mathcal{S}^{p,\pi^E}}b_{h'}(s',a^E),
\end{aligned}
$$

where at (1) we apply the Bellman's equation and the definition of $\widetilde{\pi}^m$ and π^m, at (2) we use that $|\min_x f(x)-\min_x g(x)|\leqslant\max_x|f(x)-g(x)|$, at (3) we use the definition of r in the form of Eq. 46 and Eq.47, at (4) we apply the Bellman's equation and the definition of r to recognize that $r_{h+1}(s',a')-\widehat{r}_{h+1}(s',a')=0$; moreover, we apply triangle inequality along with the usual bound $|\min_x f(x)-\min_x g(x)|\leqslant\max_x|f(x)-g(x)|$, and we recognize the definition of X_{h+1}. At (5) we proceed similarly as (4) and we use the Bellman's equation, and we observe that Y_{h+1} does not depend on s'; at (6) we upper bound the value function by HM and recognize the ℓ_1-norm, at (7) we use the concentration bound of event $\mathcal{E}_3$ in Lemma F.2 (both $\widetilde{p}^M$ and $\widetilde{p}^m$ lie at a "distance" of b from $\widehat{p}$). At (8) we have unfolded the recursion to bound the difference of value functions between transition models $\widetilde{p}^m$ and $\widetilde{p}^M$, at (9) we have unfolded the recursion on the Y terms.

Thanks to this expression, we can upper bound the difference of rewards in expert's action as:

$$\left|\widehat{r}_h(s,a^E)-r_h(s,a^E)\right|\leqslant 2MHb_h(s,a^E)+4MH^3\max_{(s',h')\in\mathcal{S}^{p,\pi^E}} b_{h'}(s',a^E).$$

With regards to visited $(s,a,h)\in\mathcal{Z}^{p,\pi^b}\setminus\mathcal{Z}^{p,\pi^E}$, we can write:

$$\begin{aligned}
\left|\widehat{r}_h(s,a)-r_h(s,a)\right| &= \Big|\mathop{\mathbb{E}}_{s'\sim\widetilde{p}_h^m(\cdot|s,a)} V_{h+1}^{\widetilde{\pi}^m}(s';\widetilde{p}^m,\widehat{r})-\mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a)} V_{h+1}^{\pi^m}(s';p^m,r)\Big| \\
&\leqslant 2MHb_h(s,a)+\mathop{\mathbb{E}}_{s'\sim\widetilde{p}_h^m(\cdot|s,a)}\left|V_{h+1}^{\widetilde{\pi}^m}(s';\widetilde{p}^m,\widehat{r})-V_{h+1}^{\pi^m}(s';p^m,r)\right| \\
&\leqslant 2MHb_h(s,a)+\max_{s'\in\mathcal{S}}\left|V_{h+1}^{\widetilde{\pi}^m}(s';\widetilde{p}^m,\widehat{r})-V_{h+1}^{\pi^m}(s';p^m,r)\right| \\
&= 2MHb_h(s,a)+Y_h \\
&\leqslant 2MHb_h(s,a)+2MH^3\max_{(s',h')\in\mathcal{S}^{p,\pi^E}} b_{h'}(s',a^E) \\
&\leqslant 2MHb_h(s,a)+4MH^3\max_{(s',h')\in\mathcal{S}^{p,\pi^E}} b_{h'}(s',a^E).
\end{aligned}$$

Obviously, for $(s,a,h)\notin\mathcal{Z}^{p,\pi^b}$, we have:

$$\left|\widehat{r}_h(s,a)-r_h(s,a)\right|=0.$$

Therefore, by Eq. 49, we can write:

$$\begin{aligned}
\mathcal{H}_d(\mathcal{R}^{\cup}_{p,\pi^E},\widetilde{\mathcal{R}}^{\cup}) &\leqslant \sup_{\widetilde{r}\in\widetilde{\mathcal{R}}^{\cup}}\frac{1}{M}\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}\left|r_h(s,a)-\widehat{r}_h(s,a)\right| \\
&\leqslant \sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}\left(2Hb_h(s,a)+4H^3\max_{(s',h')\in\mathcal{S}^{p,\pi^E}} b_{h'}(s',a^E)\right) \\
&= 2H\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} b_h(s,a)+4H^3\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)}\max_{(s',h')\in\mathcal{S}^{p,\pi^E}} b_{h'}(s',a^E) \\
&= 2H\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} b_h(s,a)+4H^3\sum_{h\in[\![H]\!]}\max_{(s',h')\in\mathcal{S}^{p,\pi^E}} b_{h'}(s',a^E) \\
&= 2H\sum_{h\in[\![H]\!]}\mathop{\mathbb{E}}_{(s,a)\sim\rho_h^{p,\pi^b}(\cdot,\cdot)} b_h(s,a)+4H^4\max_{(s',h')\in\mathcal{S}^{p,\pi^E}} b_{h'}(s',a^E).
\end{aligned}$$

This concludes the proof w.r.t. the superset. □

F.4.2. Lemmas for Theorem F.19

Lemma F.16 (Performance Decomposition Subset)**.** *Under good event $\mathcal{E}$, it holds that:*

$$\mathcal{H}_\infty(\mathcal{R}^{\cap}_{p,\pi^E},\widetilde{\mathcal{R}}^{\cap})\leqslant 2H^2\max_{(s,a,h)\in\mathcal{Z}^{p,\pi^b}} b_h(s,a)+4H^4\max_{(s,h)\in\mathcal{S}^{p,\pi^E}} b_h(s,a^E).$$

Proof Sketch. The proof is analogous to that of Lemma F.14. We can reuse the bounds for the difference between rewards proved in there and insert them into $\mathcal{H}_\infty$ to get the result. □

Lemma F.17 (Performance Decomposition Superset)**.** *Under good event $\mathcal{E}$, it holds that:*

$$\mathcal{H}_\infty(\mathcal{R}^{\cap}_{p,\pi^E},\widetilde{\mathcal{R}}^{\cap})\leqslant 2H^2\max_{(s,a,h)\in\mathcal{Z}^{p,\pi^b}} b_h(s,a)+4H^4\max_{(s,h)\in\mathcal{S}^{p,\pi^E}} b_h(s,a^E).$$

Proof Sketch. The proof is analogous to that of Lemma F.15. We can reuse the bounds for the difference between rewards proved in there and insert them into $\mathcal{H}_\infty$ to get the result. □

F.4.3. PROOFS OF THE MAIN THEOREMS

Thanks to Lemma F.14 and Lemma F.15, we can conclude the proof of the main theorem for d.

Theorem F.18. *Under the conditions of Theorem 6.1,* **PIRLO** *is (ϵ,δ)-PAC for d-IRL with a sample complexity at most:*

$$\tau^b \leqslant \widetilde{\mathcal{O}}\left(\frac{H^3 Z^{p,\pi^b}\ln\frac{1}{\delta}}{\epsilon^2}\left(\ln\frac{1}{\delta}+S^{p,\pi^b}_{\max}\right)\right.$$
$$\left.+\frac{H^8\ln\frac{1}{\delta}}{\rho^{\pi^b,\mathcal{Z}^{p,\pi^E}}_{\min}\epsilon^2}\left(\ln\frac{1}{\delta}+S^{p,\pi^b}_{\max}\right)+\frac{\ln\frac{1}{\delta}}{\ln\frac{1}{1-\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min}}}\right),$$

and τ^E is bounded as in Theorem 5.1. Furthermore, **PIRLO** *is inclusion monotonic.*

Proof Sketch. The proof for the subset and superset is completely analogous. Thanks to Lemma F.14 and Lemma F.15, we realize that we have to bound the sum of two terms, which are completely analogous to those in the proof of Theorem 6.1, with the only difference of H^4 instead of H^3. By proceeding similarly, we get the result. □

Thanks to Lemma F.16 and Lemma F.17, we can conclude the proof of the main theorem for d_∞.

Theorem F.19. *Under the conditions of Theorem 6.1,* **PIRLO** *is (ϵ,δ)-PAC for d_∞-IRL with a sample complexity at most:*

$$\tau^b \leqslant \widetilde{\mathcal{O}}\left(\frac{H^4\ln\frac{1}{\delta}}{\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min}\epsilon^2}\left(\ln\frac{1}{\delta}+S^{p,\pi^b}_{\max}\right)\right.$$
$$\left.+\frac{H^8\ln\frac{1}{\delta}}{\rho^{\pi^b,\mathcal{Z}^{p,\pi^E}}_{\min}\epsilon^2}\left(\ln\frac{1}{\delta}+S^{p,\pi^b}_{\max}\right)+\frac{\ln\frac{1}{\delta}}{\ln\frac{1}{1-\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min}}}\right),$$

and τ^E is bounded as in Theorem 5.1. Furthermore, **PIRLO** *is inclusion monotonic.*

Proof Sketch. The proof for the subset and superset is completely analogous. Thanks to Lemma F.16 and Lemma F.17, we realize that we have to bound the sum of two terms, which are completely analogous to those in the proof of Theorem 6.2, with the only difference of H^4 instead of H^3. By proceeding similarly, we get the result. □

F.5. A note on the superset without relaxation

In this section, we show that, if we use the superset definition $\widehat{\mathcal{R}}^{\cup}$ of Eq. 8, i.e., the definition without relaxation, then we are able to obtain the same performance decomposition result (see Lemma F.5) that we had for the case without pessimism, and, thus, we end up with the same sample complexity result, which is much smaller than those computed for the relaxations. Observe that we are not able to have an analogous result for the subset definition $\widehat{\mathcal{R}}^{\cap}$ of Eq. 8. Indeed, differently from the relaxations defined in Eq. 9, the subset and the superset definitions of Eq. 8 are not exactly simmetric. While $r\in\widehat{\mathcal{R}}^{\cup}$ entails, by definition, the existence of (at least) one transition model in $\mathcal{C}(\widehat{p},b)$ in which r induces an optimal policy $\pi^*\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}$, this is not true for $r'\in\widehat{\mathcal{R}}^{\cap}$. Indeed, potentially, there might exist a (worst)[18] transition model for every $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]$. Therefore, intuitively, in the reward choice lemma, we cannot make a choice of a single (worst) transition model, but we have to choose many of them. This fact "breaks" the recursion and it does not allow us to perform ℓ_1-norm bounds. Instead, since the superset is of different nature, we can. It should be remarked that a membership checker algorithm for superset $\widehat{\mathcal{R}}^{\cup}$ is inefficient to implement in practice because it requires to solve a bilinear optimization problem (see Appendix G).

We will denote by $\widehat{\mathcal{R}}^{\cup}$ the superset definition of Eq. 8.

Lemma F.20 (Reward Choice)**.** *Under good event $\mathcal{E}$, for any $\widehat{r}\in\widehat{\mathcal{R}}^{\cup}$, the reward r constructed as:*

$$\begin{cases} r_h(s,a)=\widehat{r}_h(s,a)+\sum_{s'\in\mathcal{S}}(\breve{p}_h(s'|s,a)-p_h(s'|s,a))V^*_{h+1}(s';\breve{p},\widehat{r}), & \forall(s,a,h)\in\mathcal{Z}^{p,\pi^b} \\ r_h(s,a)=\widehat{r}_h(s,a), & \forall(s,a,h)\notin\mathcal{Z}^{p,\pi^b} \end{cases},$$

[18] Worst because there is a universal quantifier $\forall$ over transition models in the definition of $\widehat{\mathcal{R}}^{\cap}$.

where $\breve{p}$ is some transition model in $\mathcal{C}(\widehat{p},b)$, belongs to $\mathcal{R}^{\cup}_{p,\pi^E}$.

Proof. By definition of $\widehat{\mathcal{R}}^{\cup}$, we have that $\widehat{r}\in\widehat{\mathcal{R}}^{\cup}$ if and only if there exists a transition model $\bar{p}\in\mathcal{C}(\widehat{p},b)$ such that $\widehat{r}\in\mathcal{R}^{\cup}_{\bar{p},\pi^E}$. Let us construct r by choosing $\breve{p}=\bar{p}$. To show that $r\in\mathcal{R}^{\cup}_{p,\pi^E}$, we are going to show that the transition model $\widetilde{p}$ defined as:

$$\begin{cases}\widetilde{p}_h(\cdot|s,a)=p_h(\cdot|s,a), & \forall(s,a,h)\in\mathcal{Z}^{p,\pi^b}\\ \widetilde{p}_h(\cdot|s,a)=\bar{p}_h(\cdot|s,a), & \forall(s,a,h)\notin\mathcal{Z}^{p,\pi^b}\end{cases},$$

belongs to $[p]_{\equiv_{\mathcal{Z}^{p,\pi^b}}}$ and is such that, for all $(s,h)\in\mathcal{S}^{p,\pi^E}$, for all $a\in\mathcal{A}\backslash\{a^E\}$:

$$Q^*_h(s,a;\widetilde{p},r)\leqslant Q^*_h(s,a^E;\widetilde{p},r).$$

Then, by Lemma E.1, we can conclude that $r\in\mathcal{R}^{\cup}_{p,\pi^E}$.

Trivially, notice that $\widetilde{p}\equiv_{\mathcal{Z}^{p,\pi^b}}p$. We proceed by induction to show that, for all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]$, the following identity holds:

$$Q^*_h(s,a;\bar{p},\widehat{r})=Q^*_h(s,a;\widetilde{p},r).$$

Then, since $\widehat{r}\in\widehat{\mathcal{R}}^{\cup}$, for all $(s,h)\in\mathcal{S}^{p,\pi^E}$ and for all $a\in\mathcal{A}\backslash\{a^E\}$, the inequality $Q^*_h(s,a;\bar{p},\widehat{r})\leqslant Q^*_h(s,a^E;\bar{p},\widehat{r})$ (Lemma E.1) entails $Q^*_h(s,a;\widetilde{p},r)\leqslant Q^*_h(s,a^E;\widetilde{p},r)$, and the thesis follows.

As case base, consider stage H. For any $(s,a)\in\mathcal{S}\times\mathcal{A}$, thanks to the definition of r, we can write:

$$\begin{aligned}Q^*_H(s,a;\widetilde{p},r)&=r_H(s,a)\\&=\widehat{r}_H(s,a)\\&=Q^*_H(s,a;\bar{p},\widehat{r}).\end{aligned}$$

Now, make the inductive hypothesis that for all $(s',a')\in\mathcal{S}\times\mathcal{A}$, it holds that $Q^*_{h+1}(s',a';\widetilde{p},r)=Q^*_{h+1}(s',a';\bar{p},\widehat{r})$, and consider stage h. For any $(s,a)\in\mathcal{Z}^{p,\pi^b}_h$, we can write:

$$\begin{aligned}Q^*_h(s,a;\widetilde{p},r)&\overset{(1)}{=}r_h(s,a)+\sum_{s'\in\mathcal{S}}\widetilde{p}_h(s'|s,a)\max_{a'\in\mathcal{A}}Q^*_{h+1}(s',a';\widetilde{p},r)\\&\overset{(2)}{=}r_h(s,a)+\sum_{s'\in\mathcal{S}}\widetilde{p}_h(s'|s,a)\max_{a'\in\mathcal{A}}Q^*_{h+1}(s',a';\bar{p},\widehat{r})\\&\overset{(3)}{=}\widehat{r}_h(s,a)+\sum_{s'\in\mathcal{S}}(\breve{p}_h(s'|s,a)-p_h(s'|s,a))\max_{a'\in\mathcal{A}}Q^*_{h+1}(s',a';\breve{p},\widehat{r})+\sum_{s'\in\mathcal{S}}\widetilde{p}_h(s'|s,a)\max_{a'\in\mathcal{A}}Q^*_{h+1}(s',a';\bar{p},\widehat{r})\\&\overset{(4)}{=}\widehat{r}_h(s,a)+\sum_{s'\in\mathcal{S}}\bar{p}_h(s'|s,a)\max_{a'\in\mathcal{A}}Q^*_{h+1}(s',a';\bar{p},\widehat{r})\\&\overset{(5)}{=}Q^*_h(s,a;\bar{p},\widehat{r}),\end{aligned}$$

where at (1) we have applied the Bellman's optimality equation, at (2) we have used the inductive hypothesis, at (3) we have inserted the definition of $r_h(s,a)$ along with the fact that $(s,a,h)\in\mathcal{Z}^{p,\pi^b}$, at (4) we have noticed that $\breve{p}=\bar{p}$ (by choice) and that $\widetilde{p}_h(\cdot|s,a)=p_h(\cdot|s,a)$ (by definition); finally, at (5), we have applied again the Bellman's optimality equation.

On the other side, for any $(s,a)\notin\mathcal{Z}^{p,\pi^b}_h$, we can write:

$$\begin{aligned}Q^*_h(s,a;\widetilde{p},r)&=r_h(s,a)+\sum_{s'\in\mathcal{S}}\widetilde{p}_h(s'|s,a)\max_{a'\in\mathcal{A}}Q^*_{h+1}(s',a';\widetilde{p},r)\\&=r_h(s,a)+\sum_{s'\in\mathcal{S}}\widetilde{p}_h(s'|s,a)\max_{a'\in\mathcal{A}}Q^*_{h+1}(s',a';\bar{p},\widehat{r})\\&\overset{(1)}{=}\widehat{r}_h(s,a)+\sum_{s'\in\mathcal{S}}\bar{p}_h(s'|s,a)\max_{a'\in\mathcal{A}}Q^*_{h+1}(s',a';\bar{p},\widehat{r})\end{aligned}$$

$$= Q_h^*(s,a;\bar{p},\widehat{r}),$$

where at (1) we have used both the definition of $r_h(s,a)$ and that $\widetilde{p}_h(\cdot|s,a) = \bar{p}_h(\cdot|s,a)$, for $(s,a,h) \notin \mathcal{Z}^{p,\pi^b}$.

This concludes the proof. □

Thanks to the reward choice lemma just presented, we obtain the following sample complexity result.

Theorem F.21. *Let $\mathcal{M}$ be an MDP without reward and let π^E be the expert's policy. Let $\mathcal{D}^E$ and $\mathcal{D}^b$ be two datasets of τ^E and τ^b trajectories collected with policies π^E and π^b in $\mathcal{M}$, respectively. Under Assumption 2.1, any algorithm $\mathfrak{A}$ that outputs $\widehat{\mathcal{R}}^{\cup}$ (defined as in Eq. 8) is such that, for any $\epsilon, \delta \in (0,1)$:*

$$\mathop{\mathbb{P}}_{(p,\pi^E,\pi^b)}\left(\left\{\mathcal{H}_c(\mathcal{R}^{\cup}_{p,\pi^E}, \widehat{\mathcal{R}}^{\cup}) \leqslant \epsilon\right\} \wedge \left\{\mathcal{R}^{\cup}_{p,\pi^E} \subseteq \widehat{\mathcal{R}}^{\cup}\right\}\right) \geqslant 1-\delta,$$

with a sample complexity at most:

$$\tau^b \leqslant \widetilde{\mathcal{O}}\left(\frac{H^3 Z^{p,\pi^b} \ln\frac{1}{\delta}}{\epsilon^2}\left(\ln\frac{1}{\delta} + S^{p,\pi^b}_{\max}\right) + \frac{\ln\frac{1}{\delta}}{\ln\frac{1}{1-\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min}}}\right),$$

$$\tau^E \leqslant \widetilde{\mathcal{O}}\left(\frac{\ln\frac{1}{\delta}}{\ln\frac{1}{1-\rho^{\pi^E,\mathcal{Z}^{p,\pi^E}}_{\min}}}\right).$$

if $c = d$, and a sample complexity at most:

$$\tau^b \leqslant \widetilde{\mathcal{O}}\left(\frac{H^4 \ln\frac{1}{\delta}}{\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min}\epsilon^2}\left(\ln\frac{1}{\delta} + S^{p,\pi^b}_{\max}\right) + \frac{\ln\frac{1}{\delta}}{\ln\frac{1}{1-\rho^{\pi^b,\mathcal{Z}^{p,\pi^b}}_{\min}}}\right),$$

$$\tau^E \leqslant \widetilde{\mathcal{O}}\left(\frac{\ln\frac{1}{\delta}}{\ln\frac{1}{1-\rho^{\pi^E,\mathcal{Z}^{p,\pi^E}}_{\min}}}\right).$$

if $c = d_\infty$.

Proof Sketch. Observe that, thanks to Lemma F.20, we are able to obtain a performance decomposition lemma analogous to Lemma F.5 for distance d (or analogous for distance d_∞). Next, following the steps in the proof of Theorem 5.1 (Theorem 5.2), we obtain the result. □

G. Implementation

In this appendix, we provide some comments on the implementation of membership checker algorithms for **IRLO**, **PIRLO**, and some comments about the subset and superset defined in Eq. 8. In Section G.1, we present the pseudocode of the membership checker algorithms. In Section G.2, we provide more details on the definitions of the relaxations $\widetilde{\mathcal{R}}^{\cap}$ and $\widetilde{\mathcal{R}}^{\cup}$. In Section G.3 we show that the implementation of a membership checker algorithm for the subset and superset defined in Eq. 8 is inefficient. Finally, in Section G.4, we give the intuition that a straightforward relaxation of the representation of the sets provided by Lemma E.1 is worse than that obtained by relaxing the representation in Theorem 3.1.

G.1. Algorithm

The pseudocode of the membership checker algorithms for **IRLO** and **PIRLO** is provided in Algorithm 2.

Algorithm 2 Membership checker for **IRLO** and **PIRLO**.

Input : Datasets $\mathcal{D}^E = \{\langle s_h^{E,i}, a_h^{E,i}\rangle_h\}_i$, $\mathcal{D}^b = \{\langle s_h^{b,i}, a_h^{b,i}\rangle_h\}_i$, candidate reward function $r \in \mathfrak{R}$
Output : True if $r \in (\widehat{\mathcal{R}}^{\cup}, \widehat{\mathcal{R}}^{\cap})$
Run lines 1-11 of Algorithm 1
Define $\mathcal{C}$ as in Eq. (6) for **IRLO** and as in Eq. (7) for **PIRLO**
Extended value iteration:
$\mathcal{A}_h(s) \leftarrow$ **if** $(s,h) \in \widehat{\mathcal{S}}^{p,\pi^E}$ **then** $\{\widehat{\pi}_h^E(s)\}$ **else** $\mathcal{A}, \forall (s,h) \in \mathcal{S} \times [\![H]\!]$
$Q_H^+(s,a), Q_H^-(s,a) \leftarrow r_H(s,a)\ \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]$
for $h = H-1$ **to** 1 **do**
  for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
    $Q_h^+(s,a) \leftarrow r_h(s,a) + \max_{p' \in \mathcal{C}} \sum_{s' \in \mathcal{S}} p'_h(s'|s,a) \max_{a' \in \mathcal{A}_{h+1}(s)} Q_{h+1}^+(s',a')$
    $Q_h^-(s,a) \leftarrow r_h(s,a) + \min_{p' \in \mathcal{C}} \sum_{s' \in \mathcal{S}} p'_h(s'|s,a) \max_{a' \in \mathcal{A}_{h+1}(s)} Q_{h+1}^-(s',a')$
  end
end
Membership test:
$\text{in}^{\cup} \leftarrow$ True, $\text{in}^{\cap} \leftarrow$ True
for $(s,h) \in \widehat{\mathcal{S}}^{p,\pi^E}$ **do**
  for $a \in \mathcal{A} \setminus \{\widehat{\pi}^E(s)\}$ **do**
    IRLO
    if $Q_h^+(s, \widehat{\pi}_h^E(s)) < Q_h^+(s,a)$ **then**
      $\text{in}^{\cup} \leftarrow$ False
    end
    if $Q_h^-(s, \widehat{\pi}_h^E(s)) < Q_h^-(s,a)$ **then**
      $\text{in}^{\cap} \leftarrow$ False
    end
    PIRLO
    if $Q_h^+(s, \widehat{\pi}_h^E(s)) < Q_h^-(s,a)$ **then**
      $\text{in}^{\cup} \leftarrow$ False
    else if $Q_h^-(s, \widehat{\pi}_h^E(s)) < Q_h^+(s,a)$ **then**
      $\text{in}^{\cap} \leftarrow$ False
    end
  end
end
return ($\text{in}^{\cup}$, $\text{in}^{\cap}$)

The idea is to find the worst (resp. best) transition model for the subset (resp. superset) among all those feasible. In practice, what we do is to exploit the representation provided in Eq. 22 for **IRLO** and the representation provided in Eq. 27 for **PIRLO**.

Observe that, inside the support $(s,h) \in \widehat{\mathcal{S}}^{p,\pi^E}$, we use the estimated expert's action $\widehat{\pi}_h^E(s)$, while outside the support we always play the action that maximizes the Q-function (both Q^+ and Q^-). Concerning the transition model, notice that, for Q^+, we consider the $p' \in \mathcal{C}$ that maximizes the expected Q^+, while for Q^- we consider the $p' \in \mathcal{C}$ that minimizes the expected Q^-. Observe also that, for **IRLO**, because of the definition of $\mathcal{C}$, inside the support $\widehat{\mathcal{Z}}^{p,\pi^b}$ we use $p' = \widehat{p}$, and outside the support we consider $p' = \arg\max_{s' \in \mathcal{S}}$ for Q^+ and $p' = \arg\min_{s' \in \mathcal{S}}$ for Q^-.

Finally, we check the Bellman optimality conditions to assess the membership of the candidate reward r in the estimated sets $\widehat{\mathcal{R}}^{\cup}$ (boolean variable $\text{in}^{\cup}$) and $\widehat{\mathcal{R}}^{\cap}$ (boolean variable $\text{in}^{\cap}$) (line 2).

G.2. A better understanding of the relaxations

To get a better understanding of why $\widetilde{\mathcal{R}}^{\cap} \subseteq \mathcal{R}^{\cap}_{p,\pi^E}$ and $\mathcal{R}^{\cup}_{p,\pi^E} \subseteq \widetilde{\mathcal{R}}^{\cup}$, under the hypothesis (good event) that the true transition model $p \in \mathcal{C}(\widehat{p}, b)$ and that $\widehat{\pi}^E = \pi^E$ in all $\mathcal{S}^{p,\pi^E}$, observe that, for the subset:

$$\mathcal{R}^{\cap}_{p,\pi^E} \supseteq \bigcap_{p' \in \mathcal{C}(\widehat{p},b)} \mathcal{R}^{\cap}_{p',\pi^E}$$
$$= \{r \in \mathfrak{R} \,|\, \forall p' \in \mathcal{C}(\widehat{p},b), \forall \overline{\pi} \in [\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}, \forall (s,h) \in \mathcal{S}^{p,\pi^E}, \forall a \in \mathcal{A}\colon Q_h^{\pi^E}(s, \pi_h^E(s); p', r) \geqslant Q_h^{\overline{\pi}}(s,a;p',r)\}$$

$$\overset{(1)}{\supseteq}\{r\in\mathfrak{R}\,|\,\forall\bar{\pi}\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}},\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\colon \min_{p'\in\mathcal{C}(\widehat{p},b)} Q_h^{\pi^E}(s,\pi_h^E(s);p',r)\geqslant \max_{p'\in\mathcal{C}(\widehat{p},b)} Q_h^{\bar{\pi}}(s,a;p',r)\}$$

$$=\{r\in\mathfrak{R}\,|\,\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\colon \min_{\bar{\pi}\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}}\Big(\min_{p'\in\mathcal{C}(\widehat{p},b)} Q_h^{\pi^E}(s,\pi_h^E(s);p',r)-\max_{p'\in\mathcal{C}(\widehat{p},b)} Q_h^{\bar{\pi}}(s,a;p',r)\Big)\geqslant 0\}$$

$$\overset{(2)}{=}\{r\in\mathfrak{R}\,|\,\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\colon \min_{p'\in\mathcal{C}(\widehat{p},b)} Q_h^{\pi^E}(s,\pi_h^E(s);p',r)\geqslant \max_{\bar{\pi}\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}}\max_{p'\in\mathcal{C}(\widehat{p},b)} Q_h^{\bar{\pi}}(s,a;p',r)\}$$

$$=:\widetilde{\mathcal{R}}^{\cap},$$

where at (1) we have exchanged the order of the quantifiers, and we can do so because they all are of the same type, and then we have observed that $\min_x(f(x)-g(x))\geqslant\min_x f(x)-\max_x g(x)$, and at (2) we recognize that the first term does not depend on $\bar{\pi}$. W.r.t. the superset, in an analogous manner, observe that:

$$\mathcal{R}^{\cup}_{p,\pi^E}\subseteq\bigcup_{p'\in\mathcal{C}(\widehat{p},b)}\mathcal{R}^{\cup}_{p',\pi^E}$$

$$=\{r\in\mathfrak{R}\,|\,\exists p'\in\mathcal{C}(\widehat{p},b),\forall\bar{\pi}\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}},\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\colon Q_h^{\pi^E}(s,\pi_h^E(s);p',r)\geqslant Q_h^{\bar{\pi}}(s,a;p',r)\}$$

$$\overset{(1)}{\subseteq}\{r\in\mathfrak{R}\,|\,\forall\bar{\pi}\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}},\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\colon \max_{p'\in\mathcal{C}(\widehat{p},b)} Q_h^{\pi^E}(s,\pi_h^E(s);p',r)\geqslant \min_{p'\in\mathcal{C}(\widehat{p},b)} Q_h^{\bar{\pi}}(s,a;p',r)\}$$

$$=\{r\in\mathfrak{R}\,|\,\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\colon \min_{\bar{\pi}\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}}\Big(\max_{p'\in\mathcal{C}(\widehat{p},b)} Q_h^{\pi^E}(s,\pi_h^E(s);p',r)-\min_{p'\in\mathcal{C}(\widehat{p},b)} Q_h^{\bar{\pi}}(s,a;p',r)\Big)\geqslant 0\}$$

$$\overset{(2)}{=}\{r\in\mathfrak{R}\,|\,\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}\colon \max_{p'\in\mathcal{C}(\widehat{p},b)} Q_h^{\pi^E}(s,\pi_h^E(s);p',r)\geqslant \max_{\bar{\pi}\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}}\min_{p'\in\mathcal{C}(\widehat{p},b)} Q_h^{\bar{\pi}}(s,a;p',r)\}$$

$$=:\widetilde{\mathcal{R}}^{\cup},$$

where at (1) we have relaxed and at (2) we recognize that the first term does not depend on $\bar{\pi}$.

G.3. Testing the membership without relaxations

We show that the problem of testing the membership to the subset and superset defined as in Eq. 8 is equivalent to solving a bilinear optimization problem, which is in general hard. We will denote the expert's action by a^E and we use the representation provided by Lemma E.1.

Let us begin with the subset $\widehat{\mathcal{R}}^{\cap}$. A given reward r belongs to $\widehat{\mathcal{R}}^{\cap}$ if and only if:

$$\forall p'\in\mathcal{C}(\widehat{p},b),\forall(s,h)\in\widehat{\mathcal{S}}^{p,\pi^E},\forall a\in\mathcal{A}\backslash\{a^E\}\colon Q_h^*(s,a;p',r)\leqslant Q_h^*(s,a^E;p',r),$$

where $\mathcal{C}(\widehat{p},b)$ is defined in Eq. 7. By applying the Bellman optimality equation, changing the order of the quantifiers, and considering the worst possible transition model, we obtain:

$$\forall(s,h)\in\widehat{\mathcal{S}}^{p,\pi^E},\forall a\in\mathcal{A}\backslash\{a^E\}\colon$$
$$r_h(s,a)\leqslant r_h(s,a^E)+\min_{p'\in\mathcal{C}(\widehat{p},b)}\Big(\sum_{s'\in\mathcal{S}}p'_h(s'|s,a^E)V^*_{h+1}(s';p',r)-\sum_{s'\in\mathcal{S}}p'_h(s'|s,a)V^*_{h+1}(s';p',r)\Big).$$

It should be remarked that, differently from **IRLO** and **PIRLO**, to check whether a given reward r belongs to $\widehat{\mathcal{R}}^{\cap}$, we cannot optimize the value function, but we have to optimize the advantage function. Therefore, for all $(\bar{s},\bar{h})\in\widehat{\mathcal{S}}^{p,\pi^E}$, and for all $\bar{a}\in\mathcal{A}\backslash\{a^E\}$, we have to solve the optimization problem:

$$\min_{p'}\sum_{s'\in\mathcal{S}}\Big(p'_{\bar{h}}(s'|\bar{s},a^E)-p'_{\bar{h}}(s'|\bar{s},\bar{a})\Big)V^*_{\bar{h}+1}(s';p',r)$$
$$\text{s.t. }\|p'_h(\cdot|s,a)-\widehat{p}_h(\cdot|s,a)\|_1\leqslant b_h(s,a)\qquad\forall(s,a,h)\in\widehat{\mathcal{Z}}^{p,\pi^b}$$
$$p'_h(\cdot|s,a)\in\Delta^{\mathcal{S}}\qquad\forall(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]$$
$$p'_h(s'|s,a)=0\qquad\forall(s,a,h)\in\widehat{\mathcal{Z}}^{p,\pi^E}\wedge s'\notin\big(\widehat{\mathcal{S}}^{p,\pi^E}_{h+1}\cup\operatorname{supp}\widehat{p}_h(\cdot|s,a^E)\big).$$

Observe that, while the set of constraints define a convex set, the objective involves the product of optimization variables. We can introduce H variables $\{V_s\}_{s\in[\![S]\!]}$ to replace $V^*_{h+1}(s';p',r)$ by adding suitable constraints that keep into account also the presence of the maximum operator inside V^* (because of the Bellman's optimality equation). In this way, due to the product between variables p' and V, we can conclude that this is a bilinear optimization problem, which is in general difficult to solve.

W.r.t. the superset, we have that a reward r belongs to $\widehat{\mathcal{R}}^{\cup}$ if and only if:

$$\exists p'\in\mathcal{C}(\widehat{p},b),\forall(s,h)\in\widehat{\mathcal{S}}^{p,\pi^E},\forall a\in\mathcal{A}\backslash\{a^E\}: Q^*_h(s,a;p',r)\leqslant Q^*_h(s,a^E;p',r).$$

This time, we cannot bring the transition model inside because we have different quantifiers. We can formulate the problem as a feasibility problem by adding constraints because of the presence of $\forall(s,h)\in\widehat{\mathcal{S}}^{p,\pi^E},\forall a\in\mathcal{A}\backslash\{a^E\}$. In practice, the presence of the product between "optimization" variables is now in the constraints, so the problem is again a bilinear problem.

G.4. Relaxing the representation provided by Lemma E.1

We have seen that we can represent the feasible set by using Theorem 3.1 or Lemma E.1. While the two representations are equivalent, observe that a straightforward relaxation of the constraints present in Lemma E.1 provides a different relaxation of the subset and superset w.r.t. $\widetilde{\mathcal{R}}^{\cap}$ and $\widetilde{\mathcal{R}}^{\cup}$ (which are obtained by relaxing the representation in Theorem 3.1). Indeed, by relaxing the representation with the Q^* (Lemma E.1), we would obtain constraints of the form (ex. subset):

$$\begin{aligned}&\min_{p'\in\mathcal{C}(\widehat{p},b)} Q^*_h(s,\pi^E_h(s);p',r)\geqslant \max_{p'\in\mathcal{C}(\widehat{p},b)} Q^*_h(s,a;p',r)\\ \iff\quad &\min_{p'\in\mathcal{C}(\widehat{p},b)}\max_{\pi\in\Pi} Q^\pi_h(s,\pi^E_h(s);p',r)\geqslant \max_{p'\in\mathcal{C}(\widehat{p},b)}\max_{\pi\in\Pi} Q^\pi_h(s,a;p',r).\end{aligned}\tag{50}$$

Clearly, this is different from $\widetilde{\mathcal{R}}^{\cap}$, whose constraints can be written as:

$$\min_{p'\in\mathcal{C}(\widehat{p},b)} Q^{\pi^E}_h(s,\pi^E_h(s);p',r)\geqslant \max_{p'\in\mathcal{C}(\widehat{p},b)}\max_{\pi\in[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}} Q^\pi_h(s,a;p',r).$$

Indeed, $\widetilde{\mathcal{R}}^{\cap}$ puts the additional constraint that the Q^* is achieved by a policy in $[\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}$, which is not present in Eq. 50.

An analogous reasoning can be carried out also for the superset.

H. Proofs of Section 8

In this section, we provide the missing proofs of Section 8.

Proposition 8.1. *Let $\mathcal{M}$ be the usual MDP without reward with $A\geqslant 2$ and let π^E be the deterministic expert's policy. Let $\mathcal{D}^E$ be a dataset of trajectories collected by following π^E in $\mathcal{M}$. Then, for any reward in $r\in\mathcal{R}^{\cap}_{p,\pi^E}$ it holds that:*

$$\forall(s,h)\in\mathcal{S}^{p,\pi^E},\forall a\in\mathcal{A}:\quad r_h(s,\pi^E_h(s))\geqslant r_h(s,a).\tag{9}$$

Proof. Let r be an arbitrary reward function of $\mathcal{R}^{\cap}_{p,\pi^E}$. Consider a certain $(s,h)\in\mathcal{S}^{p,\pi^E}$, with expert's action $\pi^E_h(s)=a^E$, and let $a\in\mathcal{A}$ be a non-expert's action. By Lemma E.1, we know that, for any $p'\in[p]_{\equiv_{\mathcal{Z}^{p,\pi^E}}}$, it must hold:

$$\begin{aligned}r_h(s,a)&\leqslant r_h(s,a^E)+\mathop{\mathbb{E}}_{s'\sim p'_h(\cdot|s,a^E)} V^*_{h+1}(s';p',r)-\mathop{\mathbb{E}}_{s'\sim p'_h(\cdot|s,a)} V^*_{h+1}(s';p',r)\\&=r_h(s,a^E)+\mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)} V^*_{h+1}(s';p,r)-\mathop{\mathbb{E}}_{s'\sim p'_h(\cdot|s,a)} V^*_{h+1}(s';p',r),\end{aligned}$$

where we have used the definition of $[p]_{\equiv_{\mathcal{Z}^{p,\pi^E}}}$. Since $(s,a,h)\notin\mathcal{Z}^{p,\pi^E}$, then the constraint must hold $\forall p'_h(\cdot|s,a)\in\Delta^{\mathcal{S}}$. In particular, it must hold for the transition model such that:

$$r_h(s,a)\leqslant r_h(s,a^E)+\underbrace{\mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a^E)} V^*_{h+1}(s';p,r)-\max_{s'\in\mathcal{S}} V^*_{h+1}(s';p',r)}_{\leqslant 0},$$

from which the thesis follows. □

Proposition 8.2. *Under the conditions of Proposition 8.1, assume that $p_h(\cdot|s,a)$ is known, where $a \in \mathcal{A}$ is a non-expert's action in $(s,h) \in \mathcal{S}^{p,\pi^E}$. Then, if $p_h(\cdot|s,a) \neq p_h(\cdot|s,\pi_h^E(s))$, there exists a reward $r \in \mathcal{R}^{\cap}_{p,\pi^E}$ such that:*

$$r_h(s,\pi_h^E(s)) < r_h(s,a).$$

Proof. Let $a^E := \pi_h^E(s)$. Similarly to the proof of Proposition 8.1, we can write: $p_h(\cdot|s,a)$:

$$\begin{aligned} r_h(s,a) &\leqslant r_h(s,a^E) + \mathop{\mathbb{E}}_{s' \sim p'_h(\cdot|s,a^E)} V^*_{h+1}(s';p',r) - \mathop{\mathbb{E}}_{s' \sim p'_h(\cdot|s,a)} V^*_{h+1}(s';p',r) \\ &= r_h(s,a^E) + \mathop{\mathbb{E}}_{s' \sim p_h(\cdot|s,a^E)} V^*_{h+1}(s';p,r) - \mathop{\mathbb{E}}_{s' \sim p_h(\cdot|s,a)} V^*_{h+1}(s';p',r), \end{aligned}$$

where we have used that we have access to samples about $p_h(\cdot|s,a)$. By hypothesis, $p_h(\cdot|s,a) \neq p_h(\cdot|s,a^E)$, therefore, by taking r such that $\mathbb{E}_{s' \sim p_h(\cdot|s,a^E)} V^*_{h+1}(s';p,r) > \mathbb{E}_{s' \sim p_h(\cdot|s,a)} V^*_{h+1}(s';p',r)$, we can obtain a reward r in $\mathcal{R}^{\cap}_{p,\pi^E}$ such that $r_h(s,\pi_h^E(s)) < r_h(s,a)$.

□

I. A relaxed triangle inequality

In this section, we show that both our notions of distance d, d_∞, defined in Section 4, are semimetrics, and that they satisfy a ρ-relaxed triangle inequality (see Fagin & Stockmeyer, 1998) with finite $\rho > 1$ for any pair of rewards $r, r' \in \mathfrak{R}$. Furthermore, we show that the Hausdorff distance $\mathcal{H}$, when applied to the sets of rewards considered in this work, inherits the relaxed triangle inequality property. It should be remarked that we need the ρ-relaxed triangle inequality property with finite ρ just for the *learnability* proofs of Appendix C. Moreover, notice that we do not care about a tight value of ρ, but only that it is finite. Instead, if we wanted to compute a minimax lower bound, then we would need a tight value of ρ in order to obtain a tight lower bound.

I.1. d and d_∞ satisfy a relaxed triangle inequality

In the following, for the sake of simplicity, we denote reward functions by vectors $x, y, z, \ldots \in \mathbb{R}^k$. Moreover, for any pair $x, y \in \mathbb{R}^k$, we will consider distance d for some distribution $q \in \Delta^{[\![k]\!]}$ as:

$$d(x,y) = \frac{\sum_{i \in [\![k]\!]} q_i |x_i - y_i|}{\max\{\|x\|_\infty, \|y\|_\infty\}},$$

and distance d_∞ as:

$$d_\infty(x,y) = \frac{\|x-y\|_\infty}{\max\{\|x\|_\infty, \|y\|_\infty\}}.$$

First of all, let us see that neither d nor d_∞ are metrics:

Proposition I.1. *Both the functions d and d_∞ do not satisfy the triangle inequality.*

Proof. To show that the triangle inequality property is not satisfied, we simply provide some counterexamples. For the sake of simplicity, let $k = 2$.

W.r.t. distance d, let the vectors $x, y, z \in \mathbb{R}^2$ be defined as:

$$\begin{cases} x = [1,0]^\intercal \\ y = [-1,-1]^\intercal \\ z = [-2,-1]^\intercal \end{cases},$$

and observe that, for any $q \in \Delta^{[\![2]\!]}$ such that $q_2 > 0$:

$$d(x,y) = 2q_1 + q_2 \overset{?}{\leqslant} d(x,z) + d(y,z) = 3/2q_1 + q_2/2 + q_1/2 = 2q_1 + q_2/2$$

$$\iff q_2 \leqslant 0.$$

If $q_2 = 0$, we can take the second component of z to be arbitrary large so that the inequality is not verified.

As far as d_∞ is concerned, let $x, y, z \in \mathbb{R}^2$ be the vectors defined as:

$$\begin{cases} x = [0,2]^\intercal \\ y = [2,2]^\intercal \\ z = [1,3]^\intercal \end{cases}.$$

We have:

$$\begin{aligned} d_\infty(x,y) = \frac{2}{2} = 1 \overset{?}{\leqslant} d_\infty(x,z) + d_\infty(y,z) = 1/3 + 1/3 = 2/3 \\ \iff 1 \leqslant 2/3, \end{aligned}$$

which is clearly false.

Notice that it is possible to generate counterexamples for higher dimensions $k > 2$ by using a simple script of code. □

Having verified that distances d, d_∞ do not satisfy the triangle inequality, we can conclude that they are *semimetrics*, as it is easy to check the other three properties of positivity, simmetry, and that the distance between two points is zero if and only if the two points coincide. We are interested in verifying whether they satisfy a relaxed form of triangle inequality (Fagin & Stockmeyer, 1998). Specifically, for any finite $\rho \in \mathbb{R}$ with $\rho > 1$, we say that a function d satisfies the ρ-relaxed triangle inequality if, for any $x, y, z \in \mathbb{R}^k$:

$$d(x,y) \leqslant \rho\big(d(x,z) + d(y,z)\big).$$

We aim to show that both d and d_∞ satisfy the ρ-relaxed triangle inequality for some ρ. Let us begin with a useful lemma.

Lemma I.2. *Let $d_2 : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ be the function that, for any pair $x, y \in \mathbb{R}^k$, it returns:*

$$d_2(x,y) := \frac{\|x-y\|_2}{\max\{\|x\|_2, \|y\|_2\}}.$$

Then, d_2 is a metric.

Proof. It is easy to observe that $d_2(x,y) = 0$ if and only if $x = y$. Moreover, notice that $d_2(x,y) \geqslant 0$ for all $x, y \in \mathbb{R}^k$, and also that $d_2(x,y) = d_2(y,x)$.

It remains to prove that d_2 satisfies the triangle inequality property, i.e., for any $x, y, z \in \mathbb{R}^k$, it satisfies:

$$d_2(x,y) \leqslant d_2(x,z) + d_2(y,z).$$

We distinguish two cases, one in which $\max\{\|x\|_2, \|y\|_2, \|z\|_2\} \neq \|z\|_2$ and the other in which $\max\{\|x\|_2, \|y\|_2, \|z\|_2\} = \|z\|_2$.

Let us begin with the former case. W.l.o.g., assume that $\arg\max\{\|x\|_2, \|y\|_2, \|z\|_2\} = y$. Then, we can write:

$$\begin{aligned} d_2(x,y) &:= \frac{\|x-y\|_2}{\max\{\|x\|_2, \|y\|_2\}} \\ &= \frac{\|x-y\|_2}{\max\{\|x\|_2, \|y\|_2, \|z\|_2\}} \\ &\overset{(1)}{\leqslant} \frac{\|x-z\|_2}{\max\{\|x\|_2, \|y\|_2, \|z\|_2\}} + \frac{\|y-z\|_2}{\max\{\|x\|_2, \|y\|_2, \|z\|_2\}} \\ &\overset{(2)}{=} \frac{\|x-z\|_2}{\max\{\|x\|_2, \|y\|_2, \|z\|_2\}} + \frac{\|y-z\|_2}{\max\{\|y\|_2, \|z\|_2\}} \\ &\overset{(3)}{\leqslant} \frac{\|x-z\|_2}{\max\{\|x\|_2, \|z\|_2\}} + \frac{\|y-z\|_2}{\max\{\|y\|_2, \|z\|_2\}} \end{aligned}$$

$$=: d_2(x,z) + d_2(y,z),$$

where at (1) we apply triangle inequality of the $\|\cdot\|_2$ norm, at (2) we use that $\max\{\|x\|_2, \|y\|_2, \|z\|_2\} = \max\{\|y\|_2, \|z\|_2\} = \|y\|_2$, at (3) we use that, since $\max\{\|x\|_2, \|y\|_2, \|z\|_2\} = \|y\|_2$, then $\max\{\|x\|_2, \|y\|_2, \|z\|_2\} \geqslant \max\{\|x\|_2, \|z\|_2\}$.

Now, w.l.o.g., consider the case in which $\|x\|_2 \leqslant \|y\|_2 \leqslant \|z\|_2$. Since the normed vector space $\mathbb{R}^k$ with $\|\cdot\|_2$ is an inner product space, then the Ptolemy's inequality (Steele, 2004) holds:

$$\begin{aligned}\|x-y\|_2\|z\|_2 &\leqslant \|x-z\|_2\|y\|_2 + \|y-z\|_2\|x\|_2\\ &\leqslant \|x-z\|_2\|y\|_2 + \|y-z\|_2\|y\|_2\\ &= \|y\|_2\big(\|x-z\|_2 + \|y-z\|_2\big).\end{aligned}$$

By dividing both sides of the inequality by $\|z\|_2$ and $\|y\|_2$, we can write:

$$\begin{aligned}&\frac{\|x-y\|_2}{\|y\|_2} \leqslant \frac{\|x-z\|_2}{\|z\|_2} + \frac{\|y-z\|_2}{\|z\|_2}\\ \iff &\frac{\|x-y\|_2}{\max\{\|x\|_2, \|y\|_2\}} \leqslant \frac{\|x-z\|_2}{\max\{\|x\|_2, \|z\|_2\}} + \frac{\|y-z\|_2}{\max\{\|y\|_2, \|z\|_2\}}.\end{aligned}$$

This concludes the proof. □

It should be remarked that the Ptolemy's inequality holds in inner product spaces only, and that the unique p-normed vector space to be an inner product space is that with $p=2$. This is why our proof of Lemma I.2 works for function d_p defined as:

$$d_p(x,y) := \frac{\|x-y\|_p}{\max\{\|x\|_p, \|y\|_p\}},$$

if and only if $p=2$. Thanks to Lemma I.2, we are able to prove the main theorem of this section.

Theorem I.3. *Let $q \in \Delta^{[\![k]\!]}$ such that $q_i > 0$ for all $i \in [\![k]\!]$, and denote $q_{\min} := \min_{i\in[\![k]\!]} q_i$. Then, both the semimetrics d and d_∞ satisfy the ρ-relaxed triangle inequality with ρ upper bounded, respectively, by $k/q_{\min}^2$ and k.*

Proof. First, we prove the statement of the theorem for d_∞, and then we use it to prove the statement for d.

Observe that, for any $x \in \mathbb{R}^k$:

$$\|x\|_\infty \leqslant \|x\|_2 \leqslant \sqrt{k}\|x\|_\infty. \tag{51}$$

Let us consider any three vectors $x, y, z \in \mathbb{R}^k$. If $\arg\max\{\|x\|_\infty, \|y\|_\infty, \|z\|_\infty\} \neq z$, then we can proceed as in the first part of the proof of Lemma I.2 to show that $d_\infty(x,y) \leqslant d_\infty(x,z) + d_\infty(y,z)$. Therefore, w.l.o.g., we consider the case in which $\arg\max\{\|x\|_\infty, \|y\|_\infty, \|z\|_\infty\} = z$. We can write:

$$\begin{aligned}d_\infty(x,y) &= \frac{\|x-y\|_\infty}{\max\{\|x\|_\infty, \|y\|_\infty\}}\\ &\overset{(1)}{\leqslant} \frac{\|x-y\|_2}{\max\{\|x\|_\infty, \|y\|_\infty\}}\\ &\overset{(2)}{\leqslant} \frac{\sqrt{k}\|x-y\|_2}{\max\{\|x\|_2, \|y\|_2\}}\\ &= \sqrt{k}d_2(x,y)\\ &\overset{(3)}{\leqslant} \sqrt{k}d_2(x,z) + \sqrt{k}d_2(y,z)\\ &= \sqrt{k}\frac{\|x-z\|_2}{\max\{\|x\|_2, \|z\|_2\}} + \sqrt{k}\frac{\|y-z\|_2}{\max\{\|y\|_2, \|z\|_2\}}\\ &\overset{(4)}{\leqslant} \sqrt{k}\frac{\|x-z\|_2}{\max\{\|x\|_\infty, \|z\|_\infty\}} + \sqrt{k}\frac{\|y-z\|_2}{\max\{\|y\|_\infty, \|z\|_\infty\}}\end{aligned}$$

$$\overset{(5)}{\leqslant} k\frac{\|x-z\|_\infty}{\max\{\|x\|_\infty,\|z\|_\infty\}}+k\frac{\|y-z\|_\infty}{\max\{\|y\|_\infty,\|z\|_\infty\}}$$
$$=k\big(d_\infty(x,z)+d_\infty(y,z)\big),$$

where at (1) and at (2) we use Eq. 51, at (3) we use the result in Lemma I.2, and at (4) and at (5) we use again Eq. 51.

Now, we move to prove the statement concerning d. In a similar way as in the proof of Proposition 4.1, we have that, for any $x,y\in\mathbb{R}^k$:

$$\begin{aligned}d(x,y)\leqslant d_\infty(x,y)&:=\frac{\|x-y\|_\infty}{\max\{\|x\|_\infty,\|y\|_\infty\}}\\&=\frac{\max_{i\in[\![k]\!]}\frac{q_i}{q_i}|x_i-y_i|}{\max\{\|x\|_\infty,\|y\|_\infty\}}\\&\leqslant\frac{\max_{i\in[\![k]\!]}\frac{q_i}{q_{\min}}|x_i-y_i|}{\max\{\|x\|_\infty,\|y\|_\infty\}}\\&\leqslant\frac{\sum_{i\in[\![k]\!]}q_i|x_i-y_i|}{q_{\min}\max\{\|x\|_\infty,\|y\|_\infty\}}\\&=\frac{d(x,y)}{q_{\min}}.\end{aligned}$$

By using this relation in place of that in Eq. 51, we can carry out the same derivation made for d_∞ using d_2 for the semimetric d using d_∞. □

It should be remarked that we are not claiming here that the values of ρ provided in Theorem I.3 are tight[19].

I.2. The Hausdorff distance inherits the relaxed triangle inequality property

First, we show that, thanks to the definitions of d and d_∞, if we apply the Hausdorff distance to *closed* sets, then the (relaxed) triangle inequality property is satisfied. Next, we show that the sets of rewards we work with are *closed*.

Let us begin with the following proposition.

Proposition I.4. *Let $\mathcal{H}_d$ and $\mathcal{H}_\infty$ be defined as in Section 4. The* closedness *of the sets to which these distances are applied is a sufficient condition for the (relaxed) triangle inequality property to hold.*

Proof Sketch. We will not provide an exhaustive proof, since it is completely analogous to the proof that shows that *compactness* is a sufficient condition for the Hausdorff distance with inner metric to satisfy triangle inequality. Instead, we simply give an idea of why for d and d_∞ *closedness* (instead of *compactness*) suffices.

In practice, the compactness requirement is just needed to guarantee that the *infimum* is actually a *minimum* over the sets in input to the Hausdorff distance. For a generic notion of inner distance, closedness is not sufficient because the infimum might be at ∞ and, thus, the minimum would not exist. However, observe that both d and d_∞ contain the normalization term $1/M$ (see Section 4), therefore, for any finite vector $x\in\mathbb{R}^k$, getting to infinity $\lim_{y\to\infty}\|x-y\|_\infty/M=1$ worsens the distance to x w.r.t. any other finite z in the set containing y. This shows that boundedness is not required anymore, but closedness suffices. This concludes the proof. □

In this work we consider unbounded sets of rewards, so clearly compactness does not hold. The following proposition shows the closedness of some sets of rewards.

Proposition I.5. *The following sets are closed:*

$$\overline{\mathcal{R}}_{p,\pi^E},\mathcal{R}_{p,\pi^E},\mathcal{R}^{\cap}_{p,\pi^E},\mathcal{R}^{\cup}_{p,\pi^E},\widetilde{\mathcal{R}}^{\cap},\widetilde{\mathcal{R}}^{\cup}.$$

[19] Indeed, we do not believe so. By using a script to generate a large number of vectors, and using the intuition that the diagonal of the unit square ($\|\cdot\|_\infty$) is $\sqrt{2}$ the radius of the unit circle ($\|\cdot\|_2$), we conjecture that a tighter value of ρ for d_∞ is $\rho=2$, irrespective of the dimension.

Proof. From Theorem 3 of Ng & Russell (2000), we observe that the old feasible set $\overline{\mathcal{R}}_{p,\pi^E}$ is closed because it is defined by linear less than or *equal to* $\leqslant$ inequalities.

The new feasible set $\mathcal{R}_{p,\pi^E}$ can be expressed, from Corollary D.1, as an arbitrary union of closed sets $\mathcal{R}_{p,\pi^E} = \bigcup_{\pi' \in [\pi^E]_{\equiv_{\mathcal{S}^{p,\pi^E}}}} \overline{\mathcal{R}}_{p,\pi'}$. However, observe that the feasible sets $\overline{\mathcal{R}}_{p,\pi'}$ with stochastic π' are contained in the feasible sets of some deterministic policies. Since there is a finite number of deterministic policies, then $\mathcal{R}_{p,\pi^E}$ can be expressed as a finite union of closed sets, so it is closed.

The subset $\mathcal{R}^{\cap}_{p,\pi^E}$ is an arbitrary intersection of $\mathcal{R}_{p,\pi^E}$, i.e., closed sets, thus it is closed.

The superset $\mathcal{R}^{\cup}_{p,\pi^E}$ is an arbitrary union of $\mathcal{R}_{p,\pi^E}$, so, potentially, it might not be non-closed. However, thanks to the definitions of p^m and π^m in Eq. 20 and Eq. 21, we know that the arbitrary union representing $\mathcal{R}^{\cup}_{p,\pi^E}$ coincides with the feasible set $\mathcal{R}_{p^m,\pi^m}$, which is closed, thus $\mathcal{R}^{\cup}_{p,\pi^E}$ is closed.

In an analogous manner, by using Eq. 25 and Eq. 26, we observe that the relaxations $\widetilde{\mathcal{R}}^{\cap}$ and $\widetilde{\mathcal{R}}^{\cup}$ can be expressed by a finite number of linear less than or *equal to* $\leqslant$ constraints, thus they are closed. □

J. Technical Lemmas

In this section, we report some technical lemmas that are useful in the analysis of the sample complexity of **IRLO** and **PIRLO** (see Appendix F). Lemma J.1 and Lemma J.2 are taken from other works, while Lemma J.3 takes inspiration from Lemma B.9 of Metelli et al. (2021).

Lemma J.1 (Lemma A.1 of (Xie et al., 2021))**.** *Suppose that $N \sim Bin(n,p)$ is a binomially distributed random variable, with $n \geqslant 1$ and $p \in [0,1]$. Then, with probability at least $1-\delta$, we have that:*

$$\frac{p}{N \vee 1} \leqslant \frac{8 \ln \frac{1}{\delta}}{n}.$$

Lemma J.2 (Lemma 8 of (Kaufmann et al., 2021))**.** *Let $X_1, X_2, \dots, X_n, \dots$ be i.i.d. samples from a distribution supported over $[\![m]\!]$, of probabilities given by $p \in \Delta^{[\![m]\!]}$. We denote by $\widehat{p}_n$ the empirical vector of probabilities, i.e., for all $k \in [\![m]\!]$:*

$$\widehat{p}_{n,k} = \frac{1}{n} \sum_{l=1}^{n} \mathbb{1}\{X_l = k\}.$$

For all $p \in \Delta^{[\![m]\!]}$, for all $\delta \in [0,1]$:

$$\mathbb{P}\left(\exists n \in \mathbb{N}_{\geqslant 0},\, n KL(\widehat{p}_n \| p) > \ln(1/\delta) + (m-1) \ln\big(e(1 + n/(m-1))\big) \right) \leqslant \delta.$$

Lemma J.3. *Let $a, b, c, d > 0$ such that $2bc > e$. Then, the inequality $x \geqslant a + b \ln(cx + d)$ is satisfied by all $x \geqslant 2a + 3b \ln(2bc) + d/c$.*

Proof. Observe that, since function x grows faster than function $a + b\ln(cx+d)$, then there exists $\bar{x}$ such that, for all $x \geqslant \bar{x}$, the inequality is satisfied. Our goal here is to show that such $\bar{x}$ can be upper bounded by $2a + 3b\ln(2bc) + d/c$.

Let us consider any $x \geqslant 2a + d/c$. We can write:

$$\begin{aligned}
x \geqslant a + b\ln(cx+d) \quad &\iff \quad \frac{x-a}{b} \geqslant \ln(c(x \pm a) + d) \\
&\iff \quad e^{\frac{x-a}{b}} \geqslant c(x-a) + ca + d \\
&\overset{(1)}{\impliedby} \quad e^{\frac{x-a}{b}} \geqslant 2c(x-a) \\
&\iff \quad \frac{a-x}{b} e^{\frac{a-x}{b}} \geqslant -\frac{1}{2bc}, \qquad \text{(I)}
\end{aligned}$$

where at (1) we have used that, since $x \geqslant 2a + d/c$, then $c(x-a) \geqslant ca + d$, and thus we have replaced the constraint with a stronger one.

Hypothesis $2bc > e$ entails that $-\frac{1}{2bc} \geqslant -\frac{1}{e}$, thus we can apply the Lambert function, which provides as solution to inequality I all the x such that:

$$\frac{a-x}{b} \leqslant W_{-1}\left(-\frac{1}{2bc}\right) \quad \text{or} \quad \frac{a-x}{b} \geqslant W_0\left(-\frac{1}{2bc}\right),$$

where W_0 is the principal component of the Lambert W function. Consider the first inequality. We can write:

$$\begin{aligned} x &\geqslant a - bW_{-1}\left(-\frac{1}{2bc}\right) \\ &\overset{(1)}{\leqslant} a + b + b\sqrt{2\ln(2bc) - 2} + b\ln(2bc) - b \\ &\leqslant a + 3b\ln(2bc), \end{aligned}$$

where at (1) we have applied the inequality $W_{-1}(-e^{-u-1}) \geqslant -1 - \sqrt{2u} - u$ from (Chatzigeorgiou, 2013).

To obtain the result, we use that $\max\{a,b\} \leqslant a+b$ for any $a,b \geqslant 0$ to upper bound:

$$\max\{2a + d/c, a + 3b\ln(2bc)\} = a + \max\{a + d/c, 3b\ln(2bc)\} \leqslant 2a + 3b\ln(2bc) + d/c.$$

□

K. Illustrative Experiment

We have applied **PIRLO** to the highway driving application domain. To this aim, we have used the data[20] gathered by Likmeta et al. (2021).

Data Description The dataset consists of trajectories of $H = 400$ stages collected by 10 different human experts driving in a simulator. The highway has 3 lanes. The goal of each expert is to change lane in order to drive safely and to minimize the trip time. The action space $\mathcal{A}$ is made of 3 actions: Turn left, turn right, continue forward. The state space $\mathcal{S}$ is continue, and it is represented by 25 features, keeping into account the speed and position of the car, and the speed and position of the surrounding cars.

Data Preprocessing We have to transform the data to obtain a tabular MDP. To this aim, we construct 5 discrete features from the 25 present in the original data: We use three binary features, free left, free right, free forward, that say whether there is a vehicle on the left, on the right, or in front of our car; next, we use a binary feature that says whether the car is changing lane, and a discrete feature with 5 possible values for the speed of the vehicle. In this way, we obtain a tabular MDP with $S = 80$.

Experiments Design As mentioned by Likmeta et al. (2021), this lane-change scenario represents a multi-objective task, because humans consider several objectives while driving. We manually design some reward functions coherent with the most common driving objectives and we use **PIRLO** to verify whether they are compatible w.h.p. with the demonstrations of behavior provided by the 10 experts in the dataset. First, we construct a single behavioral dataset $\mathcal{D}^b$ by joining the trajectories of all the 10 experts, and then we consider one expert at a time to construct $\mathcal{D}^E$. Next, we design the reward functions and we give them in input to the membership checker implementation of **PIRLO**.

Experiments Results We design 3 kinds of reward functions:

- reward r_{BC}, i.e., the "behavioral cloning" reward, which is the reward that assigns positive values to actions played by the expert's policy;
- reward r, which is coherent with the observations provided in Section 5.3 of Likmeta et al. (2021). In words, it assigns negative reward when (i) the right lane is not free, (ii) there is a car in front of us (and so it decreases our speed), (iii) we change lane;

[20] The data is publicly available at https://github.com/amarildolikmeta/irl_real_life/tree/main/datasets/highway.

	Alice	Bob	Carol	Chuck	Craig	Dan	Erin	Eve	Grace	Judy
r_{BC}	Y,Y	Y,Y	Y,Y	Y,Y	Y,Y	Y,Y	Y,Y	Y,Y	Y,Y	Y,Y
r	N,N	Y,N	Y,N	Y,N	N,N	Y,N	Y,N	N,N	N,N	Y,N
$\overline{r}$	N,N	N,N	N,N	N,N	N,N	N,N	N,N	N,N	N,N	N,N

Table 2. The output of **PIRLO** when fed with the rewards designed for the highway driving task. The first letter refers to the superset, while the second letter refers to the subset. "N" means that the reward does not belong to the set, while "Y" means that it belongs to the set.

- reward $\overline{r}$, which is $-r$, i.e., it assigns positive reward to all the bad actions;

We provide the output of **PIRLO** in Table 2. Some comments are in order. First, our reduction to a smaller state space has caused the policies of the agents to be (more) stochastic. Moreover, this reduction has increased the number of times that the corner case described in Appendix D.5 takes place. Since this corner case is outside the good event, we have removed such data from $\mathcal{D}^b$; in this way, we improve the performances of **PIRLO**.

Observe that the behavioral cloning reward r_{BC} belongs to the subset and superset for all the experts. This is reasonable since it assigns positive reward only to expert's actions in the support of the expert's policy. However, it should be remarked that if we had not removed the "corner-case" samples, then r_{BC} would not belong to the subsets.

The reward r compatible with the analysis provided in Likmeta et al. (2021) belongs to the superset of some experts only. Specifically, for the experts Alice, Eve, Grace, and Craig, that belong to the clusters 1 and 3 of Table 1 of Likmeta et al. (2021), the reward r is not in the superset. However, it should be remarked that reward r is not exactly the same as the reward described by Likmeta et al. (2021), and also that we are working with a more aggregated state space.

Notice that, as expected, reward $\overline{r} = -r$, which rewards "bad" actions, does not belong neither to the subset nor to the superset of any expert.