
How does Inverse RL Scale to Large State Spaces? A Provably Efficient Approach

Filippo Lazzati
Politecnico di Milano
Milan, Italy
filippo.lazzati@polimi.it

Mirco Mutti
Technion
Haifa, Israel

Alberto Maria Metelli
Politecnico di Milano
Milan, Italy

Abstract

In online Inverse Reinforcement Learning (IRL), the learner can collect samples about the dynamics of the environment to improve its estimate of the reward function. Since IRL suffers from identifiability issues, many theoretical works on online IRL focus on estimating the entire set of rewards that explain the demonstrations, named the *feasible reward set*. However, none of the algorithms available in the literature can scale to problems with large state spaces. In this paper, we focus on the online IRL problem in Linear Markov Decision Processes (MDPs). We show that the structure offered by Linear MDPs is not sufficient for efficiently estimating the feasible set when the state space is large. As a consequence, we introduce the novel framework of *rewards compatibility*, which generalizes the notion of feasible set, and we develop **CATY-IRL**, a sample efficient algorithm whose complexity is independent of the cardinality of the state space in Linear MDPs. When restricted to the tabular setting, we demonstrate that **CATY-IRL** is minimax optimal up to logarithmic factors. As a by-product, we show that Reward-Free Exploration (RFE) enjoys the same worst-case rate, improving over the state-of-the-art lower bound. Finally, we devise a unifying framework for IRL and RFE that may be of independent interest.

1 Introduction

Inverse Reinforcement Learning (IRL) is the problem of inferring the reward function given demonstrations of an optimal behavior, i.e., from an *expert* agent. [49, 42]. Since its formulation, much of the research effort has been put into the design of efficient algorithms for solving the IRL problem [6, 4]. Indeed, the solution of the IRL problem opens the door to a variety of interesting applications, including Apprenticeship Learning (AL) [2, 1], reward design [16], interpretability of the expert’s behavior [17], and transferability to new environments [15].

Nowadays, the factor that most negatively impacts the adoption of IRL solutions in real-world applications is the intrinsic *ill-posedness* of its formulation. The IRL problem has been historically defined as the problem of recovering *the* reward function underlying the demonstrations [49, 42], even though mere demonstrations can be equivalently explained by a *variety* of rewards. In other words, the IRL problem is underconstrained, even in the limit of infinite demonstrations [42, 39].

To overcome this weakness and to come up with a *single* reward function, three main approaches are commonly adopted in the literature. (i) The first approach consists of the use of a *heuristic* to select a specific reward function from the set of all the rewards that explain the demonstrations. Implicitly, these works re-define IRL as the problem of recovering *the* reward function explaining the demonstrations *and* complying with the heuristic. As an example, [42, 48] select the reward that maximizes some notion of margin, and [70] implicitly chooses the reward returned by the optimization algorithm among those that maximize the likelihood. However, these approaches may generate issues in applications [56, 15]. (ii) In the second approach, additional *constraints* beyond

mere demonstrations are enforced to guarantee the uniqueness of the reward function to recover. In “reward identifiability” works, the additional information commonly concerns some structure of the environment [25], or multiple demonstrations across various environments [5, 10]. In Reward Learning (ReL) works [19], demonstrations of optimal behavior are combined with other kinds of expert feedback, like comparisons [65]. (iii) As a third approach, recently, [39, 38] proposed the alternative formulation of IRL as the problem of recovering *all* the reward functions compatible with the demonstrations, i.e., the *feasible reward set*. In this manner, we are not subject to the limitations of the first approach, and we do not depend on additional information like in the second approach.

In practical applications, the chosen IRL formulation has to be tackled by algorithms that use a *finite* number of demonstrations and a limited knowledge of the dynamics of the environment. In the common *online* IRL scenario, the learner explores the (unknown) environment, and exploits this additional information to improve its performance on the IRL task [e.g., 39, 33, 38, 68, 31]. On this basis, the IRL approach (iii) based on the *feasible set* [39, 38] displays desirable properties since “postpones” the choice of the heuristic and/or enforcement of additional constraints, with the advantage of analyzing the intrinsic complexity of the IRL problem only, without being obfuscated by other factors. In other words, this recent formulation of the IRL problem paves the way for the design and analysis of provably efficient IRL algorithms, endowed with solid theoretical guarantees.

However, the algorithms designed for learning the feasible set currently available in the literature [e.g., 39, 33, 38, 68, 31] struggle when attempting to scale them to IRL problems with *large state spaces*. This is apparent because their sample complexity exhibits an explicit dependence on the cardinality of the state space. This inevitably represents a major limitation since most real-world scenarios concern problems with large, or even continuous, state spaces [e.g., 15, 7, 40, 14].

In this context, function approximation represents an essential tool to tackle the curse of dimensionality and enforce generalization [54, 41]. Linear Markov Decision Processes (MDPs) [23, 67] offer a simple but powerful structure, in which we assume the reward function and the transition model can be expressed as linear combinations of known features, that permits theoretical analysis of the sample complexity. Even though many extensions have been developed [64, 22, 13], the Linear MDPs framework typically represents one of the first function approximation settings to analyze when focusing on a novel problem, before moving to more complex settings [e.g., 63, 61].

In this paper, we aim to shed light on the challenges of scaling the feasible reward set to large-scale problems. Motivated by its limitations when dealing with large state spaces, we introduce the novel *Rewards Compatibility* framework. Being a generalization of the notion of feasible set, it allows us to define the new *IRL Classification Problem*, a fourth approach to cope with the ill-posedness of the IRL formulation. This permits the development of **CATY-IRL** (CompATibility for IRL), a provably efficient IRL algorithm for Linear MDPs characterized by large or even continuous state spaces.

Original Contributions. The main contributions of the current work can be summarized as follows:

- We prove that the notion of feasible set can *not* be learned efficiently in MDPs with large/continuous state spaces, even under the structure enforced by Linear MDPs. Nevertheless, we show that this problem disappears under the *assumption* that the expert’s policy is known, by providing a sample efficient algorithm for such setting (Section 3).
- To overcome the need for knowing the expert’s policy exactly, we propose *Rewards Compatibility*, a novel framework that formalizes the intuitive notion of *compatibility* of a reward function with expert demonstrations. It generalizes the feasible set and allows us to define an original learning setting, *IRL classification*, based on a new formulation of IRL *classification* task (Section 4).
- For the newly-devised framework, we develop **CATY-IRL** (CompATibility for IRL), a new sample and computationally efficient IRL algorithm for both tabular and Linear MDPs. Remarkably, this **CATY-IRL** does not require the additional assumption that the expert’s policy is known (Section 5).
- In the tabular setting, we prove a tight minimax lower bound to the sample complexity of the IRL classification problem of $\Omega\left(\frac{H^3 SA}{\epsilon^2} \left(S + \log \frac{1}{\delta}\right)\right)$ episodes, where S and A are the cardinalities of the state and action spaces, H is the horizon, ϵ the accuracy and δ the failure probability. This bound is *matched* by **CATY-IRL**, up to logarithmic factors. Exploiting a similar construction, we show that a lower bound with the same rate holds also for the Reward-Free Exploration (RFE) problem, improving by an H factor over the RFE state-of-the-art lower bound [21] (Section 6.1).
- Finally, we formulate a novel *Objective-Free Exploration* (OFE) setting that isolates the challenges of exploration beyond Reinforcement Learning (RL), by unifying RFE and IRL (Section 6.2).

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

2 Preliminaries

Notation. Given an integer $N \in \mathbb{N}$, we define $\llbracket N \rrbracket := \{1, \dots, N\}$. Given sets \mathcal{X} and \mathcal{Y} , we denote $\mathcal{H}_d(\mathcal{X}, \mathcal{Y}) := \max\{\sup_{x \in \mathcal{X}} \inf_{y \in \mathcal{Y}} d(x, y), \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} d(y, x)\}$ their Hausdorff distance with inner distance d . We denote by $\Delta^{\mathcal{X}}$ the probability simplex over \mathcal{X} , and by $\Delta_{\mathcal{Y}}^{\mathcal{X}}$ the set of functions from \mathcal{Y} to $\Delta^{\mathcal{X}}$. Sometimes, we denote the dot product between vectors x, y as $\langle x, y \rangle := x^\top y$. We employ $\mathcal{O}, \Omega, \Theta$ for the common asymptotic notation and $\tilde{\mathcal{O}}, \tilde{\Omega}, \tilde{\Theta}$ to omit logarithmic terms.

Markov Decision Processes. A finite-horizon Markov Decision Process (MDP) without reward [45] is defined as a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, H, d_0, p)$, where \mathcal{S} and \mathcal{A} are the measurable state and action spaces, $H \in \mathbb{N}$ is the horizon, $d_0 \in \Delta^{\mathcal{S}}$ is the initial-state distribution, and $p \in \mathcal{P} := \Delta_{\mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket}^{\mathcal{S}}$ is the transition model. Given a (deterministic) reward function $r \in \mathfrak{R} := [-1, 1]^{\mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket}$, we denote by $\overline{\mathcal{M}} := \mathcal{M} \cup \{r\}$ the MDP obtained by pairing \mathcal{M} and r . Each policy $\pi \in \Pi := \Delta_{\mathcal{S} \times \llbracket H \rrbracket}^{\mathcal{A}}$ induces in $\overline{\mathcal{M}}$ a state-action probability distribution $d^{p, \pi} := \{d_h^{p, \pi}\}_{h \in \llbracket H \rrbracket}$ (we omit d_0 for simplicity) that assigns, to each subset $\mathcal{Z} \subseteq \mathcal{S} \times \mathcal{A}$, the probability of being in \mathcal{Z} at stage $h \in \llbracket H \rrbracket$ when playing π in $\overline{\mathcal{M}}$. We denote with $\mathcal{S}_h^{p, \pi}$ the set of states supported by $d_h^{p, \pi}$ for any action at stage h , and with $\mathcal{S}^{p, \pi}$ the disjoint union of sets $\{\mathcal{S}_h^{p, \pi}\}_{h \in \llbracket H \rrbracket}$. The Q -function of policy π in MDP $\overline{\mathcal{M}}$ is defined at every $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$ as $Q_h^\pi(s, a; p, r) := \mathbb{E}_{p, \pi}[\sum_{t=h}^H r_t(s_t, a_t) | s_h = s, a_h = a]$, and the optimal Q -function as $Q_h^*(s, a; p, r) := \sup_{\pi \in \Pi} Q_h^\pi(s, a; p, r)$, where the expectation $\mathbb{E}_{p, \pi}$ is computed over the stochastic process generated by playing policy π in the MDP $\overline{\mathcal{M}}$. Similarly, we define the V -function of policy π at (s, h) as $V_h^\pi(s; p, r) := \mathbb{E}_{p, \pi}[\sum_{t=h}^H r_t(s_t, a_t) | s_h = s]$, and the optimal V -function as $V_h^*(s; p, r) := \sup_{\pi \in \Pi} V_h^\pi(s; p, r)$. We define the utility of π as $J^\pi(r; p) := \mathbb{E}_{s \sim d_0}[V_1^\pi(s; p, r)]$, and the optimal utility as $J^*(r; p) := \mathbb{E}_{s \sim d_0}[V_1^*(s; p, r)]$. A forward (sampling) model of the environment permits to collect samples starting from $s \sim d_0$ and following some policy. A generative (sampling) model consists in an oracle that, given an arbitrary state-action-stage triple s, a, h in input, returns a sampled next state $s' \sim p_h(\cdot | s, a)$.

Linear MDPs. Based on [23], we say that an MDP $\overline{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, H, d_0, p, r)$ is a *Linear MDP* with a (known) feature map $\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$, if for every $h \in \llbracket H \rrbracket$, there exist $d \in \mathbb{N}$ unknown (signed) measures $\mu_h = [\mu_h^1, \dots, \mu_h^d]^\top$ over \mathcal{S} and an unknown vector $\theta_h \in \mathbb{R}^d$, such that for every $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have $p_h(\cdot | s, a) = \langle \phi(s, a), \mu_h(\cdot) \rangle$ and $r_h(s, a) = \langle \phi(s, a), \theta_h \rangle$. Without loss of generality, we assume $\|\phi(s, a)\|_2 \leq 1$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, and $\max\{\|\theta_h\|_2, \|\mu_h\|(\mathcal{S})\|_2\} \leq \sqrt{d}$.¹ \mathcal{M} is a *Linear MDP without reward* if its transition model satisfies the assumption described above.

BPI and RFE. In both Best-Policy Identification (BPI) [37] and Reward-Free Exploration (RFE) [21], the learner has to explore the *unknown* MDP to optimize a certain reward function. In BPI, the learner observes the reward function r during exploration, and its goal is to output a policy $\hat{\pi}$ such that, in the true MDP with transition model p we have $\mathbb{P}(J^*(r; p) - J^{\hat{\pi}}(r; p) \leq \epsilon) \geq 1 - \delta$ for every $\epsilon, \delta \in (0, 1)$. RFE considers the setting in which the reward to optimize is revealed *a posteriori* of the exploration phase. Thus the goal of the agent in RFE is to compute an estimate \hat{p} of the true dynamics p so that $\mathbb{P}(\sup_{r \in \mathfrak{R}} \{J^*(r; p) - J^{\hat{\pi}_r}(r; p)\} \leq \epsilon) \geq 1 - \delta$ for every $\epsilon, \delta \in (0, 1)$, where $\hat{\pi}_r$ is the optimal policy in the MDP with \hat{p} as transition model and r as reward function.

Online IRL. We consider the online² IRL setting [39, 33, 68, 66, 53] in which, similarly to the online AL setting [53, 66], we are given a dataset $\mathcal{D}^E = \{(s_1^i, a_1^i, \dots, s_{H-1}^i, a_{H-1}^i, s_H^i)\}_{i \in \llbracket \tau^E \rrbracket}$ of $\tau^E \in \mathbb{N}$ trajectories collected by executing the expert's policy π^E in a certain (unknown) MDP $\overline{\mathcal{M}} = \mathcal{M} \cup \{r^E\}$. We make the assumption that π^E is optimal under the true (unknown) reward r^E in $\overline{\mathcal{M}}$. Since the dynamics of $\overline{\mathcal{M}}$ is unknown, we are allowed to actively explore the environment through a *forward* model to collect a new state-action dataset \mathcal{D} . The goal is to use the latter and demonstrations in \mathcal{D}^E to estimate a reward function that makes the expert's policy π^E optimal. Sometimes, we will denote an IRL instance as $\mathcal{M} \cup \{\pi^E\}$, and a Linear IRL instance with recovered reward r as an IRL instance in which $\mathcal{M} \cup \{r\}$ is a Linear MDP.

¹ $\|\mu_h\|(\mathcal{B})$ denotes the vector containing the variation of each measure μ_h^i over the measurable set \mathcal{B} .

² "Online" refers to how we estimate the transition model p , not to the expert's policy π^E , for which we assume to have access to a batch dataset. This is justified by the fact that most of IRL real-world applications involve the presence of a fixed dataset of expert demonstrations previously collected and the agent can explore the environment in order to reconstruct one (or more) reward functions compatible with those demonstrations.

3 Limitations of the Feasible Set

In this section, after having characterized the feasible set formulation in Linear MDPs, we show that it suffers from *statistical* (and *computational*) inefficiency in problems with large state spaces, even under the Linear MDP assumption. We will provide a solution to these issues in Section 4.

The Feasible Set. According to the standard definition [e.g., 39, 33, 38, 68, 31], the feasible set contains the rewards that make the expert’s policy π^E optimal, as defined below.

Definition 3.1 (Feasible Set [31]). *Let \mathcal{M} be an MDP without reward and let π^E be the expert’s policy. The feasible set \mathcal{R}_{p,π^E} of rewards compatible with π^E in \mathcal{M} is defined as:*

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

Without function approximation, the feasible set contains a variety of rewards for any deterministic policy. In Linear MDPs, due to the feature map, the feasible set might exhibit some degeneracy.³ Definition 3.1 can be adapted to Linear MDPs with feature map ϕ as: $\mathcal{R}_{\phi,p,\pi^E} := \{r \in \mathfrak{R} \mid J^{\pi^E}(r; p) = J^*(r; p) \wedge \exists \theta : \llbracket H \rrbracket \rightarrow \mathbb{R}^d, \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket : r_h(s, a) = \langle \phi(s, a), \theta_h \rangle\}$. We omit ϕ in $\mathcal{R}_{\phi,p,\pi^E}$ for notational simplicity.

Proposition 3.1. *Let \mathcal{M} be a Linear MDP without reward with a finite state space, and let ϕ be a feature mapping. Let $\{\Phi_h^{\pi^E}\}_{h \in \llbracket H \rrbracket}$ and $\{\bar{\Phi}_h\}_{h \in \llbracket H \rrbracket}$ be the sets of expert’s and non-expert’s features, defined for every $h \in \llbracket H \rrbracket$ as:*

$$\Phi_h^{\pi^E} := \{\phi(s, a^E) \mid s \in \mathcal{S}_h^{p,\pi^E}, a^E \in \mathcal{A}_h^E(s)\}, \quad \bar{\Phi}_h := \{\phi(s, a) \mid s \in \mathcal{S}_h^{p,\pi^E}, a \in \mathcal{A} \setminus \mathcal{A}_h^E(s)\},$$

where $\mathcal{A}_h^E(s) := \{a \in \mathcal{A} \mid \pi_h^E(\cdot \mid s) > 0\}$ for every $s \in \mathcal{S}$. If for none of the H pairs of sets $(\Phi_h^{\pi^E}, \bar{\Phi}_h)$ there exists a separating hyperplane, then $\mathcal{R}_{p,\pi^E} = \{\bar{r}\}$, with $\bar{r}_h(s, a) = 0 \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$ i.e., the feasible set with linear rewards in ϕ contains only the reward function that assigns zero reward everywhere.

Intuitively, expert’s actions must have the largest optimal Q -value among all actions, and linearity imposes the “separability” requirement. The result holds also for MDPs with linear rewards only. We exemplify Proposition 3.1 in Appendix B.1.

Learning the Feasible Set. In order to highlight the challenges of learning the feasible set with large-scale MDPs, based on [38, 31], we devise the following PAC requirement.

Definition 3.2 (PAC Algorithm). *Let $\epsilon, \delta \in (0, 1)$, and let \mathfrak{A} be an algorithm that collects τ^E samples about π^E using a generative model, and τ episodes from a Linear MDP without reward $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$ using a forward model. Let $\hat{\mathcal{R}}$ be the estimate of the feasible set \mathcal{R}_{p,π^E} outputted by \mathfrak{A} . Then, \mathfrak{A} is (ϵ, δ) -PAC for IRL if $\mathbb{P}_{\mathcal{M}, \mathfrak{A}}(\mathcal{H}_d(\mathcal{R}_{p,\pi^E}, \hat{\mathcal{R}}) \leq \epsilon) \geq 1 - \delta$, where $\mathbb{P}_{\mathcal{M}, \mathfrak{A}}$ is the probability measure induced by \mathfrak{A} in \mathcal{M} , and $d(r, \hat{r}) \propto \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s,a) \sim d_h^{\pi}(\cdot, \cdot)} |r_h(s, a) - \hat{r}_h(s, a)|$.⁴ The sample complexity is the pair (τ^E, τ) .*

It is worth noting that in Definition 3.2, we are considering a generative model for collecting samples from the expert’s policy, which represents the easiest learning scenario. The following result shows that, even in this convenient setting, estimating the feasible set is statistically inefficient.

Theorem 3.2 (Statistical Inefficiency). *Let $\mathcal{M} \cup \{\pi^E\}$ be a Linear IRL instance with finite state space \mathcal{S} and deterministic expert’s policy, and let $\epsilon, \delta \in (0, 1)$. If an algorithm \mathfrak{A} is (ϵ, δ) -PAC, then $\tau^E = \Omega(S)$, where $S := |\mathcal{S}|$ is the cardinality of the state space.*

In other words, even under the easiest learning conditions (i.e., generative model and deterministic expert), the sample complexity scales directly with the cardinality of the state space S , thus, it is infeasible when S is large or even infinite. Observe that this result extends to any class of MDPs that contains Linear MDPs. In Appendix B.2, we analyze if additional assumptions can drop the $\Omega(S)$ dependence. Nevertheless, if π^E is known, it is possible to construct sample efficient algorithms. Algorithm 1 (whose pseudocode is presented in Appendix B.3), under the assumption that π^E is known, makes use of an inner RFE routine (Algorithm 1 of [62]) to recover the feasible set.

³We exemplify this proposition in Appendix B.1. In Appendix B.4 we generalize to infinite state spaces.

⁴For simplicity, we provide the full expression of distance d in Appendix B.4, Equation (1).

Theorem 3.3. Assume that π^E (along with its support \mathcal{S}^{p,π^E}) is known. Then, for any $\epsilon, \delta \in (0, 1)$, Algorithm 1 is (ϵ, δ) -PAC for IRL with a number of episodes τ upper bounded by:

$$\tau \leq \tilde{\mathcal{O}}\left(\frac{H^5 d}{\epsilon^2} \left(d + \log \frac{1}{\delta}\right)\right).$$

Limitations of the Feasible Set. We can now conclude that the feasible set suffers from two main limitations. (i) *Sample Inefficiency*: If π^E is unknown, it requires a number of samples that depends on the cardinality of the state space (Theorem 3.2). (ii) *Lack of Practical Implementability*: It contains a continuum of rewards, thus, no practical algorithm can explicitly compute it. We will discuss in the next section how to overcome both these issues.

4 Rewards Compatibility

In this section, we present the main contribution of this work: *Rewards Compatibility*, a novel framework for IRL that allows us to conveniently rephrase the learning from demonstrations problem as a classification task. We anticipate that the presentation of the framework is completely general and independent of structural assumptions of the MDP (e.g., Linear MDP).

4.1 Compatible Rewards

In the following, for ease of presentation, we consider the exact setting, i.e., when d_0, p , and π^E are known. In addition, we will drop the dependence on p when clear from the context.

In IRL, an expert agent demonstrates policy π^E assumed optimal under some (unknown) reward function r^E , i.e., $J^*(r^E) = J^{\pi^E}(r^E)$. The task is to recover a reward r such that $J^*(r) = J^{\pi^E}(r)$. By definition, IRL tells us that r^E makes the demonstrated policy π^E optimal, but what about other policies? We *do not* and *cannot* know. Since there are (infinite) rewards making π^E optimal (but they differ in the performance attributed to other policies) we realize that there are many rewards equally “compatible” with π^E .⁵ Clearly, with no additional information, we are unable to identify r^E .

The feasible set considers only these rewards, i.e., $r \in \mathfrak{R}$ for which $J^*(r) = J^{\pi^E}(r)$, and it refuses all the others. This can be interpreted as the feasible set carrying out a *classification* of rewards based on a “hard” notion of *compatibility* with demonstrations. In other words, rewards r satisfying condition $J^*(r) = J^{\pi^E}(r)$ are compatible with π^E , and the others are not. Nevertheless, our insight is that some rewards are “*more*” compatible with π^E than others.

Example 4.1. Consider an MDP with one state and $H = 1$ in which the expert has three actions: Eating a muffin (M), a cake (C), or some (bad) vegetable soup (S). The true reward r^E assigns $r^E(M) = +1, r^E(C) = +0.99$ and $r^E(S) = -1$, i.e., the expert has a (weak) preference for the muffin over the cake, while she hates the soup; thus, she will demonstrate $\pi^E = M$. Let r_g, r_b be:

$$r_g(M) = +0.99, r_g(C) = +1, r_g(S) = -1, \quad r_b(M) = -1, r_b(C) = -1, r_b(S) = +1.$$

Intuitively, r_g is “more” compatible with π^E than r_b , because it establishes that M and C are much better than S , while reward r_b reverses the preferences. Clearly, we make a small error if we model the preferences of the expert with r_g instead of the true reward r^E . However, the notion of feasible set is completely blind to the difference between r_g and r_b at modeling r^E , and it refuses both of them.

We propose the following “soft” definition of (non)compatibility to capture this intuition.⁶

Definition 4.1 (Rewards (non)Compatibility). Let $\mathcal{M} \cup \{\pi^E\}$ be an IRL instance, and let $r \in \mathfrak{R}$ be any reward. We define the (non)compatibility $\bar{\mathcal{C}}_{p,\pi^E} : \mathfrak{R} \rightarrow \mathbb{R}_{\geq 0}$ of reward r w.r.t. $\mathcal{M} \cup \{\pi^E\}$ as:

$$\bar{\mathcal{C}}_{p,\pi^E}(r) := J^*(r; p) - J^{\pi^E}(r; p).$$

⁵See Appendix C.1 for a visual intuition.

⁶In Appendix C.2, a *multiplicative* alternative definition is presented.

In words, the (non)compatibility of reward r w.r.t. policy π^E in problem \mathcal{M} quantifies the *suboptimality* of π^E in the MDP $\mathcal{M} \cup \{r\}$. By definition, rewards r belonging to the feasible set (i.e., $r \in \mathcal{R}_{p,\pi^E}$) satisfy $\bar{\mathcal{C}}_{p,\pi^E}(r) = 0$, i.e., they have zero non-compatibility with π^E in \mathcal{M} .⁷

Example 4.1 (Continued). *(Non)compatibility discriminates between r_g and r_b . Indeed, we have that $\bar{\mathcal{C}}_{p,\pi^E}(r^E) = 0$, $\bar{\mathcal{C}}_{p,\pi^E}(r_g) = 0.01$, and $\bar{\mathcal{C}}_{p,\pi^E}(r_b) = 2$. In words, reward r_g suffers from very small (non)compatibility, while r_b suffers from large (non)compatibility, thus we say that reward r_g is more compatible with π^E than r_b , as expected.*

By definition of IRL, the true reward r^E makes the observed π^E optimal, but reveals no information about the other policies. Thus, it is meaningful that $\bar{\mathcal{C}}_{p,\pi^E}$ considers the suboptimality of π^E only, because demonstrations from π^E do not provide information about other policies, as illustrated below.

Example 4.2. *Let r'_b be such that $r'_b(M) = +0.99$, $r'_b(C) = -1$, $r'_b(S) = +1$. Clearly, r'_b is much worse than r_g at modeling r^E , because it does not capture the fact that the expert appreciates the cake but she hates the soup. However, demonstrations from π^E alone do not provide information about C or S , but only about $\pi^E = M$ (i.e., the expert always eats the muffin). Thus, we have that $\bar{\mathcal{C}}_{p,\pi^E}(r_g) = \bar{\mathcal{C}}_{p,\pi^E}(r'_b) = 0.01$, i.e., r_g and r'_b are equally compatible with the given demonstrations.*

For a discussion on comparing the (non)compatibility of different rewards, see Appendix C.4.

4.2 The IRL Classification Formulation

Our goal is to overcome the limitations of the feasible set highlighted in Section 3. Drawing inspiration from the notion of “membership checker” algorithm in [31], we propose a novel formulation of IRL.

Definition 4.2 (IRL Classification Problem and IRL Algorithm). *An IRL Classification Problem instance is made of a tuple $(\mathcal{M}, \pi^E, \mathcal{R}, \Delta)$, where \mathcal{M} is an MDP without reward, π^E is the expert’s policy, $\mathcal{R} \subseteq \mathfrak{R}$ is a set of rewards to classify, and $\Delta \in \mathbb{R}_{\geq 0}$ is some threshold. The goal is to classify all and only the rewards $r \in \mathcal{R}$ based on their (non)compatibility with π^E in \mathcal{M} w.r.t. Δ . In symbols:*

$$\forall r \in \mathcal{R} : \text{ if } \bar{\mathcal{C}}_{p,\pi^E}(r) \leq \Delta \text{ then return True, else return False.}$$

An IRL algorithm takes in input a reward $r \in \mathcal{R}$ and outputs a boolean saying whether $\bar{\mathcal{C}}_{p,\pi^E}(r) \leq \Delta$.

Given $r \in \mathcal{R}$, we output whether it makes the expert’s policy π^E at most Δ -suboptimal or not. Intuitively, we classify rewards in \mathcal{R} based on how good π^E performs w.r.t. them. A Δ -(non)compatible reward guarantees that, among its Δ -optimal policies, there is π^E , but the optimal policy might be different from π^E (see Appendix C.3 for how this relates to (forward) RL). Note that we allow for $\mathcal{R} \neq \mathfrak{R}$ to manage scenarios in which we have some prior knowledge on r^E , i.e., $r^E \in \mathcal{R} \subset \mathfrak{R}$.

Remark 4.1. *Permitting non-zero (non)compatibility is equivalent to enlarging the feasible set. Let $\mathcal{R} = \mathfrak{R}$, and define the set of rewards positively classified as \mathcal{R}_Δ , i.e., $\mathcal{R}_\Delta := \{r \in \mathcal{R} \mid \bar{\mathcal{C}}_{p,\pi^E}(r) \leq \Delta\}$. For any Δ, Δ' s.t. $0 \leq \Delta \leq \Delta' \leq 2H$, we have: $\mathcal{R}_{p,\pi^E} = \mathcal{R}_0 \subseteq \mathcal{R}_\Delta \subseteq \mathcal{R}_{\Delta'} \subseteq \mathcal{R}_{2H} = \mathfrak{R}$.*

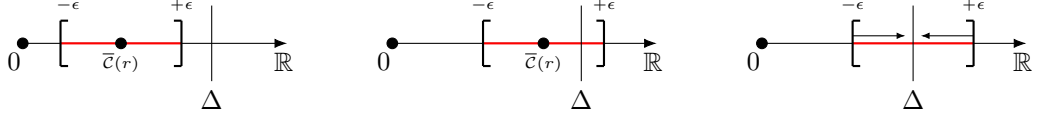
Discussion on Reward Compatibility. It should be remarked that:

- *The limits of the rewards compatibility framework are the same as the limits of the feasible set.* We cannot identify r^E from the feasible set or among the rewards with small (non)compatibility. As aforementioned, this is an inherent limit of IRL and cannot be overcome with a more refined objective formulation, unless further information on r^E is available (e.g., preferences).
- *Rewards compatibility offers advantages over feasible set.* Differently from the feasible set, as we will see in Section 5, it is possible to *practically* implement algorithms that solve the IRL classification problem, with guarantees of sample efficiency even when the state space is large.

4.3 A Learning Framework for Online IRL Classification

In this section, we combine the online IRL setting presented in Section 2 with the IRL classification problem of Definition 4.2. Intuitively, the performance of an algorithm depends on its accuracy at estimating the (non)compatibility of the rewards, as formalized by the following PAC requirement.

⁷We use (non)compatibility since a reward $r \in \mathfrak{R}$ is maximally compatible when $\bar{\mathcal{C}}_{p,\pi^E}(r) = 0$. Thus, the larger $\bar{\mathcal{C}}_{p,\pi^E}(r)$, the more r is non-compatible. In this sense, $\bar{\mathcal{C}}_{p,\pi^E}(r)$ quantifies the non-compatibility of r .



(a) Reward r is classified correctly. (b) Reward r can be misclassified. (c) Range of uncertain (non)compatibility values.

Figure 2: The axis represents (estimated) (non)compatibility values. (a) Rewards r whose true (non)compatibility $\bar{C}(r) := \bar{C}_{p, \pi^E}(r)$ is far from threshold Δ by at least ϵ , are correctly classified, while (b) in the opposite case, rewards can be misclassified. (c) The red interval $[\Delta - \epsilon, \Delta + \epsilon]$ exemplifies the set of rewards $\{r \in \mathcal{R} \mid |\bar{C}(r) - \Delta| \leq \epsilon\}$ that are (potentially) misclassified. The length of the interval reduces with ϵ .

Definition 4.3 (PAC Framework). Let $\epsilon, \delta \in (0, 1)$, and let \mathcal{D}^E be a dataset of τ^E expert’s trajectories. An algorithm \mathfrak{A} exploring for τ episodes is (ϵ, δ) -PAC for the IRL classification problem if:

$$\mathbb{P}_{\mathcal{M}, \pi^E, \mathfrak{A}} \left(\sup_{r \in \mathcal{R}} |\bar{C}_{p, \pi^E}(r) - \hat{C}(r)| \leq \epsilon \right) \geq 1 - \delta,$$

where $\mathbb{P}_{\mathcal{M}, \pi^E, \mathfrak{A}}$ is the joint probability measure induced by π^E and \mathfrak{A} in \mathcal{M} , and \hat{C} is the estimate of \bar{C}_{p, π^E} computed by \mathfrak{A} . The sample complexity is defined by the pair (τ^E, τ) .

Intuitively, our goal is to estimate the (non)compatibility of the rewards in \mathcal{R} with sufficient accuracy, so that, given a threshold $\Delta \geq 0$, we are able to classify “most” of them correctly w.h.p. (with high probability). The concept is exemplified in Figure 2. Note that the estimation problem is independent of the threshold Δ , which can be appropriately selected to cope with noise in the demonstrations, (unknown) expert suboptimality, or to manage the amount of “false negatives” and “false positives”.

Remark 4.2. For $\eta \geq 0$, let $\mathcal{R}_\eta := \{r \in \mathcal{R} \mid \bar{C}_{p, \pi^E}(r) \leq \eta\}$ and $\hat{\mathcal{R}}_\eta := \{r \in \mathcal{R} \mid \hat{C}(r) \leq \eta\}$ denote the sets of rewards positively classified using, respectively, the true (non)compatibility \bar{C}_{p, π^E} and the estimate \hat{C} constructed by an (ϵ, δ) -PAC algorithm. Then, with probability $1 - \delta$, it holds that: $\hat{\mathcal{R}}_{\Delta - \epsilon} \subseteq \mathcal{R}_\Delta \subseteq \hat{\mathcal{R}}_{\Delta + \epsilon}$. Thus, we can trade-off the amount of “false negatives” (resp. “false positives”) by, e.g., choosing the threshold $\Delta \leftarrow \Delta + \epsilon$ (resp. $\Delta \leftarrow \Delta - \epsilon$).

5 CATY-IRL: A Provably Efficient Algorithm for IRL

In this section, we present **CATY-IRL** (CompATibility for IRL), a provably efficient algorithm for solving the *online IRL classification* problem. We consider three different kinds of structure for the MDPs: tabular MDPs, tabular MDPs with linear rewards, and Linear MDPs. Similarly to RFE, our online IRL classification setting is made of two phases: (i) an *exploration* phase, in which the algorithm explores the environment using the knowledge of \mathcal{R} and of the expert’s dataset \mathcal{D}^E to collect samples about the dynamics of the MDP, and (ii) a *classification* phase, in which it performs the classification of a reward $r \in \mathcal{R}$ without interactions with the environment. A flow-chart is reported in Figure 1 (pseudocode in Appendix D).

Exploration phase. The *exploration* phase collects a dataset \mathcal{D} in a way that depends on the structure of the MDP and of the set of rewards \mathcal{R} to be classified. Specifically, for Linear MDPs,

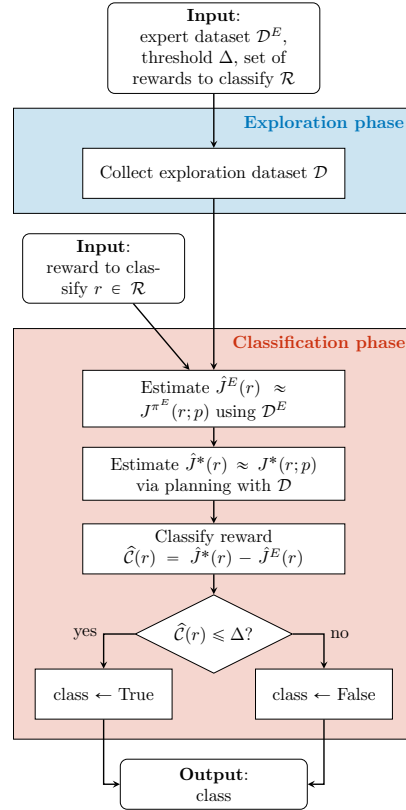


Figure 1: Flow-chart of **CATY-IRL**.

CATY-IRL executes RFLin [62]. Instead, for tabular MDPs (with or without linear reward), **CATY-IRL** instantiates either BPI-UCBVI [37] for each reward $r \in \mathcal{R}$ (when $|\mathcal{R}| = \Theta(1)$, i.e., a “small” constant w.r.t. to the size of the MDP, where “small” depends on the size of the state space, see Appendix D.2) or RF-Express [37]. Note that **CATY-IRL** in this phase does not use the expert’s dataset \mathcal{D}^E .

Classification phase. The *classification* performs the estimation $\hat{C}(r)$ of the (non)compatibility term $\bar{C}_{p,\pi^E}(r)$ for the single input reward $r \in \mathcal{R}$ by splitting it into two independent estimates: $\hat{J}^E(r) \approx J^{\pi^E}(r; p)$, which is computed with \mathcal{D}^E only, and $\hat{J}^*(r) \approx J^*(r; p)$, which is computed with \mathcal{D} only. Concerning $\hat{J}^E(r)$, when the reward is linear $r_h(s, a) = \langle \phi(s, a), \theta_h \rangle$, **CATY-IRL** uses \mathcal{D}^E to construct an empirical estimate $\hat{\psi}^E \approx \psi^{p,\pi^E}$ of the expert’s expected feature count [6]. Otherwise, it directly estimates $\hat{d}^E \approx d^{p,\pi^E}$ the expert’s occupancy measure. Such estimates can be used to derive $\hat{J}^E(r)$ straightforwardly. Regarding $\hat{J}^*(r)$, **CATY-IRL** exploits the *planning* phase of the corresponding RFE (or BPI) algorithm adopted at exploration phase.⁸ Finally, **CATY-IRL** applies the (potentially negative) input threshold Δ to the difference $\hat{J}^*(r) - \hat{J}^E(r)$ to perform the classification. See Appendix D for the full pseudo-code. Clearly, **CATY-IRL** can be implemented in practice, since it considers a single reward at a time instead of computing the full feasible set, and it is computationally efficient in linear MDPs, since it uses a computationally efficient algorithm as subroutine (see [62]).

Sample Efficiency. The next result analyzes the sample complexity (Definition 4.3) of **CATY-IRL**.

Theorem 5.1 (Sample Complexity of **CATY-IRL**). *Let $\epsilon, \delta \in (0, 1)$. Then **CATY-IRL** is (ϵ, δ) -PAC for IRL with a sample complexity upper bounded by:*

$$\text{Tabular MDPs:} \quad \tau^E \leq \tilde{O}\left(\frac{H^3 SA}{\epsilon^2} \log \frac{1}{\delta}\right), \quad \tau \leq \tilde{O}\left(\frac{H^3 SA}{\epsilon^2} \left(N + \log \frac{1}{\delta}\right)\right),$$

$$\text{Tabular MDPs with linear rewards:} \quad \tau^E \leq \tilde{O}\left(\frac{H^3 d}{\epsilon^2} \log \frac{1}{\delta}\right), \quad \tau \leq \tilde{O}\left(\frac{H^3 SA}{\epsilon^2} \left(N + \log \frac{1}{\delta}\right)\right),$$

$$\text{Linear MDPs:} \quad \tau^E \leq \tilde{O}\left(\frac{H^3 d}{\epsilon^2} \log \frac{1}{\delta}\right), \quad \tau \leq \tilde{O}\left(\frac{H^5 d}{\epsilon^2} \left(d + \log \frac{1}{\delta}\right)\right),$$

where $N = 0$ if $|\mathcal{R}| = \Theta(1)$, and $N = S$ otherwise.

Some observations are in order. We conjecture that the d^2 dependence when $|\mathcal{R}| = \Theta(1)$ is unavoidable in Linear MDPs because of the lower bound for BPI in [62]. In tabular MDPs with deterministic expert, one might use the results in [66] to reduce the rate of τ^E from $\tilde{O}(SAH^3 \log(\delta^{-1})/\epsilon^2)$ to $\tilde{O}(SH^{3/2} \log(\delta^{-1})/\epsilon^2)$. Finally, note that the choice $\Delta = \epsilon$ allows us to positively classify all the rewards in the feasible set \mathcal{R}_{p,π^E} w.h.p. and, in this case, other rewards positively classified have true (non)compatibility at most 2ϵ w.h.p. In light of this result we conclude that *rewards compatibility* framework allows the *practical* development of *sample efficient* algorithms (e.g., **CATY-IRL**) in Linear MDPs with large/continuous state spaces.

6 Statistical Barriers and Objective-Free Exploration

In this section, we show that **CATY-IRL** is minimax optimal for the number of exploration episodes in tabular MDPs, and that RFE and IRL share the same theoretical sample complexity. This allows us to formulate *Objective-Free Exploration*, a unifying setting for exploration problems.

6.1 The Theoretical Limits of IRL (and RFE) in the Tabular Setting

In **CATY-IRL**, we use a minimax optimal RFE algorithm for exploration. However, this does not entail that **CATY-IRL** is minimax optimal for the IRL classification problem. There might exist another PAC algorithm with a sample complexity smaller than **CATY-IRL**. The following result states that, in the tabular setting, the bound in Theorem 5.1 is tight for the number of episodes τ .

⁸RFE/BPI algorithms, at planning phase, return a policy, and not its estimated performance. Since BPI-UCBVI, RF-Express, and RFLin each compute an estimate of $J^*(r; p)$ as an intermediate step, with negligible abuse of notation, we assume that they output such estimate.

Theorem 6.1 (IRL Classification - Lower Bound). *Let \mathfrak{A} be an (ϵ, δ) -PAC algorithm for the IRL classification in tabular MDPs. Let τ be the number of exploration episodes. Then, there exists an IRL classification instance such that:*

$$\text{if } |\mathcal{R}| \geq 1 : \tau \geq \Omega\left(\frac{H^3 SA}{\epsilon^2} \log \frac{1}{\delta}\right), \quad \text{if } \mathcal{R} = \mathfrak{R} : \tau \geq \Omega\left(\frac{H^3 SA}{\epsilon^2} \left(S + \log \frac{1}{\delta}\right)\right).$$

In both cases, the lower bound is *matched* by **CATY-IRL**, up to logarithmic factors. Note that **CATY-IRL** explores without using \mathcal{D}^E , thus, minimax optimality for τ can be achieved without the knowledge of \mathcal{D}^E at exploration phase. As a by-product, we observe that a similar lower bound construction can be made also for RFE, leading to the following result.

Theorem 6.2 (RFE - Refined Lower Bound). *Let \mathfrak{A} be an (ϵ, δ) -PAC algorithm for RFE in tabular MDPs. Let τ be the number of exploration episodes. Then, there exists an RFE instance such that:*

$$\tau \geq \Omega\left(\frac{H^3 SA}{\epsilon^2} \left(S + \log \frac{1}{\delta}\right)\right).$$

This bound improves the state-of-the-art RFE lower bound $\Omega\left(\frac{H^3 SA}{\epsilon^2} \left(\frac{S}{H} + \log \frac{1}{\delta}\right)\right)$ (obtained combining the bounds in [21] and [12]) by one H factor, and it is matched by RF-Express [37].

6.2 Objective-Free Exploration (OFE)

What is the most efficient exploration strategy that can be performed in an unknown environment? It *depends* on the subsequent task that shall be solved. However, if the task is unknown at the exploration phase, we need a strategy that suffices for all the tasks that one might be interested in solving. Let us denote by \mathcal{F} the set of RL and IRL classification tasks. Since **CATY-IRL** is a sample efficient algorithm for the IRL classification problem, and it uses RFE as a subroutine, we conclude that the RFE exploration strategy is sufficient (and also minimax optimal in tabular MDPs) to obtain guarantees for class \mathcal{F} . Are there other problems for which RFE exploration suffices when the specific problem instance is revealed *a posteriori* of the exploration phase? We believe so, and in Appendix E, we identify two additional problems, i.e., Matching Performance (MP) and Imitation Learning from Observations alone (ILfO) [34], that represent potential candidates to belong to \mathcal{F} .

More in general, we formulate the *Objective-Free Exploration (OFE)* problem as follows:

Definition 6.1 (Objective-Free Exploration). *Given a tuple $(\mathcal{M}, \mathcal{F}, (\epsilon, \delta))$, where \mathcal{M} is an unknown environment (e.g., MDP without reward), and \mathcal{F} is a certain class of tasks (e.g., all RL and IRL problems), the Objective-Free Exploration (OFE) problem aims to find an exploration of the environment \mathcal{M} (e.g., RFE exploration) that permits to solve any task $f \in \mathcal{F}$ in an (ϵ, δ) -correct manner.*

This problem is called “objective-free” because it does not require the knowledge of the specific “objective” $f \in \mathcal{F}$ to be solved. In Appendix F, we describe a use case for OFE. We believe this is an interesting problem to be studied in future.

7 Conclusions

In this paper, we have shown that the feasible set cannot be learned efficiently in problems with large/continuous state spaces even under the strong structure provided by Linear MDPs. For this reason, we have introduced the powerful framework of *compatible rewards*, which formalizes the intuitive notion of compatibility of a reward function with expert demonstrations, and it allows us to formulate the IRL problem as a *classification* task. In this context, we have devised **CATY-IRL**, a provably efficient IRL algorithm for Linear MDPs with large/continuous state spaces. Furthermore, in tabular MDPs, we have demonstrated the minimax optimality of **CATY-IRL** at exploration by presenting a novel lower bound to the IRL classification problem. As a by-product, our construction improves the current state-of-the-art lower bound for RFE. Finally, we have introduced OFE, a unifying problem setting for exploration problems, which generalizes both RFE and IRL.

Limitations. A limitation of our contributions concerns the adoption of the *Linear MDP* model, whose assumptions are overly strong to be consistently applied to real-world applications. Nevertheless, while the rewards compatibility framework is general and not tied to Linear MDPs, we believe that Linear MDPs represent an important initial step toward the development of provably efficient IRL

algorithms with more general function approximation structures. Although a lower bound for Linear MDPs is missing, we believe that it represents an interesting direction for future works. Finally, we note that the *empirical validation* of the proposed algorithm is out of the scope of this work.

Future Directions. Promising directions for future works concern the extension of the analysis of the *rewards compatibility* framework beyond Linear MDPs to general function approximation and to the offline setting. In addition, it might be fascinating to extend the notion of reward compatibility to other kinds of expert feedback (in the context of ReL), and to other IRL settings (e.g., suboptimal experts). Finally, we believe that OFE should be analysed in-depth given its practical importance.

Acknowledgments and Disclosure of Funding

AI4REALNET has received funding from European Union’s Horizon Europe Research and Innovation programme under the Grant Agreement No 101119527. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

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

- [1] Pieter Abbeel, Adam Coates, Morgan Quigley, and Andrew Ng. An application of reinforcement learning to aerobatic helicopter flight. In *Advances in Neural Information Processing Systems 19 (NeurIPS)*, 2006.
- [2] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning 21 (ICML)*, 2004.
- [3] Pieter Abbeel and Andrew Y. Ng. Exploration and apprenticeship learning in reinforcement learning. In *International Conference on Machine Learning 22 (ICML)*, 2005.
- [4] Stephen Adams, Tyler Cody, and Peter A. Beling. A survey of inverse reinforcement learning. *Artificial Intelligence Review*, 55:4307–4346, 2022.
- [5] Kareem Amin and Satinder Singh. Towards resolving unidentifiability in inverse reinforcement learning, 2016.
- [6] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2018.
- [7] Matt Barnes, Matthew Abueg, Oliver F. Lange, Matt Deeds, Jason Trader, Denali Molitor, Markus Wulfmeier, and Shawn O’Banion. Massively scalable inverse reinforcement learning in google maps, 2024.
- [8] P. Dimitri Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.
- [9] Michael Bowling, John D. Martin, David Abel, and Will Dabney. Settling the reward hypothesis. In *International Conference on Machine Learning 40 (ICML)*, 2023.
- [10] Haoyang Cao, Samuel Cohen, and Lukasz Szpruch. Identifiability in inverse reinforcement learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 12362–12373, 2021.
- [11] Gregory Dexter, Kevin Bello, and Jean Honorio. Inverse reinforcement learning in a continuous state space with formal guarantees. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 6972–6982, 2021.
- [12] Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *International Conference on Algorithmic Learning Theory 32 (ALT)*, volume 132, pages 578–598, 2021.

- [13] Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning 38 (ICML)*, volume 139, pages 2826–2836, 2021.
- [14] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning 33 (ICML)*, volume 48, pages 49–58, 2016.
- [15] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. In *International Conference on Learning Representations 5 (ICLR)*, 2017.
- [16] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.
- [17] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, 2016.
- [18] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, 2016.
- [19] Hong Jun Jeon, Smitha Milli, and Anca Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 4415–4426, 2020.
- [20] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning 34 (ICML)*, volume 70, pages 1704–1713, 2017.
- [21] Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning 37 (ICML)*, volume 119, pages 4870–4879, 2020.
- [22] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 13406–13418, 2021.
- [23] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory 33 (COLT)*, volume 125, pages 2137–2143, 2020.
- [24] Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. In *International Conference on Algorithmic Learning Theory 32 (ALT)*, volume 132, pages 865–891, 2021.
- [25] Kuno Kim, Shivam Garg, Kirankumar Shiragur, and Stefano Ermon. Reward identification in inverse reinforcement learning. In *International Conference on Machine Learning 38 (ICML)*, pages 5496–5505, 2021.
- [26] Kuno Kim, Yihong Gu, Jiaming Song, Shengjia Zhao, and Stefano Ermon. Domain adaptive imitation learning. In *International Conference on Machine Learning 37 (ICML)*, volume 119, pages 5286–5295, 2020.
- [27] Edouard Klein, Matthieu Geist, Bilal Piot, and Olivier Pietquin. Inverse reinforcement learning through structured classification. In *Advances in Neural Information Processing Systems 25 (NeurIPS)*, 2012.
- [28] Abi Komanduru and Jean Honorio. On the correctness and sample complexity of inverse reinforcement learning. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.
- [29] Abi Komanduru and Jean Honorio. A lower bound for the sample complexity of inverse reinforcement learning. In *International Conference on Machine Learning 38 (ICML)*, volume 139, pages 5676–5685, 2021.

- [30] David M. Kreps. *Notes On The Theory Of Choice*. Westview Press, 1988.
- [31] Filippo Lazzati, Mirco Mutti, and Alberto Maria Metelli. Offline inverse rl: New solution concepts and provably efficient algorithms. In *International Conference on Machine Learning 41 (ICML)*, 2024.
- [32] Gen Li, Yuling Yan, Yuxin Chen, and Jianqing Fan. Minimax-optimal reward-agnostic exploration in reinforcement learning, 2023.
- [33] David Lindner, Andreas Krause, and Giorgia Ramponi. Active exploration for inverse reinforcement learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 5843–5853, 2022.
- [34] YuXuan Liu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Imitation from observation: Learning to imitate behaviors from raw video via context translation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1118–1125, 2018.
- [35] Zhihan Liu, Yufeng Zhang, Zuyue Fu, Zhuoran Yang, and Zhaoran Wang. Learning from demonstration: Provably efficient adversarial policy imitation with linear function approximation. In *International Conference on Machine Learning 39 (ICML)*, volume 162, pages 14094–14138, 2022.
- [36] Manuel Lopes, Francisco Melo, and Luis Montesano. Active learning for reward estimation in inverse reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 31–46, 2009.
- [37] Pierre Menard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning 38 (ICML)*, volume 139, pages 7599–7608, 2021.
- [38] Alberto Maria Metelli, Filippo Lazzati, and Marcello Restelli. Towards theoretical understanding of inverse reinforcement learning. In *International Conference on Machine Learning 40 (ICML)*, volume 202, pages 24555–24591, 2023.
- [39] Alberto Maria Metelli, Giorgia Ramponi, Alessandro Concetti, and Marcello Restelli. Provably efficient learning of transferable rewards. In *International Conference on Machine Learning 38 (ICML)*, volume 139, pages 7665–7676, 2021.
- [40] Bernard Michini, Mark Cutler, and Jonathan P. How. Scalable reward learning from demonstration. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 303–308, 2013.
- [41] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013.
- [42] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning 17 (ICML)*, pages 663–670, 2000.
- [43] Donald Ornstein. On the existence of stationary optimal strategies. *Proceedings of the American Mathematical Society*, 20(2):563–569, 1969.
- [44] Riccardo Poiani, Gabriele Curti, Alberto Maria Metelli, and Marcello Restelli. Inverse reinforcement learning with sub-optimal experts, 2024.
- [45] Martin Lee Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.
- [46] Nived Rajaraman, Yanjun Han, Lin Yang, Jingbo Liu, Jiantao Jiao, and Kannan Ramchandran. On the value of interaction and function approximation in imitation learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, volume 34, pages 1325–1336, 2021.
- [47] Nived Rajaraman, Lin Yang, Jiantao Jiao, and Kannan Ramchandran. Toward the fundamental limits of imitation learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 2914–2924, 2020.

- [48] Nathan D. Ratliff, J. Andrew Bagnell, and Martin A. Zinkevich. Maximum margin planning. In *International Conference on Machine Learning 23 (ICML)*, pages 729–736, 2006.
- [49] Stuart Russell. Learning agents for uncertain environments (extended abstract). In *Conference on Computational Learning Theory 11 (COLT)*, pages 101–103, 1998.
- [50] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edition, 2010.
- [51] Rohin Shah, Noah Gundotra, Pieter Abbeel, and Anca Dragan. On the feasibility of learning, rather than assuming, human biases for reward inference. In *International Conference on Machine Learning 36 (ICML)*, volume 97, pages 5670–5679, 2019.
- [52] Mehran Shakerinava and Siamak Ravanbakhsh. Utility theory for sequential decision making. In *International Conference on Machine Learning 39 (ICML)*, volume 162, pages 19616–19625, 2022.
- [53] Lior Shani, Tom Zahavy, and Shie Mannor. Online apprenticeship learning. In *AAAI Conference on Artificial Intelligence 36 (AAAI)*, pages 8240–8248, 2022.
- [54] David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503, 2016.
- [55] Joar Skalse and Alessandro Abate. Misspecification in inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence 37 (AAAI)*, pages 15136–15143, 2023.
- [56] Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *International Conference on Machine Learning 40 (ICML)*, volume 202, pages 32033–32058, 2023.
- [57] Wen Sun, Anirudh Vemula, Byron Boots, and Drew Bagnell. Provably efficient imitation learning from observation alone. In *International Conference on Machine Learning 36 (ICML)*, volume 97, pages 6036–6045, 2019.
- [58] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, 2018.
- [59] Gokul Swamy, Nived Rajaraman, Matt Peng, Sanjiban Choudhury, J. Bagnell, Steven Z. Wu, Jiantao Jiao, and Kannan Ramchandran. Minimax optimal online imitation learning via replay estimation. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, volume 35, pages 7077–7088, 2022.
- [60] Umar Syed and Robert E Schapire. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems 20 (NeurIPS)*, 2007.
- [61] Luca Viano, Stratis Skoulakis, and Volkan Cevher. Imitation learning in discounted linear mdps without exploration assumptions, 2024.
- [62] Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. Reward-free RL is no harder than reward-aware RL in linear Markov decision processes. In *International Conference on Machine Learning 39 (ICML)*, volume 162, pages 22430–22456, 2022.
- [63] Ruosong Wang, Simon S Du, Lin Yang, and Russ R Salakhutdinov. On reward-free reinforcement learning with linear function approximation. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 17816–17826, 2020.
- [64] Ruosong Wang, Ruslan Salakhutdinov, and Lin F. Yang. Reinforcement learning with general value function approximation: provably efficient approach via bounded eluder dimension. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 6123–6135, 2020.

- [65] Christian Wirth, Riad Akrouf, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18:1–46, 2017.
- [66] Tian Xu, Ziniu Li, Yang Yu, and Zhimin Luo. Provably efficient adversarial imitation learning with unknown transitions. In *Conference on Uncertainty in Artificial Intelligence 39 (UAI)*, volume 216, pages 2367–2378, 2023.
- [67] Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning 36 (ICML)*, volume 97, pages 6995–7004, 2019.
- [68] Lei Zhao, Mengdi Wang, and Yu Bai. Is inverse reinforcement learning harder than standard reinforcement learning? In *International Conference on Machine Learning 41 (ICML)*, 2024.
- [69] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory 34 (COLT)*, pages 4532–4576, 2021.
- [70] Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence 23 (AAAI)*, pages 1433–1438, 2008.

A Related Works

In this appendix, we report and describe the literature that most relates to this paper. Theoretical works concerning the online IRL problem can be grouped in works that concern the feasible set, and works that do not.

Let us begin with works related to the feasible set. While the notion of feasible set has been introduced implicitly in [42], the first paper that analyses the sample complexity of estimating the feasible set in online IRL is [39]. Authors in [39] adopt the simple generative model in tabular MDPs, and devise two sample efficient algorithms. [33] focuses on the same problem as [39], but adopts a forward model in tabular MDPs. By adopting RFE exploration algorithms, they devise sample efficient algorithms. However, as remarked in [68], paper [33] suffers from a limitation in the definition of the dissimilarity between feasible sets. [38] builds upon [39] to construct the first minimax lower bound for the problem of estimating the feasible set using a generative model. The lower bound is in the order of $\Omega\left(\frac{H^3 SA}{\epsilon^2} \left(S + \log \frac{1}{\delta}\right)\right)$, where S and A are the cardinality of the state and action spaces, H is the horizon, ϵ is the accuracy and δ the failure probability. In addition, [38] develops US-IRL, an efficient algorithm whose sample complexity matches the lower bound. [44] analyze a setting analogous to that of [38], in which there is availability of a single optimal expert and multiple suboptimal experts with known suboptimality. [31] analyse the problem of estimating the feasible set when no active exploration of the environment is allowed, but the learner is given a batch dataset collected by some behavior policy π^b . Interestingly, [31] focuses on two novel learning targets that are suited for the offline setting, i.e., a subset and a superset of the feasible set. Authors in [31] demonstrate that such sets are the tightest learnable subset and superset of the feasible set, and propose a pessimistic algorithm, PIRLO, to estimate them. [68] analyses the same offline setting as [31], but instead of focusing on the notion of feasible set directly, it considers the notion of reward mapping, which considers reward functions as parametrized by their value and advantage functions, and whose image coincides with the feasible set.

With regards to online IRL works that do not consider the feasible set, we mention [36], which analyses an active learning framework for IRL. However, [36] assumes that the transition model is known, and its goal is to estimate the expert policy only. Works [28] and [29] provide, respectively, an upper bound and a lower bound to the sample complexity of IRL for β -strict separable problems in the tabular setting. However, both the setting considered and the bound obtained are fairly different from ours. Analogously, [11] provides a sample efficient IRL algorithm for β -strict separable problems with continuous state space. However, their setting is different from ours since they assume that the system can be modelled using a basis of orthonormal functions.

A.1 Additional Related Works

In this section, we collect additional related works that deserve to be mentioned.

Identifiability and Reward Learning. As aforementioned, the IRL problem is ill-posed, thus, to retrieve a single reward, additional constraints shall be imposed. [5] analyses the setting in which demonstrations of an optimal policy for the same reward function are provided across environments with different transition models. In this way, authors can reduce the experimental unidentifiability, and recover the state-only reward function. [10] and [26] concern reward identifiability but in entropy-regularized MDPs [70, 15]. Such setting is in some sense easier than the common IRL setting, because entropy-regularization permits a unique optimal policy for any reward function. [10] uses expert demonstrations from multiple transition models and multiple discount factors to retrieve the reward function, while [26] analyses properties of the dynamics of the MDP to increase the constraints. With regards to the more general field of Reward Learning (ReL), we mention [19], which introduces a framework that formalizes the constraints imposed by various kinds of human feedback (like demonstrations or preferences [65]). Intuitively, multiple feedbacks about the same reward represent additional constraints beyond mere demonstrations. [56] characterizes the partial identifiability of the reward function based on various reward learning data sources.

Linear MDPs and Extensions. As explained for instance in [23], since lower bounds to the sample complexity of various RL tasks in tabular MDPs depend explicitly on the cardinality state space S , then we need to add structure to the problem if we want to develop efficient algorithms that scale to large state spaces. For this reason, the works [67, 23] analyze the Linear MDP model, which enforces

some linearity constraints to the common MDP model. In this way, authors are able to provide efficient algorithms for RL in problems with large/continuous state spaces. However, there are other settings beyond Linear MDPs that are analysed in the RL literature. [20] introduces the notion of Bellman rank as complexity measure, and provides a sample efficient algorithm for problems with small Bellman rank. [64] analyzes general value function approximation when the function class has a low eluder dimension. [22] generalizes both the eluder dimension and Bellman rank complexity measures by defining the Bellman eluder dimension and providing a provably efficient algorithm. [13] introduces bilinear classes, a structural framework that, among the others, generalizes Linear MDPs.

Reward-Free Exploration (RFE) in Tabular and Linear MDPs. The RFE problem was introduced in [21], where authors provided a sample efficient algorithm and a lower bound for tabular MDPs. Later on, the state-of-the-art sample-efficient algorithms for RFE in tabular MDPs have been developed in [24, 37, 32]. It should be remarked that RFE requires more samples than common RL in tabular MDPs. [63] proposes a sample efficient algorithm for RFE in linear MDPs. [62] improves the algorithm of [63] and, interestingly, demonstrates that RFE is no harder than RL in Linear MDPs.

Online Apprenticeship Learning (AL). The first works that provide a theoretical analysis of the AL setting when the transition model is unknown are [3, 60]. Recently, [53] formulates the online AL problem, which closely resembles the online IRL problem. The main difference is that in online AL the ultimate goal is to imitate the expert, while in IRL is to recover a reward function. [66] improves the results in [53] by combining an RFE algorithm with an efficient algorithm for the estimation of the visitation distribution of the deterministic expert’s policy in tabular MDPs, presented in [47]. We mention also [46, 59] for the sample complexity of estimating the expert’s policy in problems with linear function approximation. In the context of Imitation Learning from Observation alone (ILfO) [34], the work [57] proposes a probably efficient algorithm for large-scale MDPs with unknown transition model. [35] provides an efficient AL algorithm based on GAIL [18] in Linear Kernel Episodic MDPs [69] with unknown transition model.

Others. We mention work [27], which considers a classification approach for IRL. However, this is fairly different from our IRL problem formulation in Section 4.

B Additional Results and Proofs for Section 3

In this section, we provide additional results beyond those presented in Section 3, and then we report the missing proofs. Specifically, in Appendix B.1, we provide two numerical examples that explain Proposition 3.1, in Appendix B.2 we show that some additional regularity assumptions beyond the Linear MDP cannot remove the dependence on the cardinality of the state space in the sample complexity. In Appendix B.3, we report and describe the sample efficient algorithm mentioned in Section 3, while in Appendix B.4 we collect all the missing proofs of this section.

B.1 Some Examples for Proposition 3.1

The following examples aim to explain Proposition 3.1 in a simple manner.

Example B.1 (Non-degenerate feasible set). *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0) \cup \{\pi^E\}$ be an IRL instance such that $\mathcal{S} = \{s_1, s_2\}$, $\mathcal{A} = \{a_1, a_2\}$, $H = 1$, $d_0(s_1) = d_0(s_2) = 1/2$, $\pi^E(s_1) = \pi^E(s_2) = a_1$. Consider the feature mapping ϕ_1 s.t. $\phi_1(s, a) = \mathbb{1}\{a = a_1\}$ for all $s \in \mathcal{S}$. Then, we have $\Phi^{\pi^E} = \{1\}$ and $\bar{\Phi} = \{0\}$. Clearly, these sets can be separated by any hyperplane $w \in \mathbb{R}_{>0}$, since $1 \cdot w > 0 \cdot w$, and so $\mathcal{R}_{p, \pi^E} \neq \{\bar{r}\}$, with $\bar{r}_h(s, a) = 0 \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$. Actually, $\mathcal{R}_{p, \pi^E} = \{r \in \mathfrak{R} \mid \exists \theta \in (0, 1] : r_1(s, a) = \langle \phi(s, a), \theta \rangle \forall (s, a) \in \mathcal{S} \times \mathcal{A}\}$.*

Example B.2 (Degenerate feasible set). *Consider the same IRL instance as in the previous example, but this time consider the feature mapping ϕ_2 s.t. $\phi_2(s_1, a) = \mathbb{1}\{a = a_1\}$, and $\phi_2(s_2, a) = \mathbb{1}\{a = a_2\}$. Then, we have $\Phi^{\pi^E} = \{0, 1\}$ and $\bar{\Phi} = \{0, 1\}$. Clearly, the two sets coincide, thus they cannot be separated, and $\mathcal{R}_{p, \pi^E} = \{\bar{r}\}$, with $\bar{r}_h(s, a) = 0 \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$.*

B.2 Additional Regularity Assumptions of the State Space do not Make the Problem Learnable

In tabular MDPs with small state space \mathcal{S} , collecting samples from every state $s \in \mathcal{S}$ is feasible, and it is exactly what previous works do:

- Under the assumption that π^E is deterministic, [39, 38] collect one sample from every $(s, h) \in \mathcal{S} \times \llbracket H \rrbracket$ using a generative model, obtaining π^E exactly.
- If π^E is stochastic, under the assumption that all actions in the support of the expert’s policy are played with probability at least π_{\min} (see Assumption D.1 of [38]), both [38, 68] are able to learn the support of π^E exactly w.h.p. using $\propto 1/\pi_{\min}$ samples in the online setting.⁹
- In the offline setting, assuming that the occupancy measure of the expert’s policy is at least d_{\min} in all reachable $(s, a) \in \mathcal{S} \times \mathcal{A}$, then [31] learns the support of π^E exactly w.h.p. using $\propto 1/d_{\min}$ episodes.

However, when \mathcal{S} is large, even under the Linear MDP assumption, this is not possible. In Section 3, we have formalized this fact with the following proposition:

Theorem 3.2 (Statistical Inefficiency). *Let $\mathcal{M} \cup \{\pi^E\}$ be a Linear IRL instance with finite state space \mathcal{S} and deterministic expert’s policy, and let $\epsilon, \delta \in (0, 1)$. If an algorithm \mathfrak{A} is (ϵ, δ) -PAC, then $\tau^E = \Omega(S)$, where $S := |\mathcal{S}|$ is the cardinality of the state space.*

Theorem 3.2 tells us that the Linear MDP assumption is too weak for the feasible set to be learnable using the PAC framework of Definition 3.2 with a number of samples independent of the cardinality of the state space. Therefore, we can try to introduce an additional assumption on the structure of the IRL problem $\mathcal{M} \cup \{\pi^E\}$ and see whether it helps in alleviating the issue. Let us consider the following first assumption.

Assumption B.1. *We assume a Lipschitz continuity property between features and states:*

$$\forall (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} : \quad \|\phi(s, a) - \phi(s', a)\|_2 \leq L\|s - s'\|,$$

for some $L > 0$ and some distance $\|\cdot - \cdot\|$ in \mathcal{S} .

The intuition is that, based on the fact that in Linear MDPs the Q -function of any policy π is linear in the feature mapping $Q_h^\pi(\cdot, \cdot) = (\phi(\cdot, \cdot), w_h^\pi)$ for some parameter vector $w_h^\pi \in \mathbb{R}^d$ (see [23]), then if we are able to ϵ -cover the state space \mathcal{S} , we can approximate the Q -function $Q_h^\pi(s, \cdot)$ in any $s \in \mathcal{S}$ with the Q -function $Q_h^\pi(s', \cdot)$ of the closest point s' into the covering, so that $|Q_h^\pi(s, a) - Q_h^\pi(s', a)| = |(\phi(s, a) - \phi(s', a))^\top w_h^\pi| \leq \|\phi(s, a) - \phi(s', a)\|_2 \|w_h^\pi\|_2 \leq L\epsilon \|w_h^\pi\|_2$. However, this assumption is not sufficient.

Proposition B.1. *Under the setting of Proposition 3.2, even under Assumption B.1, then an algorithm is (ϵ, δ) -PAC only if $\tau^E = \Omega(S)$.*

Assumption B.1 fails because it does not provide any information about how the knowledge of the expert’s policy at a state can be “transferred” to other states, and thus we still need to sample almost all the states of $\mathcal{S}^{\mathcal{P}, \pi^E}$ to get an acceptable feasible set.

We devise another assumption to attempt to fix this issue.

Assumption B.2. *We assume the following Lipschitz continuity property:*

$$\forall (s, s') \in \mathcal{S} \times \mathcal{S} : \quad \|\phi(s, \pi_h^E(s)) - \phi(s', \pi_h^E(s'))\|_2 \leq L\|s - s'\|,$$

for some $L > 0$ and some distance $\|\cdot - \cdot\|$ in \mathcal{S} .

This assumption says that states that are close to each other cannot have the features corresponding to the expert’s action too far away from each other. From a high-level point of view, it says that the features are “somehow” regular with π^E , so that when the expert lies in s' which is really close to s , then she plays an action which has the same “effect” (i.e., same transition model and same reward, due to the Linear MDP assumption) as the expert’s action in s .

Assumption B.2 is not comparable with Assumption B.1 since, on the one hand, it does not hold for all actions in \mathcal{A} , but only for those corresponding to π^E , but, on the other hand, provides information on how to transfer knowledge about π^E to neighbor states.

⁹Actually, [68] makes use of a concentrability assumption too.

Let $\Delta' := \min_{s \in \mathcal{S}, a, a' \in \mathcal{A}: \phi(s, a) \neq \phi(s, a')} \|\phi(s, a) - \phi(s, a')\|_2$, i.e., the smallest non-zero distance between the features of different actions. Clearly, when \mathcal{S} is finite, since in Linear MDPs also $\mathcal{A} := |\mathcal{A}|$ is finite, then Δ' is finite too. So we can define a new quantity Δ to be any number $0 < \Delta < \Delta'$.

Proposition B.2. *Under the setting of Proposition 3.2, under Assumption B.2, then a number of samples $\tau^E = |\mathcal{N}(\frac{\Delta}{2L}; \mathcal{S}, \|\cdot\|)|$ is sufficient to recover π^E exactly in any $(s, h) \in \mathcal{S}$, where $|\mathcal{N}(\frac{\Delta}{2L}; \mathcal{S}, \|\cdot\|)|$ is the $\Delta/(2L)$ -covering number of space \mathcal{S} w.r.t. distance $\|\cdot\|$.*

Intuitively, by constructing a covering with a sufficiently small radius in the state space \mathcal{S} , then we are able to retrieve the exact expert’s action in the neighborhood of each state of the covering. Doing so, we are able to construct ϵ -correct estimates of the feasible set. Of course, this is possible as long as Δ' is not too small, and L is not too large. When \mathcal{S} is infinitely large or continuous, it might be possible to construct feature mappings in which $\Delta' \rightarrow 0$, and so the approach would still require too many samples.

However, even for cases with finite and not too small Δ' , the result in Proposition B.2 is not satisfactory, because it just allows to retrieve π^E under a stronger assumption than Linear MDPs, but not to perform an interesting learning process. We observe that the feasible set is an “unstable” concept, in the sense that, based on Proposition 3.1, changing the expert action in a single state might reduce the feasible set from a continuum of rewards to a singleton, or vice versa.

Remark B.1. *If we want to be able to recover the exact feasible set efficiently, we need to recover the exact expert’s policy almost everywhere.*

B.3 Algorithm

By exploiting an RFE algorithm as sub-routine like that of Algorithm 1 in [63] or Algorithm 1 in [62], we are able to construct estimates of the transition model \hat{p} , that can be used to compute an “empirical” estimate of the feasible set $\hat{\mathcal{R}} \approx \mathcal{R}_{\hat{p}, \pi^E}$ (since ϕ and π^E are known). The algorithm is presented in Algorithm 1.

Algorithm 1: IRL for Linear MDPs (known expert’s policy)

Data: failure probability $\delta > 0$, error tolerance $\epsilon > 0$, expert policy π^E , all sets $\mathcal{Z} \subseteq \mathcal{S} \times \llbracket H \rrbracket$
that coincide with \mathcal{S}^{p, π^E} almost everywhere based on measure d^{p, π^E}

```

1  $\mathcal{D} \leftarrow \text{RFE\_Exploration}(\delta, \epsilon)$  /* Various choices */
2 for  $h$  in  $\{H, H - 1, \dots, 2, 1\}$  do
3    $\Lambda_h \leftarrow I + \sum_{k=1}^{\tau} \phi(s_h^k, a_h^k) \phi(s_h^k, a_h^k)^\top$ 
4    $\hat{\mu}_h(\cdot) \leftarrow \Lambda_h^{-1} \sum_{k=1}^{\tau} \phi(s_h^k, a_h^k) \delta(\cdot, s_{h+1}^k)$ 
5 end
6  $\hat{p}_h(\cdot | s, a) \leftarrow \langle \phi(s, a), \hat{\mu}_h(\cdot) \rangle$  for all  $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$ 
7  $\hat{\mathcal{R}} \leftarrow \{\hat{r} \in \mathfrak{R} \mid \exists \mathcal{Z}, \forall (s, h) \in \mathcal{Z}, \forall a \in \mathcal{A} : \mathbb{E}_{a' \sim \pi_h^E(\cdot | s)} Q_h^*(s, a'; \hat{p}, \hat{r}) \geq Q_h^*(s, a; \hat{p}, \hat{r})\}$ 
8 Return  $\hat{\mathcal{R}}$ 
```

Simply put, Algorithm 1 uses the dataset collected by an RFE algorithm to compute a least-squares estimate of the transition model \hat{p} , and then it returns the feasible set defined according to it (recall that ϕ and π^E are known). Notice that this algorithm cannot be implemented in practice due to various reasons, like the presence of the Dirac delta δ measure in the definition of some quantities (see Appendix B.4.3), and the fact that the feasible set is, potentially, a set containing infinite rewards. Nevertheless, Theorem 3.3 states that this algorithm is sample efficient. The proof of the theorem is provided in Appendix B.4.3.

It should be remarked that Algorithm 1 takes in input also the true support of the visit distribution of the expert policy \mathcal{S}^{p, π^E} in case \mathcal{S} is finite, and all the possible sets \mathcal{Z} that agree with \mathcal{S}^{p, π^E} a.e. based on the measure d^{p, π^E} in case \mathcal{S} is infinite. Intuitively, this set (\mathcal{S}^{p, π^E}) of (s, h) pairs represents

the domain in which π^E is defined. Indeed, since the expert in the true problem p never visits pairs $(s', h') \notin \mathcal{S}^{p, \pi^E}$, its expert policy might reasonably be non well-defined there. When \mathcal{S} is infinite, we require all sets \mathcal{Z} because otherwise we cannot know which are the sets $\mathcal{S}^{p, \pi^E} \setminus \mathcal{Z}$ with zero measure, i.e., in which the reward can induce an optimal action different from the expert's one, since the overall contribution to the expected return is zero.

The proof of Theorem 3.3 is obtained by using Algorithm 1 of [62] at Line 1 of Algorithm 1. In Appendix B.4.3, we demonstrate an upper bound also if we use Algorithm 1 in [63].

B.4 Missing Proofs

Before diving into the proofs, we recall some important properties of the feasible set and of the Linear MDPs that will be useful in the proofs. First, we provide an explicit form for the feasible set presented at Definition 3.1.

Lemma B.3 (Lemma E.1 in [31]). *In the setting of Definition 3.1, if \mathcal{S} is finite, then the feasible set \mathcal{R}_{p, π^E} satisfies:*

$$\mathcal{R}_{p, \pi^E} = \left\{ r \in \mathfrak{R} \mid \forall (s, h) \in \mathcal{S}^{p, \pi^E}, \forall a \in \mathcal{A} : \mathbb{E}_{a' \sim \pi_h^E(\cdot|s)} Q_h^*(s, a'; p, r) \geq Q_h^*(s, a; p, r) \right\}.$$

Notice that we have extended Lemma E.1 in [31] to consider stochastic expert policies (the extension is trivial). We can easily extend it to problems with large/continuous \mathcal{S} .

Lemma B.4 (Feasible Set Explicit). *In the setting of Definition 3.1, then the feasible set \mathcal{R}_{p, π^E} satisfies:*

$$\mathcal{R}_{p, \pi^E} = \left\{ r \in \mathfrak{R} \mid \forall h \in \llbracket H \rrbracket, \exists \bar{\mathcal{S}} \subseteq \mathcal{S}_h^{p, \pi^E} : d_h^{p, \pi^E}(\bar{\mathcal{S}}) = 0 \wedge \forall s \notin \bar{\mathcal{S}}, \forall a \in \mathcal{A} : \mathbb{E}_{a' \sim \pi_h^E(\cdot|s)} Q_h^*(s, a'; p, r) \geq Q_h^*(s, a; p, r) \right\}.$$

Simply, Lemma B.4 improves on Lemma B.3 by allowing the reward to enforce the “wrong” action (i.e., different from the expert’s action) in a subset with zero measure based on the visitation distribution.

Proof. The proof is completely analogous to that of Lemma E.1 in [31]. We just need to observe that if set $\bar{\mathcal{S}}$ has zero measure (and the set of rewards \mathfrak{R} contains bounded rewards), then it does not affect the expected return. \square

Another useful property that we need is that the Q -function is always linear in the feature map for any policy in Linear MDPs.

Proposition B.5 (Proposition 2.3 in [23]). *For a Linear MDP, for any policy π , there exist weights $\{w_h^\pi\}_{h \in \llbracket H \rrbracket}$ such that, for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$, we have $Q_h^\pi(s, a) = \langle \phi(s, a), w_h^\pi \rangle$.*

We can combine the results of Lemma B.4 and Proposition B.5 to obtain the following characterization of the feasible set in Linear MDPs.

Lemma B.6. *In the setting of Definition 3.1, the feasible set \mathcal{R}_{p, π^E} satisfies:*

$$\begin{aligned} \mathcal{R}_{p, \pi^E} = \left\{ r \in \mathfrak{R} \mid \right. & \exists \{w_h\}_{h \in \llbracket H \rrbracket}, \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket : r_h(s, a) = \langle \phi(s, a), \theta_h \rangle \\ & \wedge \forall h \in \llbracket H \rrbracket, \exists \bar{\mathcal{S}} \subseteq \mathcal{S}_h^{p, \pi^E} : d_h^{p, \pi^E}(\bar{\mathcal{S}}) = 0 \wedge \forall s \notin \bar{\mathcal{S}}, \forall a^E \in \mathcal{A}_h^E(s) : \\ & \left. \langle \phi(s, a^E), w_h \rangle = \max_{a \in \mathcal{A}} \langle \phi(s, a), w_h \rangle \right\}, \end{aligned}$$

where $\theta_h := w_h - \int_{\mathcal{S}} \max_{a' \in \mathcal{A}} \langle \phi(s', a'), w_{h+1} \rangle d\mu_h(s')$ for all $h \in \llbracket H \rrbracket$, and $\mathcal{A}_h^E(s) := \{a \in \mathcal{A} \mid \pi_h^E(a|s) > 0\}$.

Proof. From [45], we know that in any MDP there exists an optimal policy. Therefore, thanks to Proposition B.5, we know that the optimal Q -function Q^* is linear in the feature map too. So, there

exist parameters $\{w_h\}_h$ such that, for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$, the optimal Q -function can be rewritten as $Q_h^*(s, a) = \langle \phi(s, a), w_h \rangle$. From the Bellman equation, we know that:

$$\begin{aligned} Q_h^*(s, a; p, r) &= r_h(s, a) + \int_{\mathcal{S}} V_{h+1}^*(s'; p, r) dp_h(s'|s, a) \\ &= \langle \phi(s, a), \theta_h \rangle + \langle \phi(s, a), \int_{\mathcal{S}} \max_{a' \in \mathcal{A}} \langle \phi(s', a'), w_{h+1} \rangle d\mu_h(s') \rangle. \end{aligned}$$

By rearranging this equation, and removing the dot product with $\phi(s, a)$, we obtain that:

$$\theta_h = w_h - \int_{\mathcal{S}} \max_{a' \in \mathcal{A}} \langle \phi(s', a'), w_{h+1} \rangle d\mu_h(s').$$

Now, this holds in any Linear MDP. If we desire to enforce the constraints in Lemma B.4, we simply have to impose the constraint on the optimal Q -function using parameters $\{w_h\}_h$ outside some $\bar{\mathcal{S}}$. This concludes the proof. \square

It is useful to introduce the following definitions. First we define the set of parameters that induce a Q -function compatible with π^E :

$$\begin{aligned} \mathcal{W}_{p, \pi^E} := \left\{ w : \llbracket H \rrbracket \rightarrow \mathbb{R}^d \mid \forall h \in \llbracket H \rrbracket, \exists \bar{\mathcal{S}} \subseteq \mathcal{S}_h^{p, \pi^E} : d_h^{p, \pi^E}(\bar{\mathcal{S}}) = 0 \wedge \forall s \notin \bar{\mathcal{S}}, \forall a^E \in \mathcal{A}_h^E(s) : \right. \\ \left. \langle \phi(s, a^E), w_h \rangle = \max_{a \in \mathcal{A}} \langle \phi(s, a), w_h \rangle \right\}. \end{aligned}$$

Next, we define the set of parameters of the reward function obtained by using Q -functions parametrized by $w \in \mathcal{W}_{p, \pi^E}$:

$$\Theta_{p, \pi^E} := \left\{ \theta : \llbracket H \rrbracket \rightarrow \mathbb{R}^d \mid \exists \{w_h\}_h \in \mathcal{W}_{p, \pi^E} : \theta_h = w_h - \int_{\mathcal{S}} \max_{a' \in \mathcal{A}} \langle \phi(s', a'), w_{h+1} \rangle d\mu_h(s') \right\}.$$

Irrespective of the transition model $\{\mu_h\}_h$ and the feature map ϕ , we see that it is always possible to construct a surjective map from Θ_{p, π^E} to \mathcal{W}_{p, π^E} (the map in the definition of Θ_{p, π^E}). Thanks to these definitions, the feasible set can be rewritten as:

$$\mathcal{R}_{p, \pi^E} = \{r \in \mathfrak{R} \mid \exists \{\theta_h\}_h \in \Theta_{p, \pi^E}, \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket : r_h(s, a) = \langle \phi(s, a), \theta_h \rangle\}.$$

We are now ready to provide the proofs of the various results of this section.

B.4.1 Proof of Proposition 3.1

Proposition 3.1. *Let \mathcal{M} be a Linear MDP without reward with a finite state space, and let ϕ be a feature mapping. Let $\{\Phi_h^{\pi^E}\}_{h \in \llbracket H \rrbracket}$ and $\{\bar{\Phi}_h\}_{h \in \llbracket H \rrbracket}$ be the sets of expert's and non-expert's features, defined for every $h \in \llbracket H \rrbracket$ as:*

$$\Phi_h^{\pi^E} := \{\phi(s, a^E) \mid s \in \mathcal{S}_h^{p, \pi^E}, a^E \in \mathcal{A}_h^E(s)\}, \quad \bar{\Phi}_h := \{\phi(s, a) \mid s \in \mathcal{S}_h^{p, \pi^E}, a \in \mathcal{A} \setminus \mathcal{A}_h^E(s)\},$$

where $\mathcal{A}_h^E(s) := \{a \in \mathcal{A} \mid \pi_h^E(\cdot|s) > 0\}$ for every $s \in \mathcal{S}$. If for none of the H pairs of sets $(\Phi_h^{\pi^E}, \bar{\Phi}_h)$ there exists a separating hyperplane, then $\mathcal{R}_{p, \pi^E} = \{\bar{r}\}$, with $\bar{r}_h(s, a) = 0 \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$ i.e., the feasible set with linear rewards in ϕ contains only the reward function that assigns zero reward everywhere.

Proof. From [8], we recall that two sets $\mathcal{Y}_1, \mathcal{Y}_2$ are separated by a hyperplane $H = \{x \mid a^\top x = b\}$ if each lies in a different closed halfspace associated with H , i.e., if either:

$$a^\top y_1 \leq b \leq a^\top y_2, \quad \forall y_1 \in \mathcal{Y}_1, \forall y_2 \in \mathcal{Y}_2,$$

or:

$$a^\top y_2 \leq b \leq a^\top y_1, \quad \forall y_1 \in \mathcal{Y}_1, \forall y_2 \in \mathcal{Y}_2.$$

By definition of \mathcal{W}_{p, π^E} , for each stage $h \in \llbracket H \rrbracket$, we are looking for vectors $w_h \in \mathbb{R}^d$ such that $\forall (s, h) \in \mathcal{S}^{p, \pi^E}$, it holds that:

$$w_h^\top \phi(s, a) \leq w_h^\top \phi(s, a^E) \quad \forall a^E \in \mathcal{A}_h^E(s), \forall a \in \mathcal{A} \setminus \mathcal{A}_h^E(s).$$

In words, for each $(s, h) \in \mathcal{S}^{p, \pi^E}$, we are looking for non-affine separating hyperplanes between features of expert and non-expert actions. However, since the hyperplane parameter w_h is common to all states $s \in \mathcal{S}_h^{p, \pi^E}$, then it must separate expert from non-expert actions at all states. This is equivalent to finding the separating hyperplanes to the sets $\Phi_h^{\pi^E}$ and $\bar{\Phi}_h$ which contain all the points. Clearly, when the separating hyperplanes do not exist at all $h \in \llbracket H \rrbracket$, then the condition in \mathcal{W}_{p, π^E} is satisfied by the zero vector alone. As a consequence, set Θ_{p, π^E} contains only the zero vector, and so does \mathcal{R}_{p, π^E} . \square

Remark B.2. *By using the result of Lemma B.4, we can easily convert Proposition 3.1 into a more general result by considering the impossibility of separating any pair of sets constructed by varying at will some subsets with zero measure. We will not provide such result explicitly.*

B.4.2 Proofs of Proposition 3.2 and Appendix B.2

In the PAC framework of Definition 3.2, we have not specified formally the inner distance d :

$$d(r, \hat{r}) := \frac{1}{M_{r, \hat{r}}} \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s, a) \sim d_h^{p, \pi}(\cdot, \cdot)} |r_h(s, a) - \hat{r}_h(s, a)|, \quad (1)$$

where:

$$M_{r, \hat{r}} := \max\{\sqrt{d}, \max_{h \in \llbracket H \rrbracket} \|\theta_h\|_2, \max_{h \in \llbracket H \rrbracket} \|\hat{\theta}_h\|_2\} / \sqrt{d},$$

where $\{\theta_h\}_h$ and $\{\hat{\theta}_h\}_h$ are the (unbounded) parameters of rewards r and \hat{r} . As explained in [31], such normalization term allows us to work with unbounded reward functions. In practice, we are relaxing the Linear MDP assumption presented in Section 2 about the boundedness of the parameters θ of the rewards to avoid the issue described in [38] and [31]. We still assume that the feature mapping is bounded. Observe that this relaxation does *not* affect the results we present, which would hold even if we considered bounded parameters θ . Indeed, as visible in the proofs, the instances do not need to be constructed with unbounded θ .

Theorem 3.2 (Statistical Inefficiency). *Let $\mathcal{M} \cup \{\pi^E\}$ be a Linear IRL instance with finite state space \mathcal{S} and deterministic expert's policy, and let $\epsilon, \delta \in (0, 1)$. If an algorithm \mathfrak{A} is (ϵ, δ) -PAC, then $\tau^E = \Omega(S)$, where $S := |\mathcal{S}|$ is the cardinality of the state space.*

Proof. We construct two problem instances that lie at a finite Hausdorff distance, and show that, with less than S calls to the sampling oracle, we are not able to discriminate between the two instances.

Let \mathcal{S} be the finite state space with cardinality S , $\mathcal{A} = \{a_1, a_2\}$, $H = 1$, $d_0(s) = 1/S \forall s \in \mathcal{S}$, $\phi(s, a) = \mathbb{1}\{a = a_1\}$, and consider two deterministic expert's policies $\pi_1^E(s) = a_1 \forall s \in \mathcal{S}$, and $\pi_2^E(s) = a_1 \forall s \in \mathcal{S} \setminus \{\bar{s}\}$, and $\pi_2^E(\bar{s}) = a_2$, for a certain $\bar{s} \in \mathcal{S}$. The set of parameters compatible with π_1^E is:

$$\Theta_{p, \pi_1^E} = \{\theta \in \mathbb{R} \mid \theta \geq 0\},$$

since $Q^{\pi_1^E}(s, a_1) \geq Q^{\pi_1^E}(s, a_2) \iff r(s, a_1) \geq r(s, a_2) \iff \phi(s, a_1)\theta \geq \phi(s, a_2)\theta \iff 1 \cdot \theta \geq 0 \cdot \theta$. Observe that, for π_2^E , due to the presence of \bar{s} , we have:

$$\Theta_{p, \pi_2^E} = \{\theta \in \mathbb{R} \mid \theta = 0\},$$

since \bar{s} imposes $\theta \leq 0$, and the other states impose $\theta \geq 0$.

Therefore, the Hausdorff distance between the two problems is:

$$\mathcal{H}(\mathcal{R}_{\pi_1^E}, \mathcal{R}_{\pi_2^E}) = \sup_{\theta \geq 0} \frac{1}{\max\{1, \theta, 0\}} \theta = \sup_{\theta \geq 0} \frac{1}{\max\{1, \theta\}} \theta = 1.$$

Obviously, we need a $\Omega(S)$ samples to spot, if it exists, state \bar{s} , and thus distinguish between $\mathcal{R}_{\pi_1^E}$ and $\mathcal{R}_{\pi_2^E}$. \square

Proposition B.1. *Under the setting of Proposition 3.2, even under Assumption B.1, then an algorithm is (ϵ, δ) -PAC only if $\tau^E = \Omega(S)$.*

Proof. The same proof of Proposition 3.2 works here.

In particular, we now show that Assumption B.1 does not help. The Hausdorff distance between the instances in the proof of Proposition 3.2 can be written as:

$$\begin{aligned}
\mathcal{H}(\mathcal{R}_{\pi_1^E}, \mathcal{R}_{\pi_2^E}) &= \sup_{\theta_1 \geq 0} \inf_{\theta_2 = 0} \frac{1}{\max\{1, \theta_1, \theta_2\}} \sup_{\pi \in \Pi} \mathbb{E}_{s \sim d_0(\cdot), a \sim \pi(\cdot|s)} |r_1(s, a) - r_2(s, a)| \\
&= \sup_{\theta_1 \geq 0} \inf_{\theta_2 = 0} \frac{1}{\max\{1, \theta_1\}} \sup_{\pi \in \Pi} \mathbb{E}_{s \sim d_0(\cdot), a \sim \pi(\cdot|s)} |\phi(s, a)\theta_1 - \phi(s, a)\theta_2 \\
&\quad \pm \phi(s', a)\theta_1 \pm \phi(s', a)\theta_2| \\
&\leq \sup_{\pi \in \Pi} \mathbb{E}_{s \sim d_0(\cdot), a \sim \pi(\cdot|s)} |\phi(s, a) - \phi(s', a)| + 0 \\
&\quad + \sup_{\pi \in \Pi} \sup_{\theta_1 \geq 0} \frac{1}{\max\{1, \theta_1\}} \inf_{\theta_2 = 0} \mathbb{E}_{s \sim d_0(\cdot), a \sim \pi(\cdot|s)} |\phi(s', a)\theta_1 - \phi(s', a)\theta_2| \\
&= \sup_{\pi \in \Pi} \mathbb{E}_{s \sim d_0(\cdot), a \sim \pi(\cdot|s)} |\phi(s, a) - \phi(s', a)| + \sup_{\pi \in \Pi} \mathbb{E}_{s \sim d_0(\cdot), a \sim \pi(\cdot|s)} \phi(s', a),
\end{aligned}$$

where s' is the state in the covering closest to state s ; while the first term can be bounded, the assumption does not help us with the second term. \square

Proposition B.2. *Under the setting of Proposition 3.2, under Assumption B.2, then a number of samples $\tau^E = |\mathcal{N}(\frac{\Delta}{2L}; \mathcal{S}, \|\cdot\|)|$ is sufficient to recover π^E exactly in any $(s, h) \in \mathcal{S}$, where $|\mathcal{N}(\frac{\Delta}{2L}; \mathcal{S}, \|\cdot\|)|$ is the $\Delta/(2L)$ -covering number of space \mathcal{S} w.r.t. distance $\|\cdot\|$.*

Proof. For any state $s \in \mathcal{S}^{p, \pi^E}$, by definition of covering $\mathcal{N}(\frac{\Delta}{2L}; \mathcal{S}, \|\cdot\|)$, there always exist another state $s' \in \mathcal{N}(\frac{\Delta}{2L}; \mathcal{S}, \|\cdot\|)$ such that $\|s' - s\| \leq \frac{\Delta}{2L}$. By Assumption B.2 we know that:

$$\|\phi(s, \pi_h^E(s)) - \phi(s', \pi_h^E(s'))\|_2 \leq L\|s' - s\| \leq \frac{\Delta}{2},$$

and since $\pi_h^E(s')$ and thus $\phi(s', \pi_h^E(s'))$ is known, then the fact that Δ is finite guarantees us that $\pi_h^E(s)$ is equal to the action a that minimizes the distance to $\phi(s', \pi_h^E(s'))$. Notice that if, by contradiction, there were two actions a_1, a_2 with $\|\phi(s, a_1) - \phi(s', \pi_h^E(s'))\|_2 \leq \frac{\Delta}{2}$ and $\|\phi(s, a_2) - \phi(s', \pi_h^E(s'))\|_2 \leq \frac{\Delta}{2}$, then by triangle inequality and finiteness of Δ , we would have:

$$\begin{aligned}
\Delta &< \|\phi(s, a_1) - \phi(s, a_2)\|_2 \\
&\leq \|\phi(s, a_1) - \phi(s', \pi_h^E(s'))\|_2 + \|\phi(s, a_2) - \phi(s', \pi_h^E(s'))\|_2 \\
&\leq \frac{\Delta}{2} + \frac{\Delta}{2} = \Delta,
\end{aligned}$$

which is clearly a contradiction. \square

B.4.3 Proof of Theorem 3.3

The proof is based on deriving an upper bound to the Hausdorff distance between the true feasible set and its estimate. To do so, first, using the notation of [23], let us define the following quantities:

$$\begin{aligned}
\mathbb{P}_h(\cdot|s, a) &:= \langle \phi(s, a), \mu_h(\cdot) \rangle, \\
\hat{\mathbb{P}}_h(\cdot|s, a) &:= \phi(s, a)^\top \Lambda_h^{-1} \sum_{k=1}^{\tau} \phi(s_h^k, a_h^k) \delta(\cdot, s_{h+1}^k), \\
\bar{\mathbb{P}}_h(\cdot|s, a) &:= \phi(s, a)^\top \Lambda_h^{-1} \sum_{k=1}^{\tau} \phi(s_h^k, a_h^k) \mathbb{P}_h(\cdot|s_h^k, a_h^k),
\end{aligned}$$

where $\delta(\cdot, x)$ is the Dirac measure, and (s_h^k, a_h^k) represents the state-action pair visited at stage h of exploration episode $k \in [\tau]$. In words, \mathbb{P} denotes the true transition model, $\hat{\mathbb{P}}$ denotes the least squares estimate computed by Algorithm 1, and $\bar{\mathbb{P}}$ represents a bridge between the two. As we will see, the core of the proof consists in upper bounding the term $|(\mathbb{P}_h - \hat{\mathbb{P}})V_{h+1}(s, a)|$ at all $h \in [H]$

and reachable $(s, a) \in \mathcal{S} \times \mathcal{A}$, for all the bounded linear functions V in class \mathcal{V} , defined as:

$$\mathcal{V} := \left\{ V : \mathcal{S} \times \llbracket H \rrbracket \rightarrow [-H, +H] \mid V(\cdot) = \max_{a \in \mathcal{A}} \phi(\cdot, a)^\top w, \|w\|_2 \leq 2H\sqrt{d} \right\}. \quad (2)$$

To achieve this goal, it will be useful to apply triangle inequality and to bound the following two terms separately:

$$\left| (\mathbb{P}_h - \widehat{\mathbb{P}}) V_{h+1}(s, a) \right| \leq \left| (\mathbb{P}_h - \overline{\mathbb{P}}) V_{h+1}(s, a) \right| + \left| (\overline{\mathbb{P}}_h - \widehat{\mathbb{P}}) V_{h+1}(s, a) \right|.$$

Lemma B.7 and Lemma B.8, which we now present, serve exactly this purpose.

Lemma B.7. *For any value function V in the class \mathcal{V} , for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$, it holds that:*

$$\left| (\overline{\mathbb{P}}_h - \mathbb{P}_h) V_{h+1}(s, a) \right| \leq \min \left\{ H\sqrt{d} \|\phi(s, a)\|_{\Lambda_h^{-1}}, 2H \right\}.$$

Proof. We have:

$$\begin{aligned} (\overline{\mathbb{P}}_h - \mathbb{P}_h) V_{h+1}(s, a) &= \phi(s, a)^\top \Lambda_h^{-1} \sum_{k=1}^{\tau} \phi(s_h^k, a_h^k) \mathbb{P}_h V_{h+1}(s_h^k, a_h^k) - \mathbb{P}_h V_{h+1}(s, a) \\ &\stackrel{(1)}{=} \phi(s, a)^\top \Lambda_h^{-1} \sum_{k=1}^{\tau} \phi(s_h^k, a_h^k) \mathbb{P}_h V_{h+1}(s_h^k, a_h^k) - \phi(s, a)^\top \tilde{w}_h \\ &= \phi(s, a)^\top \Lambda_h^{-1} \sum_{k=1}^{\tau} \phi(s_h^k, a_h^k) \mathbb{P}_h V_{h+1}(s_h^k, a_h^k) - \phi(s, a)^\top \Lambda_h^{-1} \Lambda_h \tilde{w}_h \\ &= \phi(s, a)^\top \Lambda_h^{-1} \left[\sum_{k=1}^{\tau} \phi(s_h^k, a_h^k) \mathbb{P}_h V_{h+1}(s_h^k, a_h^k) - \Lambda_h \tilde{w}_h \right] \\ &\stackrel{(2)}{=} \phi(s, a)^\top \Lambda_h^{-1} \left[\sum_{k=1}^{\tau} \phi(s_h^k, a_h^k) \mathbb{P}_h V_{h+1}(s_h^k, a_h^k) \right. \\ &\quad \left. - I \tilde{w}_h - \sum_{k=1}^{\tau} \phi(s_h^k, a_h^k) \phi(s_h^k, a_h^k)^\top \tilde{w}_h \right] \\ &\stackrel{(3)}{=} \phi(s, a)^\top \Lambda_h^{-1} \left[\sum_{k=1}^{\tau} \phi(s_h^k, a_h^k) \mathbb{P}_h V_{h+1}(s_h^k, a_h^k) \right. \\ &\quad \left. - \sum_{k=1}^{\tau} \phi(s_h^k, a_h^k) \mathbb{P}_h V_{h+1}(s_h^k, a_h^k) - \tilde{w}_h \right] \\ &= -\phi(s, a)^\top \Lambda_h^{-1} \tilde{w}_h, \end{aligned}$$

where at (1) we have defined vector $\tilde{w}_h := \int_{\mathcal{S}} V_{h+1}(s') d\mu_h(s')$, at (2) we have used the definition of Λ_h , and at (3) we have recognized that $\phi(s_h^k, a_h^k)^\top \tilde{w}_h = \mathbb{P}_h V_{h+1}(s_h^k, a_h^k)$.

By taking the absolute value, we can write:

$$\begin{aligned} \left| (\overline{\mathbb{P}}_h - \mathbb{P}_h) V_{h+1}(s, a) \right| &= \left| \phi(s, a)^\top \Lambda_h^{-1} \tilde{w}_h \right| \\ &\stackrel{(4)}{\leq} \|\tilde{w}_h\|_{\Lambda_h^{-1}} \|\phi(s, a)\|_{\Lambda_h^{-1}} \\ &\stackrel{(5)}{\leq} \|\tilde{w}_h\|_2 \|\phi(s, a)\|_{\Lambda_h^{-1}} \\ &\stackrel{(6)}{\leq} H\sqrt{d} \|\phi(s, a)\|_{\Lambda_h^{-1}}, \end{aligned}$$

where at (4) we have applied Cauchy-Schwarz's inequality, at (5) we have bounded the quadratic form with the 2-norm and the largest eigenvector of the matrix, i.e., $\|\tilde{w}_h\|_{\Lambda_h^{-1}} = \sqrt{\tilde{w}_h^\top \Lambda_h^{-1} \tilde{w}_h} \leq \sqrt{\sigma} \|\tilde{w}_h\|_2$, where σ is the largest eigenvalue of matrix Λ_h^{-1} , and then we have upper bounded $\sigma \leq 1$, since 1 is the smallest eigenvalue of invertible matrix Λ_h (see [23]); finally, at (6) we have used the fact that $|V_{h+1}(\cdot)| \leq H$, and so that $\|\tilde{w}_h\|_2 = \|\int_{\mathcal{S}} V_{h+1}(s') d\mu_h(s')\|_2 \leq H \|\mu_h(\mathcal{S})\|_2 \leq H\sqrt{d}$.

The result follows by noticing that the quantity to bound cannot be larger than $2H$. \square

Lemma B.8. *Let $\delta \in (0, 1)$. For any value function V in the class \mathcal{V} , for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$, with probability at least $1 - \delta/2$, it holds that:*

$$\left| \left(\widehat{\mathbb{P}}_h - \overline{\mathbb{P}}_h \right) V_{h+1}(s, a) \right| \leq \min \left\{ cH \sqrt{d \log(1 + \tau) + \log \frac{H}{\delta}} \|\phi(s, a)\|_{\Lambda_h^{-1}}, 2H \right\},$$

for some constant c .

Proof. We can write:

$$\begin{aligned} \left| \left(\widehat{\mathbb{P}}_h - \overline{\mathbb{P}}_h \right) V_{h+1}(s, a) \right| &= \left| \phi(s, a)^\top \Lambda_h^{-1} \sum_{k=1}^{\tau} \phi(s_h^k, a_h^k) \left[V_{h+1}(s_{h+1}^k) - \mathbb{P}_h V_{h+1}(s_h^k, a_h^k) \right] \right| \\ &\stackrel{(1)}{\leq} \left\| \sum_{k=1}^{\tau} \phi(s_h^k, a_h^k) \left[V_{h+1}(s_{h+1}^k) - \mathbb{P}_h V_{h+1}(s_h^k, a_h^k) \right] \right\|_{\Lambda_h^{-1}} \|\phi(s, a)\|_{\Lambda_h^{-1}} \\ &\stackrel{(2)}{\leq} \sqrt{4H^2 \left(\frac{d}{2} \log(1 + \tau) + \log \frac{2\mathcal{N}_\epsilon}{\delta} \right) + 8\tau^2 \epsilon^2} \|\phi(s, a)\|_{\Lambda_h^{-1}} \\ &\stackrel{(3)}{\leq} \sqrt{4H^2 \left(\frac{d}{2} \log(1 + \tau) + 2d \log \left(1 + \frac{H\sqrt{d}}{\epsilon} \right) + \log \frac{1}{\delta} \right) + 8\tau^2 \epsilon^2} \|\phi(s, a)\|_{\Lambda_h^{-1}} \\ &\stackrel{(4)}{\leq} \sqrt{4H^2 \left(\frac{d}{2} \log(1 + \tau) + 2d \log(1 + 4\tau) + \log \frac{1}{\delta} \right) + 8H^2 d} \|\phi(s, a)\|_{\Lambda_h^{-1}} \\ &\stackrel{(5)}{\leq} cH \sqrt{d \log(1 + \tau) + \log \frac{1}{\delta}} \|\phi(s, a)\|_{\Lambda_h^{-1}}, \end{aligned}$$

where at (1) we have applied Cauchy-Schwarz's inequality, at (2) we have applied Lemma B.13, at (3) we have upper bounded \mathcal{N}_ϵ using Lemma B.12, at (4), similarly to [62], unlike [23], we see that no union bound is needed (because there is no dependence on Λ), thus by choosing $\epsilon = H\sqrt{d}/\tau$, we get the passage. Passage (5) follows for some constant c .

The result follows by a union bound over $h \in \llbracket H \rrbracket$, and by noticing that the quantity to bound cannot be larger than $2H$. \square

We are now ready to upper bound the Hausdorff distance using the two lemmas just presented. Recall that we work with unbounded rewards (parameters θ), and that the definition of inner distance d is provided in Equation (1).

Lemma B.9. *With probability at least $1 - \delta/2$, the Hausdorff distance between the true feasible set \mathcal{R}_{p, π^E} and its estimate $\widehat{\mathcal{R}}$ returned by Algorithm 1 can be upper bounded by:*

$$\mathcal{H}(\mathcal{R}_{p, \pi^E}, \widehat{\mathcal{R}}) \leq 4J^*(u; p),$$

where $u_h(s, a) := \min\{\beta \|\phi(s, a)\|_{\Lambda_h^{-1}}, H\}$ for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$, and $\beta := cH \sqrt{d \log(1 + \tau) + \log(H/\delta)}$ for some absolute constant $c > 0$.

Proof. Let us begin to bound the first branch of the Hausdorff distance.

$$\begin{aligned} \sup_{r \in \mathcal{R}_{p, \pi^E}} \inf_{\widehat{r} \in \widehat{\mathcal{R}}} d(r, \widehat{r}) &= \sup_{r \in \mathcal{R}_{p, \pi^E}} \inf_{\widehat{r} \in \widehat{\mathcal{R}}} \frac{1}{M_{r, \widehat{r}}} \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s, a) \sim d_h^{p, \pi}(\cdot, \cdot)} |r_h(s, a) - \widehat{r}_h(s, a)| \\ &\stackrel{(1)}{=} \sup_{r \in \mathcal{R}_{p, \pi^E}} \inf_{\widehat{r} \in \widehat{\mathcal{R}}} \frac{1}{M_{r, \widehat{r}}} \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s, a) \sim d_h^{p, \pi}(\cdot, \cdot)} \left| Q_h^*(s, a; p, r) - \mathbb{P}_h V_{h+1}^*(s, a; p, r) \right. \\ &\quad \left. - Q_h^*(s, a; \widehat{p}, \widehat{r}) + \widehat{\mathbb{P}}_h V_{h+1}^*(s, a; \widehat{p}, \widehat{r}) \right| \\ &\stackrel{(2)}{\leq} \sup_{r \in \mathcal{R}_{p, \pi^E}} \frac{1}{M_{r, \widehat{r}}} \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s, a) \sim d_h^{p, \pi}(\cdot, \cdot)} \left| Q_h^*(s, a; p, r) - \mathbb{P}_h V_{h+1}^*(s, a; p, r) \right. \\ &\quad \left. - Q_h^*(s, a; \widehat{p}, \widehat{r}) + \widehat{\mathbb{P}}_h V_{h+1}^*(s, a; \widehat{p}, \widehat{r}) \right|, \end{aligned}$$

$$\begin{aligned}
&\stackrel{(3)}{=} \sup_{r \in \mathcal{R}_{p, \pi^E}} \frac{1}{M_{r, \tilde{r}}} \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s, a) \sim d_h^{p, \pi}(\cdot, \cdot)} |Q_h^*(s, a; p, r) - \mathbb{P}_h V_{h+1}^*(s, a; p, r) \\
&\quad - Q_h^*(s, a; p, r) + \widehat{\mathbb{P}}_h V_{h+1}^*(s, a; p, r)| \\
&= \sup_{r \in \mathcal{R}_{p, \pi^E}} \frac{1}{M_{r, \tilde{r}}} \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s, a) \sim d_h^{p, \pi}(\cdot, \cdot)} |(\widehat{\mathbb{P}}_h - \mathbb{P}_h) V_{h+1}^*(s, a; p, r)| \\
&\stackrel{(4)}{\leq} \sup_{r \in \mathcal{R}_{p, \pi^E}} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s, a) \sim d_h^{p, \pi}(\cdot, \cdot)} \left| (\widehat{\mathbb{P}}_h - \mathbb{P}_h) \frac{V_{h+1}^*(s, a; p, r)}{\max\{1, \max_h \|\theta_h\|_2 / \sqrt{d}\}} \right| \\
&\stackrel{(5)}{=} \sup_{r \in \mathcal{R}_{p, \pi^E}} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s, a) \sim d_h^{p, \pi}(\cdot, \cdot)} \left| (\widehat{\mathbb{P}}_h - \mathbb{P}_h) V_{h+1}^*(s, a; p, \frac{r}{K}) \right| \\
&\stackrel{(6)}{\leq} \sup_{V \in \mathcal{V}} \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s, a) \sim d_h^{p, \pi}(\cdot, \cdot)} |(\widehat{\mathbb{P}}_h - \mathbb{P}_h) V_{h+1}(s, a)|,
\end{aligned}$$

where at (1) we have simply applied the Bellman optimality equation twice w.r.t. the reward function, at (2) we have upper bounded the infimum over the second set of rewards $\widehat{\mathcal{R}}$ with the specific choice of reward $\tilde{r} \in \widehat{\mathcal{R}}$ provided by Lemma B.11, at (3) we use the property of \tilde{r} described in Lemma B.11, at (4) we bring term $1/M_{r, \tilde{r}}$ inside, and then we upper bound it by: $1/M_{r, \tilde{r}} := 1/\max\{\sqrt{d}, \max_h \|\theta_h\|_2, \max_h \|\tilde{\theta}_h\|_2\}/\sqrt{d} \leq 1/\max\{1, \max_h \|\theta_h\|_2/\sqrt{d}\}$, i.e., by simply removing one of the terms inside the maximum operator at denominator. At (5) we define $K := \max\{1, \max_h \|\theta_h\|_2/\sqrt{d}\}$, and, since the value function is linear in the reward, we apply K directly to the reward. At (6) we realize that the possible optimal value functions that can be constructed in p using rewards in \mathcal{R}_{p, π^E} normalized by K are a subset of the value functions in class \mathcal{V} , i.e., of all the possible optimal value functions with parameters $\|w_h\|_2 \leq 2H\sqrt{d}$. This is not trivial since we are working with *unbounded* rewards r , and thus their parameters $\{\theta_h\}_h$ can be any. The normalization by K permits this in the following manner. For any $h \in \llbracket H \rrbracket$, we have $r_h(\cdot, \cdot)/K = \langle \phi(\cdot, \cdot), \theta_h/K \rangle = \langle \phi(\cdot, \cdot), \theta_h/\max\{1, \max_{h'} \|\theta_{h'}\|_2/\sqrt{d}\} \rangle$. Therefore, if $\max_{h'} \|\theta_{h'}\|_2 > \sqrt{d}$, then the normalization makes sure that $\max_{h'} \|\theta_{h'}\|_2 = \sqrt{d}$, while if $\max_{h'} \|\theta_{h'}\|_2 \leq \sqrt{d}$, then the normalization is by 1 and it has no effect. In this way, we see that value functions $V_{h+1}^*(s, a; p, \frac{r}{K})$ can be created by a simple r' with parameters $\{\theta'_{h'}\}_h$ with 2-norms bounded by \sqrt{d} . This guarantees that, since by hypothesis of Linear MDPs $\|\phi(\cdot, \cdot)\|_2 \leq 1$, the value function never exceeds H , and that the norm of the Q -function parameters $\{w_h^\pi\}_h$ for any policy π can be bounded as: $\|w_h^\pi\|_2 \leq \|\theta_h/K\|_2 + \|\int_{\mathcal{S}} V_{h+1}^\pi(s') d\mu_h(s')\|_2 \leq \sqrt{d} + H\|\mu_h(\mathcal{S})\|_2 \leq \sqrt{d} + H\sqrt{d} \leq 2H\sqrt{d}$ (similarly to Lemma B.1 of [23]). It should be remarked that class \mathcal{V} is more general than the actual set of optimal value functions that can be obtained using $r \in \mathcal{R}_{p, \pi^E}$ in p , since such rewards induce optimal value functions for which the optimal action in \mathcal{S}^{p, π^E} is always the expert's action/s $\pi^E(s)$.

Notice that the same derivation can be carried out also for the other branch of the Hausdorff distance, ending up with the same expression. Therefore, the last line is an upper bound to the Hausdorff distance:

$$\begin{aligned}
\mathcal{H}_d(\mathcal{R}_{p, \pi^E}, \widehat{\mathcal{R}}) &\leq \sup_{V \in \mathcal{V}} \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s, a) \sim d_h^{p, \pi}(\cdot, \cdot)} |(\widehat{\mathbb{P}}_h - \mathbb{P}_h) V_{h+1}(s, a)| \\
&\stackrel{(7)}{=} \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s, a) \sim d_h^{p, \pi}(\cdot, \cdot)} \sup_{V \in \mathcal{V}} |(\widehat{\mathbb{P}}_h - \mathbb{P}_h) V_{h+1}(s, a)| \\
&= \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s, a) \sim d_h^{p, \pi}(\cdot, \cdot)} \sup_{V \in \mathcal{V}} |(\widehat{\mathbb{P}}_h - \mathbb{P}_h) V_{h+1}(s, a) \pm \overline{\mathbb{P}}_h V_{h+1}(s, a)| \\
&\stackrel{(8)}{\leq} \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s, a) \sim d_h^{p, \pi}(\cdot, \cdot)} \sup_{V \in \mathcal{V}} |(\overline{\mathbb{P}}_h - \mathbb{P}_h) V_{h+1}(s, a)| + |(\widehat{\mathbb{P}}_h - \overline{\mathbb{P}}_h) V_{h+1}(s, a)| \\
&\stackrel{(9)}{\leq} \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s, a) \sim d_h^{p, \pi}(\cdot, \cdot)} \min \left\{ c_1 H \sqrt{d \log(1 + \tau) + \log \frac{H}{\delta}} \|\phi(s, a)\|_{\Lambda_h^{-1}}, 4H \right\}
\end{aligned}$$

$$\begin{aligned}
&\leq 4 \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s,a) \sim d_h^{p,\pi}(\cdot, \cdot)} \min \left\{ \underbrace{c_2 H \sqrt{d \log(1 + \tau) + \log \frac{H}{\delta}}}_{=:\beta} \|\phi(s, a)\|_{\Lambda_h^{-1}}, H \right\} \\
&= 4 \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s,a) \sim d_h^{p,\pi}(\cdot, \cdot)} \underbrace{\min \left\{ \beta \|\phi(s, a)\|_{\Lambda_h^{-1}}, H \right\}}_{=:u_h(s,a)} \\
&= 4 \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s,a) \sim d_h^{p,\pi}(\cdot, \cdot)} u_h(s, a) \\
&= 4J^*(u; p),
\end{aligned}$$

where at (7) we have noticed that class \mathcal{V} contains the cartesian product of H sets, one for each stage, and therefore the supremum can be brought inside the summation, at (8) we have applied triangle inequality, at (9) we have applied Lemma B.7 and Lemma B.8 and used some absolute constants $c_1, c_2 > 0$, and also the fact that for any numbers x, y, w, z , we have $\min\{x, y\} + \min\{w, z\} \leq \min\{x + w, y + z\}$. □

To conclude the proof of the main theorem, we simply have to observe that any RFE algorithm provides a bound to $J^*(u'; p)$ for some u' similar to u . Depending on the RFE algorithm instantiated as sub-routine, the sample complexity of Algorithm 1 varies.

Theorem 3.3. *Assume that π^E (along with its support S^{p,π^E}) is known. Then, for any $\epsilon, \delta \in (0, 1)$, Algorithm 1 is (ϵ, δ) -PAC for IRL with a number of episodes τ upper bounded by:*

$$\tau \leq \tilde{\mathcal{O}}\left(\frac{H^5 d}{\epsilon^2} \left(d + \log \frac{1}{\delta}\right)\right).$$

Proof. To get the result, we instantiate Algorithm 1 of [62] as RFE sub-routine. Simply, observe that [62] sets β' so that $\beta' \geq \tilde{\beta} := c' H \sqrt{d \log(1 + dH\tau) + \log(H/\delta)} \geq \beta$. By Lemma B.9, we know that:

$$\begin{aligned}
\mathcal{H}(\mathcal{R}_{p,\pi^E}, \hat{\mathcal{R}}) &\leq 4 \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s,a) \sim d_h^{p,\pi}(\cdot, \cdot)} \min \left\{ \beta \|\phi(s, a)\|_{\Lambda_h^{-1}}, H \right\} \\
&\leq 2c_1 \beta' \sum_{h \in \llbracket H \rrbracket} \sup_{\pi \in \Pi} \mathbb{E}_{(s,a) \sim d_h^{p,\pi}(\cdot, \cdot)} \|\phi(s, a)\|_{\Lambda_h^{-1}},
\end{aligned}$$

for some absolute constant $c_1 > 0$. It should be remarked that the quantity in the last line is, modulo c_1 , the quantity that [62] bound in the proof of their Theorem 1 using their algorithm. Specifically, by taking:

$$\tau \leq \tilde{\mathcal{O}}\left(\frac{H^5 d}{\epsilon^2} \left(d + \log \frac{1}{\delta}\right) + \frac{H^6 d^{9/2}}{\epsilon} \log^4 \frac{1}{\delta}\right),$$

and a union bound over the two events that hold w.p. $1 - \delta/2$, and re-setting $\epsilon \leftarrow c_1 \epsilon$, we get the result. □

Notice that if we run Algorithm 1 of [63] for exploration instead of Algorithm 1 of [62], we obtain:

Theorem B.10. *If we use Algorithm 1 of [63] at Line 1 of Algorithm 1, then for any $\epsilon, \delta \in (0, 1)$, such algorithm is (ϵ, δ) -PAC for IRL with a number of episodes τ upper bounded by:*

$$\tau \leq \tilde{\mathcal{O}}\left(\frac{H^6 d^3}{\epsilon^2} \log \frac{1}{\delta}\right).$$

Proof. By Lemma B.9, we know that:

$$\begin{aligned}
\mathcal{H}(\mathcal{R}_{p,\pi^E}, \hat{\mathcal{R}}) &\leq 4J^*(u; p) \\
&= 4 \sup_{\pi \in \Pi} \sum_{h \in \llbracket H \rrbracket} \mathbb{E}_{(s,a) \sim d_h^{p,\pi}(\cdot, \cdot)} \min \left\{ \beta \|\phi(s, a)\|_{\Lambda_h^{-1}}, H \right\},
\end{aligned}$$

for $\beta := cH\sqrt{d\log(1+\tau) + \log(H/\delta)}$. Now, let us define, similarly to Appendix A of [63], the quantities $u'_h(s, a) := \min\{\beta'\|\phi(s, a)\|_{\Lambda_h^{-1}}, H\}$ for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$, and $\beta' := c'dH\sqrt{\log(dH/\delta)/\epsilon}$ for some absolute constant $c' > 0$. In addition, set the number of exploration episodes τ to $\tau = c''d^3H^6 \log(dH\delta^{-1}\epsilon^{-1})/\epsilon^2$, and notice that, for appropriate choices of c', c'' , it holds that: $\beta' \geq c'dH\sqrt{\log(dH\tau/\delta)} \geq \beta := cH\sqrt{d\log(1+\tau) + \log(1/\delta)}$. This entails that $u'_h(s, a) \geq u_h(s, a)$ at all s, a, h , and so:

$$\begin{aligned} \mathcal{H}(\mathcal{R}_{p, \pi^E}, \widehat{\mathcal{R}}) &\leq cJ^*(u'; p) \\ &= cHJ^*(u'/H; p) \\ &\stackrel{(1)}{\leq} c_1H\sqrt{\frac{d^3H^4 \log \frac{d\tau H}{\delta}}{\tau}} \\ &\stackrel{(2)}{\leq} c_2\epsilon, \end{aligned}$$

where at (1) we have applied Lemma 3.2 of [63] (reported in Lemma B.14 for simplicity) with some new constant $c_1 > 0$, and at (2) we have simply replaced τ with its value defined in Algorithm 1 of [63].

The result follows by union bound between the two events that hold w.p. $1 - \delta/2$ to get $1 - \delta$, and by noticing that c_2 is a constant, thus setting $\epsilon \leftarrow c_2\epsilon$ provides the result. \square

Lemma B.11. *Let \mathcal{R}_{p, π^E} be the feasible set of policy π^E w.r.t. transition models p , and let $\widehat{\mathcal{R}}$ be its estimate constructed as in Algorithm 1 using the true $\pi^E, \mathcal{S}^{p, \pi^E}$ (or sets \mathcal{Z}) and some \widehat{p} . For any reward $r \in \mathcal{R}_{p, \pi^E}$, the reward \widehat{r} such that, for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$:*

$$\widehat{r}_h(s, a) = r_h(s, a) + \int_{s' \in \mathcal{S}} p_h(s'|s, a)V_{h+1}^*(s'; p, r) - \int_{s' \in \mathcal{S}} \widehat{p}_h(s'|s, a)V_{h+1}^*(s'; \widehat{p}, \widehat{r}),$$

belongs to $\widehat{\mathcal{R}}$. Moreover, observe that: $Q_h^(s, a; p, r) = Q_h^*(s, a; \widehat{p}, \widehat{r})$ at all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$. In addition, for any reward $\widehat{r} \in \widehat{\mathcal{R}}$, it is possible to construct a reward r in analogous manner so that $r \in \mathcal{R}_{p, \pi^E}$, and such that $Q_h^*(s, a; p, r) = Q_h^*(s, a; \widehat{p}, \widehat{r})$ at all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$.*

Proof. First, we consider the case when \mathcal{S} is finite. By rearranging the terms in the definition of \widehat{r} , we see that, for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$:

$$\widehat{r}_h(s, a) + \sum_{s' \in \mathcal{S}} \widehat{p}_h(s'|s, a)V_{h+1}^*(s'; \widehat{p}, \widehat{r}) = r_h(s, a) + \sum_{s' \in \mathcal{S}} p_h(s'|s, a)V_{h+1}^*(s'; p, r),$$

which, by the Bellman optimality equation, entails that $Q_h^*(s, a; p, r) = Q_h^*(s, a; \widehat{p}, \widehat{r})$.

We recall that $\widehat{\mathcal{R}}$ is defined as:

$$\widehat{\mathcal{R}} = \{\widehat{r} \in \mathfrak{R} \mid \forall (s, h) \in \mathcal{S}^{p, \pi^E}, \forall a \in \mathcal{A} : \mathbb{E}_{a' \sim \pi_h^E(\cdot|s)} Q_h^*(s, a'; \widehat{p}, \widehat{r}) \geq Q_h^*(s, a; \widehat{p}, \widehat{r})\},$$

while thanks to Lemma B.4, the feasible set \mathcal{R}_{p, π^E} can be written as:

$$\mathcal{R}_{p, \pi^E} = \{r \in \mathfrak{R} \mid \forall (s, h) \in \mathcal{S}^{p, \pi^E}, \forall a \in \mathcal{A} : \mathbb{E}_{a' \sim \pi_h^E(\cdot|s)} Q_h^*(s, a'; p, r) \geq Q_h^*(s, a; p, r)\}.$$

It is clear that, if $Q_h^*(s, a; p, r) = Q_h^*(s, a; \widehat{p}, \widehat{r})$ for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$, then since $r \in \mathcal{R}_{p, \pi^E}$ we necessarily have $\widehat{r} \in \widehat{\mathcal{R}}$.

The proof of the opposite case is completely analogous.

In the case with infinite \mathcal{S} , notice that both the feasible set \mathcal{R}_{p, π^E} in Lemma B.4 and the definition of $\widehat{\mathcal{R}}$ in Algorithm 1 make use of the same sets \mathcal{Z} . Thus, we simply make the choice of reward with same \mathcal{Z} and proceed like in the finite case. \square

Lemma B.12 (Covering Number of Class \mathcal{V}). *Let \mathcal{V} be defined as in Equation (2), and define distance dist in \mathcal{V} as $\text{dist}(V, V') := \sup_{s \in \mathcal{S}} |V(s) - V'(s)|$. Then, the ϵ -covering number $|\mathcal{N}(\epsilon; \mathcal{V}, \text{dist})|$ of*

set \mathcal{V} with distance dist can be bounded as:

$$\log |\mathcal{N}(\epsilon; \mathcal{V}, \text{dist})| \leq d \log \left(1 + \frac{4H\sqrt{d}}{\epsilon} \right).$$

Proof. The proof follows that of Lemma D.6 of [23], but is simpler because of the different form of \mathcal{V} .

For any $V_1, V_2 \in \mathcal{V}$ parametrized by w_1, w_2 , we write:

$$\begin{aligned} \text{dist}(V_1, V_2) &= \sup_{s \in \mathcal{S}} \left| \max_{a \in \mathcal{A}} \langle \phi(s, a), w_1 \rangle - \max_{a \in \mathcal{A}} \langle \phi(s, a), w_2 \rangle \right| \\ &\stackrel{(1)}{\leq} \max_{(s, a) \in \mathcal{S} \times \mathcal{A}} \left| \phi(s, a)^\top (w_1 - w_2) \right| \\ &\stackrel{(2)}{\leq} \sup_{\phi: \|\phi\|_2 \leq 1} \left| \phi^\top (w_1 - w_2) \right| \\ &\stackrel{(3)}{=} \|w_1 - w_2\|_2, \end{aligned}$$

where at (1) we have used the common bound that the absolute difference of maxima is upper bounded by the maximum of the absolute difference of the two functions, at (2) we have used the fact that the feature map is always bounded by 1 in 2-norm, and at (3) we have recognized the dual norm of the 2-norm, i.e., itself.

If we construct an ϵ -cover of $\mathcal{W} := \{w \in \mathbb{R}^d \mid \|w\|_2 \leq 2H\sqrt{d}\}$ w.r.t. the 2-norm, we get a covering number bounded by $|\mathcal{N}(\epsilon; \mathcal{W}, \|\cdot\|_2)| \leq (1 + 4H\sqrt{d}/\epsilon)^d$. Clearly, this value upper bounds the covering number of class \mathcal{V} and the result follows. \square

Lemma B.13 (Lemma D.4 of [23]). *Let $\{s_k\}_{k=1}^\infty$ be a stochastic process on state space \mathcal{S} with corresponding filtration $\{\mathcal{F}_k\}_{k=0}^\infty$. Let $\{\phi_k\}_{k=0}^\infty$ be an \mathbb{R}^d -valued stochastic process where $\phi_k \in \mathcal{F}_{k-1}$, and $\|\phi_k\|_2 \leq 1$. Let $\Lambda_\tau = I + \sum_{k=1}^\tau \phi_k \phi_k^\top$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $\tau \geq 0$, and any $V \in \mathcal{V}$ so that $\sup_{s \in \mathcal{S}} |V(s)| \leq H$, we have:*

$$\left\| \sum_{k=1}^\tau \phi_k \left(V(s_k) - \mathbb{E} [V(s_k) | \mathcal{F}_{k-1}] \right) \right\|_{\Lambda_\tau^{-1}} \leq 4H^2 \left[\frac{d}{2} \log(1 + \tau) + \log \frac{\mathcal{N}_\epsilon}{\delta} \right] + 8\tau^2 \epsilon^2,$$

where \mathcal{N}_ϵ is the ϵ -covering number of \mathcal{V} with respect to the distance $\text{dist}(V, V') := \sup_{s \in \mathcal{S}} |V(s) - V'(s)|$.

Lemma B.14 (Lemma 3.2 of [63]). *With probability $1 - \delta/2$, for the function u' defined as $u'_h(s, a) := \min \{\beta' \|\phi(s, a)\|_{\Lambda_h^{-1}}, H\}$, with $\beta' := c'dH\sqrt{\log(dH\delta^{-1}\epsilon^{-1})}$, we have:*

$$J^*(u'/H) \leq c \sqrt{\frac{d^3 H^4 \log \frac{d\tau H}{\delta}}{\tau}},$$

for some absolute constant $c > 0$.

C Additional Insights on Compatibility

In this appendix, we collect and describe additional insights to the notion of *rewards compatibility* introduced in Section 4. The appendix is organized in the following manner: Appendix C.1 provides a visual explanation to the notion of rewards compatibility, in Appendix C.2 we analyse a multiplicative alternative to the definition of rewards compatibility, and Appendix C.3 discusses the conditions under which a learned reward can be used for “forward” RL, by comparing rewards with small (non)compatibility with rewards learned in previous works.

C.1 A Visual Explanation for Rewards Compatibility

In this appendix, we aim to provide a visual intuition to the notion of rewards compatibility. For this reason, the reader should keep in mind Figure 3.

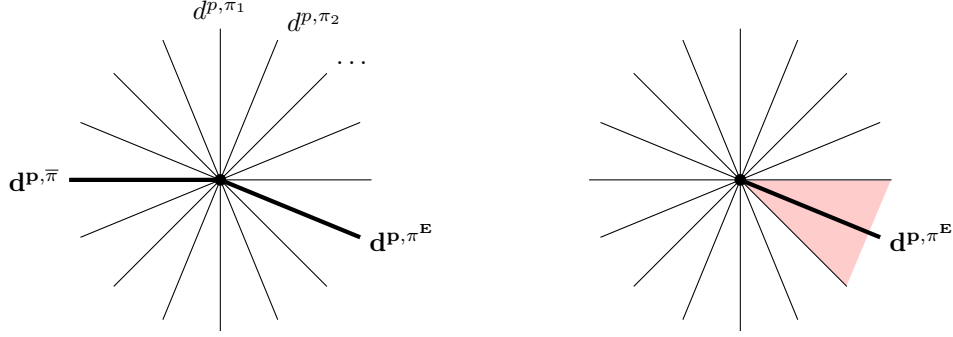


Figure 3: In this figure, the point at the center represents the initial state $s_0 = d_0$ of the environment \mathcal{M} , and each ray starting from it represents the occupancy measure $d^{p, \pi}$ of some policy π . The figure aims to provide the intuition that policies with rays close to each other induce similar visit distributions (e.g., both point towards the same direction in some grid-world), and policies with rays far away from each other point toward very different directions (i.e., they have different occupancy measures). The red area in the right denotes the set of directions (occupancy measures $d^{p, \pi}$ for some π) that are close in $\|\cdot\|_1$ norm to the direction of the expert d^{p, π^E} .

As explained in Section 4, even in the limit of infinite samples, i.e., even if we know $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p) \cup \{\pi^E\}$ exactly, and even if we assume that the expert is exactly optimal, i.e., $J^*(r^E; p) - J^{\pi^E}(r^E; p) = 0$ (where r^E is the true reward optimized by the expert), then we still do not have idea of how other policies perform. Expert demonstrations only provide information about the performance of a *single* policy, π^E , w.r.t. to the reference $J^*(r^E; p)$ under the unknown r^E , i.e., demonstrations say that π^E in r^E performs as good as $J^*(r^E; p)$. But what about other policies? Demonstrations provide no information.

To see this, consider Figure 3, in which each line exemplifies the visitation distribution induced by some policy $\pi \in \Pi$, and the point in the middle represents the starting state $s_0 = d_0$. Intuitively, observing d^{p, π^E} along with knowing that $J^{\pi^E}(r^E; p)$ is good (i.e., because of expert demonstrations), does not tell us anything about the distribution $d^{p, \bar{\pi}}$ induced by some other policy $\bar{\pi}$ potentially arbitrarily different from d^{p, π^E} . Indeed, it might be the case that $J^{\bar{\pi}}(r^E; p)$ is acceptable, or that it is as good as $J^{\pi^E}(r^E; p)$, or that it is very bad. We cannot know from demonstrations only.

For this reason, if we consider the set of rewards with 0-(non)compatibility, i.e., the feasible reward set, we notice that it contains the rewards r that make $\bar{\pi}$ optimal $J^{\bar{\pi}}(r; p) = J^*(r; p)$, but also the rewards r' that make $\bar{\pi}$ nearly optimal $J^{\bar{\pi}}(r'; p) \approx J^*(r'; p)$, and also the rewards r'' that make $\bar{\pi}$ a very bad-performing policy $J^{\bar{\pi}}(r''; p) \ll J^*(r''; p)$. Indeed, as long as both r, r', r'' make the direction pointed by d^{p, π^E} in Figure 3 a good direction, then they are in accordance with the constraint imposed by the demonstrations. The additional Degrees of Freedom (DoF) provided by policies beyond π^E (e.g., $\bar{\pi}, \dots$) permit the ill-posedness of IRL.

We said that expert demonstrations provide information just about the performance of a single policy, π^E . However, to be precise, in the context of IRL, this is not correct. Indeed, differently from the mere learning from demonstrations setting, in which we just assume that π^E is a very good-performing policy, in IRL we assume that the underlying problem is an MDP, i.e., that the expert agent is optimizing a reward function r^E .¹⁰ This additional structure (i.e., that the underlying environment is indeed an MDP), makes sure that the performances of various directions $d^{p, \pi}$ in Figure 3 are measured through a dot product with a fixed reward function r , i.e.:

$$J^\pi(r; p) = \sum_{h \in [H]} \langle d_h^{p, \pi}, r_h \rangle.$$

For this reason, we have the guarantee that the directions in the red area surrounding d^{p, π^E} are almost as good as d^{p, π^E} . Indeed, for all policies π such that $\sum_{h \in [H]} \|d_h^{p, \pi} - d_h^{p, \pi^E}\|_1 \leq \epsilon$, i.e., for all policies

¹⁰When this assumption does not hold, we incur in model misspecification [55, 51].

ϵ -close to π^E in 1-norm, we can write:

$$|J^{\pi^E}(r^E; p) - J^\pi(r^E; p)| = \left| \sum_{h \in [H]} \langle d_h^{p, \pi^E} - d_h^{p, \pi}, r_h^E \rangle \right| \leq \sum_{h \in [H]} \|d_h^{p, \pi} - d_h^{p, \pi^E}\|_1 \leq \epsilon.$$

In other words, policies π and π^E have similar performances.

However, it should be remarked that, since we aim to recover the rewards explaining the expert's preferences, then we are guaranteed that policies close in 1-norm perform similarly under any reward function (by definition of 1-norm), and so we do not risk to incur in the error of representing d^{p, π^E} and a direction $d^{p, \pi}$ inside the red area of Figure 3 with very different performances.

C.2 A Multiplicative Compatibility

In Section 4, we have defined an *additive* notion of (non)compatibility, based on the difference of performance between π^E and π^* (the optimal policy). Here, we analyze a *multiplicative* notion of (non)compatibility, based on the ratio of the performances.¹¹

We make the following observation. Any reward $r \in \mathfrak{R}$ induces, in the considered environment p , an ordering in the space of policies Π , based on the performance $J^\pi(r; p)$ of each policy $\pi \in \Pi$. It is easy to notice that for any scaling and translation parameters $\alpha \in \mathbb{R}_{>0}$, $\beta \in \mathbb{R}$, the reward constructed as $r'(\cdot, \cdot) = \alpha r(\cdot, \cdot) + \beta$ induces the same ordering as r in the space of policies.¹²

For this reason, it seems desirable to use a notion of (non)compatibility such that rewards r and $r'(\cdot, \cdot) = \alpha r(\cdot, \cdot) + \beta$ for some α, β , suffer from the same (non)compatibility w.r.t. some expert policy π^E . However, observe that, for the notion of compatibility $\bar{\mathcal{C}}$ in Definition 4.1, we have that, for any $r \in \mathfrak{R}$:

$$\begin{aligned} \bar{\mathcal{C}}_{p, \pi^E}(r + \beta) &= \bar{\mathcal{C}}_{p, \pi^E}(r) \quad \forall \beta \in \mathbb{R}, \\ \bar{\mathcal{C}}_{p, \pi^E}(\alpha r) &= \alpha \bar{\mathcal{C}}_{p, \pi^E}(r) \neq \bar{\mathcal{C}}_{p, \pi^E}(r) \quad \forall \alpha \in \mathbb{R}_{>0}. \end{aligned}$$

Simply put, for the additive notion of (non)compatibility $\bar{\mathcal{C}}$, the scale (α) of a reward matters, and rescaling the reward modifies the (non)compatibility.

To solve this issue, one might introduce a *multiplicative* notion of compatibility \mathcal{F} (defined only for non-negative rewards and setting $\mathcal{F}_{p, \pi^E}(r) = 0$ when the denominator is 0):

$$\mathcal{F}_{p, \pi^E}(r) := \frac{J^{\pi^E}(r; p)}{J^*(r; p)}.$$

Clearly, the larger $\mathcal{F}_{p, \pi^E}(r)$, the closer is the performance of π^E to the optimal performance. Observe that, for this definition, we have:

$$\begin{aligned} \mathcal{F}_{p, \pi^E}(\alpha r) &= \mathcal{F}_{p, \pi^E}(r) \quad \forall \alpha \in \mathbb{R}_{>0} \\ \mathcal{F}_{p, \pi^E}(r + \beta) &\neq \mathcal{F}_{p, \pi^E}(r) \quad \forall \beta \in \mathbb{R}, \end{aligned}$$

i.e., this definition does not care about the scaling α of the reward, but it is sensitive to the actual position β of that reward.

Therefore, both $\bar{\mathcal{C}}$ and \mathcal{F} suffer from some “rescaling” issues. Is it possible to devise a notion of compatibility, i.e., a measure of suboptimality, for a policy, that is independent of both the scale α and position β ? Formally, we are looking for a function (notion of distance) $f : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ such that, for any $J_1, J_2 \in \mathbb{R}$:

$$f(\alpha J_1 + \beta, \alpha J_2 + \beta) = f(J_1, J_2), \quad (3)$$

for all $\alpha \in \mathbb{R}_{>0}, \beta \in \mathbb{R}$. Unfortunately, this is not possible, since it is easy to show that all the functions f of this kind are of the following type:

$$\forall J_1, J_2 \in \mathbb{R} \times \mathbb{R} : f(J_1, J_2) = \begin{cases} K_+ & \text{if } J_1 > J_2 \\ K_0 & \text{if } J_1 = J_2 \\ K_- & \text{if } J_1 < J_2 \end{cases},$$

¹¹E.g., see Theorem 7.2.7 in [45], which is inspired by [43].

¹²Indeed, simply observe that, for any $\pi \in \Pi$: $J^\pi(r'; p) = J^\pi(\alpha r + \beta; p) = \alpha J^\pi(r; p) + \beta$.

for some reals K_+, K_0, K_- . In words, any function f that satisfies Equation (3) is able to express just an ordering between inputs J_1 and J_2 , but not an actual measure of sub-optimality/compatibility.

We conclude by stating that we prefer to use $\bar{\mathcal{C}}$ instead of \mathcal{F} for the following reasons:

- First, most RL literature prefers the additive notion of suboptimality towards the multiplicative one.
- The additive notion of suboptimality is simpler to analyze w.r.t. the multiplicative one.

C.3 When can a learned reward be used for “forward” RL?

In this appendix, we exploit the intuition developed in Appendix C.1 to discuss under which conditions we can exploit demonstrations alone to recover a single reward that *can be used for “forward” RL*, i.e., to recover a single reward r for which we have the guarantee that any ϵ -optimal policy π to r in the true environment p has similar performance in the same environment p under the true reward r^E , that is, policy π is an $f(\epsilon)$ -optimal policy to r^E in p , for some function f .

Applications of IRL range from Apprenticeship Learning (AL), to reward design, to interpretability of expert’s preferences. Concerning AL, it is common to “use” the reward r learned through IRL to optimize our learning agent. But what properties r should satisfy in order to obtain performance guarantees on our learning agent w.r.t. the true (unknown) r^E ? We now list and analyze various plausible requirements.

- First, we might ask that, being π^E optimal w.r.t. r^E , then $\pi^E \in \arg \max_{\pi} J^{\pi}(r)$, i.e., that the expert policy π^E is optimal under the learned reward r . However, this requirement is not satisfactory for the following reason. Reward r might induce more than one optimal policy (e.g., it might induce both $\bar{\pi}, \pi^E$ as optimal), and optimal policies other than π^E (e.g., $\bar{\pi}$) are not guaranteed to perform well under r^E (actually, $\bar{\pi}$ can be any policy in Π). Clearly, this is not satisfactory. Observe that there are rewards in the feasible set \mathcal{R}_{p, π^E} for which multiple policies are optimal (thus, not all the rewards in the feasible set are satisfactory).
- We might additionally ask that π^E is the unique optimal policy of reward r (similarly to what happens in entropy-regularized MDPs [70, 15]). However, this is not satisfactory for the following reason. In practice, it is really difficult (almost impossible) to compute the optimal policy of a given reward. Thus, what is usually done in RL, is to settle for an ϵ -optimal policy. Since any policy can be ϵ -optimal under reward r , then no guarantee we can have for such policy w.r.t. r^E .
- What if we ask that π^E is at least ϵ -optimal under r (i.e., the requirement provided by ϵ -(non)compatible rewards)? Well, this is not satisfactory because optimal policies can be any, and because there might be other ϵ -optimal policies that can perform arbitrarily bad under r^E .

All the three requirements described above on r do not provide guarantees that optimizing the considered reward r provides a policy with satisfactory performance w.r.t. the true r^E . However, as mentioned in Section 4 and in Appendix C.1, expert demonstrations *do not provide any information about the performance of policies other than π^E under r^E* .

Remark C.1. *If we want to be sure that an ϵ -optimal policy π for the learned reward r in p is if $f(\epsilon)$ -optimal for r^E in p (for some function f), then, clearly, we need that all the (at least) ϵ -optimal policies under the learned r have visitation distribution close to that of π^E in 1-norm (see Appendix C.1).*

We stress that many IRL algorithms for AL, like max-margin [2], learn a reward function just as a mere mathematical tool to compute a policy π which is close in 1-norm $\|d^{\pi} - d^{\pi^E}\|_1$ to π^E .

A remark about works on the feasible set. If we look at recent works about the feasible set [38, 31, 68], it might seem that these works are able to provide guarantees between r, r^E under distance d^{all} (see Section 3.1 of [68]), defined as:

$$d^{all}(r, r^E) := \sup_{\pi \in \Pi} |J^{\pi}(r) - J^{\pi}(r^E)|.$$

If $d^{all}(r, r^E)$ is small, then the performance of any policy in r , not just optimal policy or ϵ -optimal policy, is similar also under r^E . In other words, if we use/optimize reward r , then we have the guarantee that the performance of the retrieved policy under r^E is more or less the same as its performance in r . Therefore, clearly, rewards r with small distance to r^E w.r.t. d^{all} can be used for “forward” RL. However, we have the following result:

Proposition C.1. *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$ be a known MDP without reward, and let π^E be a known expert’s policy. Let r^E the true unknown reward optimized by the expert to construct π^E . Then, there does not exist a learning algorithm that receives in input the pair (\mathcal{M}, π^E) and outputs a single reward r such that $d^{all}(r, r^E) \leq \epsilon$ w.p. $1 - \delta$.*

Proof. The proof is trivial. Indeed, since the feasible set \mathcal{R}_{p, π^E} contains an infinite amount of reward functions along with r^E , and the learning algorithm cannot discriminate r^E inside \mathcal{R}_{p, π^E} , then the best it can do is to output an arbitrary reward function $r \in \mathcal{R}_{p, \pi^E}$. However, since \mathcal{R}_{p, π^E} contains, for any reward $r \in \mathcal{R}_{p, \pi^E}$, at least another reward $r' \in \mathcal{R}_{p, \pi^E}$ such that $d^{all}(r, r') = c$ is finite and equal to some positive constant $c > 0$,¹³ then we can simply construct the problem instance with $r^E := r'$ to make the learning algorithm not able to output rewards that can be used for forward learning. \square

Nevertheless, [38, 31, 68] seem to provide sample efficient algorithms w.r.t. d^{all} .¹⁴ By looking at Proposition C.1, we realize that this is clearly a *contradiction*. What is the right interpretation?

The trick is that the algorithms proposed in works [68, 38, 31] are *not* able to output a single reward r which is close to r^E w.r.t. d^{all} , but, for any possible reward $r^E = r^E(V, A)$ parametrized¹⁵ by some value and advantage functions V, A , they are able to output a reward r such that $d^{all}(r, r^E(V, A))$ is small. In other words, it is like if these works assume to know the V, A parametrization of the true reward r^E . Simply put, these works are able to output a reward r that can be used for “forward” RL just under such assumption. Otherwise those algorithms do not provide such guarantee.

Conclusions. To sum up, we conclude that, in general, an arbitrary reward function with small (non)compatibility can *not* be used for “forward” learning (see Proposition C.1), because we cannot know given demonstrations alone whether the performances assigned by such reward to policies other than the expert policy are meaningful. In addition, for the same reason, we realize that also an arbitrary reward with zero (non)compatibility, i.e., an arbitrary reward in the feasible set, can *not* be used for “forward” learning.

C.4 Comparing the (non)compatibility of various rewards

In Section 4, we said that rewards r with smaller values of $\bar{C}_{p, \pi^E}(r)$ are more compatible with π^E in $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$. However, one might provide the following “counter-example”:

Example C.1 (Question by Reviewer KyLX). *Let r^1, r^2 be two rewards such that $r_h^2(s, a) = 2r_h^1(s, a) \geq 0$ for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$. Then, clearly, $\bar{C}_{p, \pi^E}(r^2) = 2\bar{C}_{p, \pi^E}(r^1)$. Therefore, based on Section 4, we say that reward r^1 is more compatible than r^2 w.r.t. π^E in \mathcal{M} . However, since r^2 is just r^1 re-scaled by a constant, the two MDPs $\mathcal{M} \cup \{r^1\}$ and $\mathcal{M} \cup \{r^2\}$ should be “equivalent”, thus, r^1 and r^2 should be, intuitively, equally compatible with π^E .*

However, Example C.1 misleads the correct *interpretation* of the notion of reward function in MDPs, and in particular about the *scale* of the rewards. Let us explain better our point.

The MDP is a model, i.e., a simplified representation of reality, which is commonly applied to 2 different kinds of real-world scenarios: (i) problems in which the agent (learner in RL or expert in IRL) actually receives some kind of scalar feedback from the environment, which can be modelled as a reward function; (ii) problems in which the agent does not receive a feedback from the environment,

¹³This is immediate from the considerations in Appendix C.1.

¹⁴Actually, [38, 31] use different notions of distance, like $d_\infty(r, r') := \|r - r'\|_\infty$. However, we can write $\|r - r'\|_\infty \geq \|r - r'\|_1 / (SAH)$, and by dual norms we have that $d^{all}(r, r') = \sup_{\pi \in \Pi} |\langle d^{p, \pi}, r - r' \rangle| \leq \sup_{\bar{d}: \|\bar{d}\|_\infty \leq 1} |\langle \bar{d}, r - r' \rangle| = \|r - r'\|_1$. Therefore, the guarantees of [38, 31] can be converted too d^{all} guarantees too.

¹⁵While [68] makes this parametrization explicit, [38, 31] keep the parametrization implicit, but everything is analogous.

but its objective, i.e., its structure of preferences among state-action trajectories (which trajectories are better than others), satisfies some axioms that permit to represent it through a scalar reward [52, 9] (this is referred to as the Reward Hypothesis in literature [58]).

There is an enormous difference between scenario (i) and scenario (ii). In (i) the notion of ϵ -optimal policy is well-defined for any fixed $\epsilon > 0$, because the reward function is given and, thus, fixed. Instead, in (ii), the notion of reward function is a mere mathematical artifact used to represent preferences among trajectories, whose existence is guaranteed by a set of assumptions/axioms [52, 9, 58]. As Example C.1 shows, positive affine transformations of the reward do not affect the structure of preferences represented (see [52] or Section 16.2 of [50] or [30]). Therefore, in (ii), the notion of ϵ -optimal policy is not well-defined, because rescaling a reward function r to kr changes the suboptimality of some policy π from ϵ to $k\epsilon$. In other words, for fixed $\epsilon > 0$, any policy can be made ϵ -optimal by simply rescaling a reward r to kr for some small enough $k > 0$.

In IRL, this issue is even more influential because, although we are in setting (i), we have no idea on the scale of the true reward function. For this reason, our solution is to attach to any reward r a notion of compatibility $\bar{C}(r)$ which implicitly contains information about the scale of the reward r . Compatibilities of different rewards (e.g., r^1 and r^2 in Example C.1) cannot be compared unless the rewards have the same scale (e.g., r^1 and r^2 have different scales, thus their compatibilities shall not be compared).

It should be observed that in Appendix C.2 we discuss a notion of compatibility independent of the scale of the reward. However, we show that it suffers from major drawbacks that make the notion of compatibility introduced in the main paper (Definition 4.1) more suitable for the IRL problem.

In conclusion, to settle Example C.1, rewards r^1 and r^2 should not have the same compatibility, because they have different scales, and the notion of compatibility (i.e., suboptimality) is strictly connected to the scale of the reward. To carry out a fair comparison of compatibilities, one should rescale the compatibility of each reward based on the scale of the reward.

D Missing Proofs and Additional Results for Section 5

This appendix is organized as follows. First, we report the full pseudo-code of **CATY-IRL**. Then, we provide the proof of Theorem 5.1 in Appendix D.2.

D.1 Algorithm

In this section, we provide the extended version of **CATY-IRL** containing the explicit conditions under which we shall instantiate one BPI/RFE algorithm instead of another.

Algorithm 2: **CATY-IRL**- exploration

Data: Failure probability $\delta > 0$, target accuracy $\epsilon > 0$, expert demonstrations \mathcal{D}^E , set of rewards to classify \mathcal{R} , problem structure $\iota \in \{\text{tabular, linear rewards, Linear MDP}\}$

```

1 if  $\iota \in \{\text{tabular, linear rewards}\}$  then
2   if  $|\mathcal{R}|$  is a small constant then
3      $\mathcal{D} \leftarrow \{\}$ 
4     for  $r' \in \mathcal{R}$  do
5        $\mathcal{D} \leftarrow \mathcal{D} \cup \text{BPI\_Exploration}(\delta, \epsilon/2, r')$       /* Algorithm BPI-UCBVI [37] */
6     end
7   else
8      $\mathcal{D} \leftarrow \text{RFE\_Exploration}(\delta, \epsilon/2)$                 /* Algorithm RF-Express [37] */
9   end
10 else
11    $\mathcal{D} \leftarrow \text{RFE\_Exploration}(\delta, \epsilon/2)$                 /* Algorithm RFLin [62] */
12 end
13 Return  $\mathcal{D}$ 

```

Algorithm 3: CATY-IRL- classification

Data: Failure probability $\delta > 0$, target accuracy $\epsilon > 0$, expert demonstrations \mathcal{D}^E , classification threshold $\Delta \in \mathbb{R}$, reward to classify $r \in \mathcal{R}$, problem structure $\iota \in \{\text{tabular, linear rewards, Linear MDP}\}$, dataset \mathcal{D}

```
// Estimate the expert's performance  $\hat{J}^E(r)$ :
1 if  $\iota = \text{tabular}$  then
2    $\hat{d}^E \leftarrow$  empirical estimate of  $d^{p, \pi^E}$  from  $\mathcal{D}^E$ 
3    $\hat{J}^E(r) \leftarrow \sum_h \langle \hat{d}_h^E, r_h \rangle$ 
4 else
5    $\hat{\psi}^E \leftarrow$  empirical estimate of  $\psi^{p, \pi^E}$  from  $\mathcal{D}^E$ 
6    $\hat{J}^E(r) \leftarrow \sum_h \langle \hat{\psi}_h^E, r_h \rangle$ 
7 end

// Estimate the optimal performance  $\hat{J}^*(r)$ :
8 if  $\iota \in \{\text{tabular, linear rewards}\}$  then
9   if  $|\mathcal{R}|$  is a small constant then
10     $\hat{J}^*(r) \leftarrow \text{BPI\_Planning}(\mathcal{D}, r)$            /* Algorithm BPI-UCBVI [37] */
11   else
12     $\hat{J}^*(r) \leftarrow \text{RFE\_Planning}(\mathcal{D}, r)$        /* Algorithm RF-Express [37] */
13   end
14 else
15    $\hat{J}^*(r) \leftarrow \text{RFE\_Planning}(\mathcal{D}, r)$        /* Algorithm RFLin [62] */
16 end

// Classify the reward:
20  $\hat{\mathcal{C}}(r) \leftarrow \hat{J}^*(r) - \hat{J}^E(r)$ 
21 class  $\leftarrow$  True if  $\hat{\mathcal{C}}(r) \leq \Delta$  else False
22 return class
```

D.2 Proof of Theorem 5.1

Notice that, according to Definition 4.3, an algorithm is (ϵ, δ) -PAC for IRL if it computes an estimate ϵ -close to the true (non)compatibility w.h.p.. Such definition does not depend on the specific strategy adopted by the algorithm to actually classify the input reward using the computed estimate of (non)compatibility.

Before diving into the proof of Theorem 5.1, we make the following considerations.

In the common tabular MDPs setting without additional structure, we know that the expected utility $J^\pi(r; p)$ of policy π under reward r in environment with dynamics p can be computed as:

$$J^\pi(r; p) = \sum_{h \in \llbracket H \rrbracket} \langle r_h, d_h^{p, \pi} \rangle,$$

where $d_h^{p, \pi}$ is the occupancy measure of policy π in p . It should be remarked that both r_h and $d_h^{p, \pi}$ have SA components for all $h \in \llbracket H \rrbracket$.

In tabular MDPs with linear reward functions and in Linear MDPs, the reward function is linear in some feature map ϕ , i.e.:

$$r_h(\cdot, \cdot) = \langle \phi(\cdot, \cdot), \theta_h \rangle \quad \forall h \in \llbracket H \rrbracket,$$

where $\|\phi(s, a)\|_2 \leq 1$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $\max_h \|\theta_h\|_2 \leq \sqrt{d}$. Using this decomposition, we can rewrite the expected utility $J^\pi(r; p)$ as:

$$J^\pi(r; p) = \sum_{h \in \llbracket H \rrbracket} \langle r_h, d_h^{p, \pi} \rangle$$

$$\begin{aligned}
&= \sum_{h \in \llbracket H \rrbracket} \langle \theta_h^\top \phi, d_h^{p, \pi} \rangle \\
&= \sum_{h \in \llbracket H \rrbracket} \theta_h^\top \mathbb{E}_{(s, a) \sim d_h^{p, \pi}} \phi(s, a) \\
&= \sum_{h \in \llbracket H \rrbracket} \theta_h^\top \psi_h^{p, \pi},
\end{aligned}$$

where we have defined the feature expectations $\{\psi_h^{p, \pi}\}_{h \in \llbracket H \rrbracket}$ as $\psi_h^{p, \pi} := \mathbb{E}_{(s, a) \sim d_h^{p, \pi}} \phi(s, a)$. Observe that vector $\psi_h^{p, \pi}$ has d components instead of the SA components of each $d_h^{p, \pi}$ vector.

Since in our setting the IRL algorithm receives in input the reward function (or its parameter $\theta \in \mathbb{R}^d$), to estimate the expected utility $J^\pi(r; p)$ we must estimate the visit distributions $\{d_h^{p, \pi}\}_h$ or the feature expectations $\{\psi_h^{p, \pi}\}_h$. However, because of the different dimensionalities of such quantities (SA versus d), the estimates might require different amounts of samples.

Theorem 5.1 (Sample Complexity of **CATY-IRL**). *Let $\epsilon, \delta \in (0, 1)$. Then **CATY-IRL** is (ϵ, δ) -PAC for IRL with a sample complexity upper bounded by:*

$$\begin{aligned}
\text{Tabular MDPs:} \quad & \tau^E \leq \tilde{\mathcal{O}}\left(\frac{H^3 SA}{\epsilon^2} \log \frac{1}{\delta}\right), \quad \tau \leq \tilde{\mathcal{O}}\left(\frac{H^3 SA}{\epsilon^2} \left(N + \log \frac{1}{\delta}\right)\right), \\
\text{Tabular MDPs with linear rewards:} \quad & \tau^E \leq \tilde{\mathcal{O}}\left(\frac{H^3 d}{\epsilon^2} \log \frac{1}{\delta}\right), \quad \tau \leq \tilde{\mathcal{O}}\left(\frac{H^3 SA}{\epsilon^2} \left(N + \log \frac{1}{\delta}\right)\right), \\
\text{Linear MDPs:} \quad & \tau^E \leq \tilde{\mathcal{O}}\left(\frac{H^3 d}{\epsilon^2} \log \frac{1}{\delta}\right), \quad \tau \leq \tilde{\mathcal{O}}\left(\frac{H^5 d}{\epsilon^2} \left(d + \log \frac{1}{\delta}\right)\right),
\end{aligned}$$

where $N = 0$ if $|\mathcal{R}| = \Theta(1)$, and $N = S$ otherwise.

Proof. To prove the theorem, we aim to find a bound to the number of samples τ^E such that the estimate $\hat{J}^E(r) \approx J^{\pi^E}(r; p)$ is $\epsilon/2$ -correct with probability at least $1 - \delta/2$. Next, similarly, we aim to bound τ so that $\hat{J}^*(r) \approx J^*(r; p)$ is $\epsilon/2$ -correct with probability at least $1 - \delta/2$. Then, the conclusion follows after performing a union bound and observing that, for any $r \in \mathcal{R}$:

$$\begin{aligned}
\left| \bar{\mathcal{C}}_{p, \pi^E}(r) - \hat{\mathcal{C}}(r) \right| &= \left| \left(J^*(r; p) - J^{\pi^E}(r; p) \right) - \left(\hat{J}^*(r) - \hat{J}^E(r) \right) \right| \\
&\leq \left| J^*(r; p) - \hat{J}^*(r) \right| + \left| J^{\pi^E}(r; p) - \hat{J}^E(r) \right| \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\end{aligned}$$

Estimating $\hat{J}^E(r) \approx J^{\pi^E}(r; p)$

To estimate $J^{\pi^E}(r; p)$, **CATY-IRL** simply computes the empirical estimate of $\{d_h^{p, \pi^E}\}$ in case of tabular MDPs, and the empirical estimate of $\{\psi_h^{p, \pi^E}\}$ in case of tabular MDPs with linear rewards and Linear MDPs. Notice that by empirical estimates we mean:

$$\hat{d}_h^E(s, a) := \frac{\sum_{i \in \llbracket \tau^E \rrbracket} \mathbb{1}\{s_h^i = s \wedge a_h^i = a\}}{\sum_{i \in \llbracket \tau^E \rrbracket} \mathbb{1}\{s_h^i = s\}} \quad \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket,$$

and:

$$\hat{\psi}_h^E := \frac{\sum_{i \in \llbracket \tau^E \rrbracket} \phi(s_h^i, a_h^i)}{\tau^E} \quad \forall h \in \llbracket H \rrbracket.$$

Concerning the estimate of the visit distribution \hat{d}_h^E , we can use the result of Lemma 6 in [53] (we are working with bounded rewards), to obtain that:

$$\sum_{h \in \llbracket H \rrbracket} \|d_h^{p, \pi^E} - \hat{d}_h^E\|_1 \leq \sqrt{\frac{SAH^3 \log \frac{8SAH}{\delta}}{2\tau^E}} \leq \frac{\epsilon}{2}.$$

Solving w.r.t. τ^E we get the bound on τ^E .

In a completely analogous manner, we can bound the feature expectations as:

$$\sum_{h \in [H]} \|\psi_h^{p, \pi^E} - \hat{\psi}_h^E\|_1 \leq \sqrt{\frac{dH^3 \log \frac{8dH}{\delta}}{2\tau^E}} \leq \frac{\epsilon}{2}.$$

Again, solving w.r.t. τ^E we get the bound on τ^E .

Estimating $\hat{J}^*(r) \approx J^*(r; p)$

Let us begin with the case in which \mathcal{R} is large. As explained for instance in Definition 4 of [66], both algorithms RF-Express [37] and RFLin [62] satisfy the *uniform policy evaluation property*, i.e., they guarantee that, for any $\epsilon, \delta \in (0, 1)$, after having explored for $\tau \leq \tilde{\mathcal{O}}\left(\frac{H^3 SA}{\epsilon^2} (S + \log \frac{1}{\delta})\right)$ in case of RF-Express [37], and $\tau \leq \tilde{\mathcal{O}}\left(\frac{H^3 d}{\epsilon^2} (d + \log \frac{1}{\delta})\right)$ for the algorithm in [62] (we omit linear terms in $1/\epsilon$), they compute an estimate $\hat{p} \approx p$ of the true transition model such that:

$$\mathbb{P}\left(\sup_{r \in \mathfrak{R}, \pi \in \Pi} |J^\pi(r; p) - J^\pi(r; \hat{p})| \leq \epsilon\right) \geq 1 - \delta.$$

Clearly, if such property holds, then by computing the performance of the policy $\hat{\pi}$ outputted by the RFE algorithm we are able to obtain an $\epsilon/2$ -correct estimate of $J^*(r; p)$.¹⁶

Concerning the case in which $|\mathcal{R}|$ is a finite small constant, for tabular and tabular with linear rewards MDPs, we can simply use algorithm BPI-UCBVI of [37] as sub-routine, and run it as many times as there are rewards in \mathcal{R} . When $|\mathcal{R}|$ is a small constant, we can proceed with a union bound over \mathcal{R} :

$$\mathbb{P}\left(\sup_{r \in \mathfrak{R}, \pi \in \Pi} |J^\pi(r; p) - J^\pi(r; \hat{p})| \leq \epsilon\right) \geq 1 - \sum_{r \in \mathcal{R}} \mathbb{P}\left(\sup_{\pi \in \Pi} |J^\pi(r; p) - J^\pi(r; \hat{p})| > \epsilon\right) \geq 1 - |\mathcal{R}|\delta.$$

This allows us to formally distinguish between small and large $|\mathcal{R}|$ based on the following inequality:

$$S + \log \frac{1}{\delta} < \log \frac{|\mathcal{R}|}{\delta} \implies S < \log |\mathcal{R}|.$$

□

E Missing Proofs and Additional Results for Section 6.1

This appendix is organized as follows. First, in Appendix E.1, we introduce two problems that share similarities with RFE and IRL, and we characterize the main differences among them. In addition, we enunciate a lower bound to the sample complexity that is common to some of these 4 problems. Next, in Appendix E.2, we provide the missing proofs.

E.1 Four Problems

The 4 problems that we consider here are Reward-Free Exploration (RFE), Inverse Reinforcement Learning (IRL), Matching Performance (MP), and Imitation Learning from Demonstrations alone (ILfO). MP represents a novel generalization of RFE, while ILfO, introduced in [34], represents an exemplification of MP. Before enunciating the minimax lower bound, it is important to formally define each of these problems, as well as what we mean by learning in each problem.

E.1.1 Definition of the Problems

In all the 4 problems, the learner is placed into an *unknown* MDP without reward $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$, i.e., an environment whose dynamics (d_0, p) is unknown to the learner. For simplicity, w.l.o.g., we assume that there is a single initial state $s_0 := d_0$. In each problem, the learner can explore the environment at will to collect samples about the dynamics p , whose knowledge improves the performance of the agent at solving the task. However, at exploration phase, the learner does not

¹⁶Actually, for Linear MDPs, instead of evaluating the policy returned by Algorithm 2 of [62], we can simply consider the optimistic estimate of the V -function computed by such algorithm, which has the property of being ϵ -close to the true optimal V -function.

know which is the specific task it has to solve. It just knows that the specific task belongs to a given set of tasks \mathfrak{T} (e.g., set of reward functions). The agent can use the knowledge of \mathfrak{T} to engage in a more efficient task-driven exploration. For any $\epsilon, \delta \in (0, 1)$, the goal of the agent is to be able to outputting, for any task in \mathfrak{T} a quantity σ (e.g., a policy) that solves that specific task in an ϵ -correct manner with probability at least $1 - \delta$. The ultimate goal of exploration is to collect the least number of samples that permits (ϵ, δ) -correctness for all the tasks in \mathfrak{T} .

Now, let us see what the quantities \mathfrak{T} and σ represent in each of the 4 problems. In Table 1, we provide a sum up of the various definitions.

Reward-Free Exploration (RFE). In RFE, the learner receives a set of reward functions $\mathfrak{T} = \mathcal{R} \subseteq \mathfrak{R}$ in input, and the goal is to exploit the information about p collected at exploration phase to output, for any reward $r \in \mathcal{R}$, an ϵ -optimal policy $\sigma = \hat{\pi}_r$ w.p. $1 - \delta$. When $\mathfrak{T} = \{r\}$ is a singleton, the RFE problem is commonly termed the BPI problem. In symbols, any RFE algorithm must guarantee that:

$$\mathbb{P}\left(\sup_{r \in \mathcal{R}} J^*(r; p) - J^{\hat{\pi}_r}(r; p) \leq \epsilon\right) \geq 1 - \delta,$$

where $\hat{\pi}_r$ is the estimate of the algorithm for reward r .

Inverse Reinforcement Learning (IRL). In IRL, the learner receives in input an occupancy measure¹⁷ $\{d_h^{p, \pi^E}\}_{h \in [H]}$ and a set of reward functions $\mathcal{R} \subseteq \mathfrak{R}$: $\mathfrak{T} = (d^{p, \pi^E}, \mathcal{R})$, but it does not know which specific reward it will have to classify. Under the assumption that the occupancy measure d^{p, π^E} is known,¹⁸ the problem reduces to exploiting the information about p collected at exploration phase to output, for any reward $r \in \mathcal{R}$, an ϵ -correct estimate $\sigma = \hat{J}(r)$ of the optimal utility $J^*(r)$ w.p. $1 - \delta$. In symbols, under these conditions, any IRL algorithm must guarantee that:

$$\mathbb{P}\left(\sup_{r \in \mathcal{R}} |J^*(r; p) - \hat{J}(r)| \leq \epsilon\right) \geq 1 - \delta,$$

where $\hat{J}(r)$ is the estimate of the algorithm for reward r .

Matching Performance (MP). In MP, the learner receives in input a set of reward functions $\mathcal{R} \subseteq \mathfrak{R}$ and a measure of performance for each of them $\bar{J} : \mathcal{R} \rightarrow \mathbb{R}$: $\mathfrak{T} = (\bar{J}, \mathcal{R})$. For any $r \in \mathcal{R}$, the utility $\bar{J}(r)$ represents a performance measure for which we aim to find the policy that achieves closest performance. Thus, in MP, the goal is to exploit the information about p collected at exploration phase to output, for any reward $r \in \mathcal{R}$, a policy $\sigma = \hat{\pi}_r$ such that, if we denote the policy with performance closest to $\bar{J}(r)$ by $\bar{\pi}_r \in \arg \min_{\pi} |J^\pi(r) - \bar{J}(r)|$, then the utility of policy $\hat{\pi}_r$ is ϵ -close to the utility of policy $\bar{\pi}_r$ w.p. $1 - \delta$. In symbols, any MP algorithm must guarantee that:

$$\mathbb{P}\left(\sup_{r \in \mathcal{R}} |J^{\bar{\pi}_r}(r; p) - J^{\hat{\pi}_r}(r; p)| \leq \epsilon\right) \geq 1 - \delta,$$

where $\bar{\pi}_r \in \arg \min_{\pi} |J^\pi(r) - \bar{J}(r)|$, and $\hat{\pi}_r$ is the estimate of the algorithm for reward r .

Imitation Learning from Demonstrations alone (ILfO). In ILfO, the learner receives in input a set of *state-only* reward functions $\mathcal{R} \subset \mathfrak{R}$ and a *state-only* occupancy measure $\{\bar{d}_h\}_{h \in [H]}$: $\mathfrak{T} = (\bar{d}, \mathcal{R})$. Under the assumption that \bar{d} does not leak any information about the true transition model p , the goal is to exploit the information about p collected at exploration phase to output, for any reward $r \in \mathcal{R}$, a policy $\sigma = \hat{\pi}_r$ such that, if we denote the policy with performance closest to $\bar{J}(r) := \sum_{h \in [H]} \langle r_h, \bar{d}_h \rangle$ by $\bar{\pi}_r \in \arg \min_{\pi} |J^\pi(r) - \bar{J}(r)|$, then the utility of policy $\hat{\pi}_r$ is ϵ -close to the utility of policy $\bar{\pi}_r$ w.p. $1 - \delta$. Simply put, ILfO, as defined in this manner, exemplifies the MP setting by providing a functional form to $\bar{J} : \mathcal{R} \rightarrow \mathbb{R}$ as an inner product between a certain state-only occupancy measure and the input reward. It should be remarked that the assumption made for ILfO is mild, because it is

¹⁷Actually, as explained in Section 6.1, the knowledge of d^{p, π^E} at exploration phase is useless. The visit measure might be provided after the exploration along with the true reward to classify.

¹⁸The assumption that d^{p, π^E} is known is useful to reduce the estimation problem of the (non)compatibility of a reward $\bar{C}_{p, \pi^E}(r) := J^*(r; p) - J^{\pi^E}(r; p)$ to the problem of estimating the optimal utility $J^*(r; p)$ only. Indeed, if d^{p, π^E} is known, then, for any reward r , the utility $J^{\pi^E}(r; p)$ is known.

	BPI	IRL	MP	ILfO
Set of Tasks \mathfrak{T}	\mathcal{R}	$(d^{p, \pi^E}, \mathcal{R})$	(\bar{J}, \mathcal{R})	(\bar{d}, \mathcal{R})
Assumptions	/	d^{p, π^E} known	\bar{J} can be non-realizable	r state-only, \bar{d} no info
Output \mathfrak{o}	$\hat{\pi}$	\hat{J}	$\hat{\pi}$	$\hat{\pi}$
Goal	$J^{\hat{\pi}}(r; p) \approx J^*(r; p)$	$\hat{J} \approx J^*(r; p)$	$J^{\hat{\pi}}(r; p) \approx \bar{J}$	$J^{\hat{\pi}}(r; p) \approx \sum_h \langle \bar{d}_h, r_h \rangle$

Table 1: Summary of the problems.

satisfied by the setting in which the expert and the learner have the same state space but different action spaces (or different dynamics). Indeed, in such case, the visit distribution \bar{d} of the expert would not leak any information about p . In symbols, any ILfO algorithm must guarantee that:

$$\mathbb{P}\left(\sup_{r \in \mathcal{R}} |J^{\bar{\pi}_r}(r; p) - J^{\hat{\pi}_r}(r; p)| \leq \epsilon\right) \geq 1 - \delta,$$

where $\bar{\pi}_r \in \arg \min_{\pi} |J^{\pi}(r) - \bar{J}(r)|$ and $\bar{J}(r) := \sum_{h \in [H]} \langle r_h, \bar{d}_h \rangle$, and $\hat{\pi}_r$ is the estimate of the algorithm for reward r .

E.1.2 Lower Bound

We now present a minimax lower bound rate that is common to RFE, IRL, and MP. We report here the lower bounds presented in Section 6.1.

Theorem 6.1 (IRL Classification - Lower Bound). *Let \mathfrak{A} be an (ϵ, δ) -PAC algorithm for the IRL classification in tabular MDPs. Let τ be the number of exploration episodes. Then, there exists an IRL classification instance such that:*

$$\text{if } |\mathcal{R}| \geq 1 : \tau \geq \Omega\left(\frac{H^3 SA}{\epsilon^2} \log \frac{1}{\delta}\right), \quad \text{if } \mathcal{R} = \mathfrak{R} : \tau \geq \Omega\left(\frac{H^3 SA}{\epsilon^2} \left(S + \log \frac{1}{\delta}\right)\right).$$

Proof. The proof is similar to that of [38]. We split the proof in two parts, by considering two classes of difficult problem instances in Lemma E.2 and Lemma E.3. Next, we combine the two bounds through $\max\{a, b\} \geq (a + b)/2$ for all $a, b \geq 0$. For the proof, we will assume that the expert visit distribution is known. The obtained bound represents a lower bound to the more general setting in which it is unknown. \square

Theorem E.1 (RFE - Refined Lower Bound). *Let \mathfrak{A} be an (ϵ, δ) -PAC algorithm for RFE in tabular MDPs. Let τ be the number of exploration episodes. Then, there exists an RFE instance such that:*

$$\tau \geq \Omega\left(\frac{H^3 SA}{\epsilon^2} \left(S + \log \frac{1}{\delta}\right)\right).$$

Proof. The proof of this result is analogous to that of Theorem 6.1, and it employs Lemma E.2 and Lemma E.3. \square

Some observations are in order. First, since MP is a more general setting than RFE, then this lower bound is a lower bound for MP too. However, this is not guaranteed for ILfO. We observe that, while for RFE and IRL the bound is tight, for MP we cannot say so because we do not have the upper bound. Notice that, in case the expert state-only distribution \bar{d} was unknown at exploration phase, and revealed afterwards, then the lower bound of Theorem 6.1 holds for ILfO too, because we might a posteriori reveal the state-only distribution \bar{d} of the optimal policy, and thus, in such manner, ILfO would be reduced to RFE.

E.2 Missing proofs

Lemma E.2. *Let IRL and RFE be the learning problems defined as in Appendix E.1. Then, for each problem, any (ϵ, δ) -PAC algorithm must collect at least the following number of exploration episodes:*

$$\tau \geq \Omega\left(\frac{H^3 SA}{\epsilon^2} \log \frac{1}{\delta}\right).$$

Proof. Observe that the proof for RFE is present in [12]. Thus, we have to prove just the result for IRL. For doing so, we will use both the results of [12] and [38]. Notice that for the sake of this proof we consider $\mathcal{R} = \{r\}$, that will reduce our problem to simple RL as, in order to compute the function $\bar{\mathcal{C}}_{p, \pi^E}(r)$, we just need to compute $J^*(r; p)$, being $J^{\pi^E}(r; p)$ known from the availability of d^{p, π^E} and r .

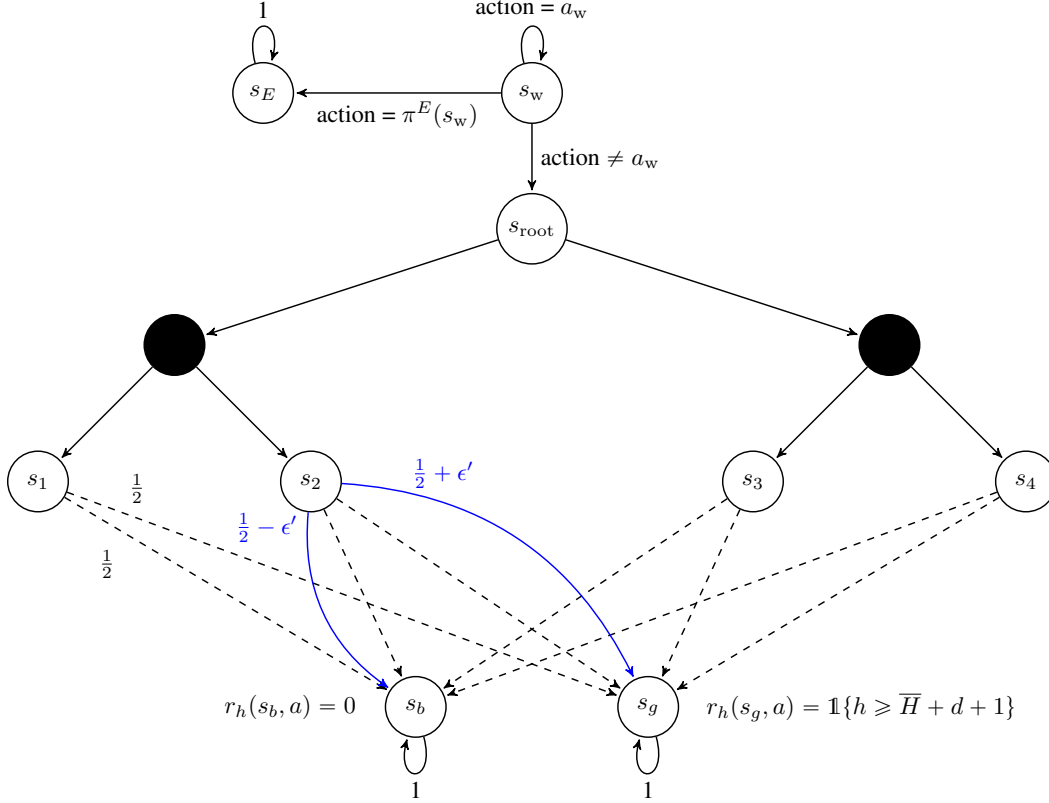


Figure 4: Hard instances.

Instances Description The hard instances considered are exactly the same as [12], and are reported in Figure 4 for simplicity. The only difference is the presence of state s_E , to which the expert’s policy π^E brings, which is absorbing. Such state is needed to make the knowledge of the expert’s visit distribution d^{p, π^E} useless at inferring information about the transition model in other parts of the state-action space. Based on [12], we describe such hard instances. Similarly to [12], we assume that $S \geq 7$, $A \geq 2$, and there exists an integer d such that $S = 4 + (A^d - 1)/(A - 1)$, and we assume that $H \geq 3d$. Note that [12] show how to relax the assumption on the existence of d .

There are the initial state s_w , from which the agent starts, and states s_g, s_b , respectively, the “good” and “bad” states which are absorbing. Moreover, there is state s_E , which is reached by the expert, and is absorbing. The remaining $S - 4$ states are arranged in a full A -ary tree of depth $d - 1$ with root s_{root} . We denote by $\bar{H} \leq H - d$ a certain integer parameter, and by $\mathcal{L} := \{s_1, s_2, \dots, s_L\}$ the set of leaves of the tree. We define $\mathcal{I} := \{1 + d, \dots, \bar{H} + d\} \times \mathcal{L} \times \mathcal{A}$. For any $i \in \mathcal{I}$, we define and MDP \mathcal{M}_i as follows. In any state of the tree, i.e., in states $\mathcal{S} \setminus \{s_w, s_g, s_b, s_E\}$, the transitions are deterministic, and the a -th action of a state brings to the a -th child of that node.

The transitions from s_w are given by

$$p_h(s_w | s_w, a) := \mathbb{1}\{a = a_w, h \leq \bar{H}\} \quad \text{and} \quad p_h(s_{\text{root}} | s_w, a) := 1 - p_h(s_w | s_w, a).$$

In other words, action a_w allows the agent to remain in the initial state s_w up to stage \overline{H} . After stage \overline{H} , the agent is forced to leave s_w and to traverse the tree down to the leaves. Action $a_E = \pi_1^E(s_w)$ is the only action that brings to state s_E , which is absorbing. The transitions from any leaf $s_i \in \mathcal{L}$ are given, as in [12], by:

$$p_h(s_g|s_i, a) := \frac{1}{2} + \Delta_{(h^*, \ell^*, a^*)}(h, s_i, a) \quad \text{and} \quad p_h(s_b|s_i, a) := \frac{1}{2} - \Delta_{(h^*, \ell^*, a^*)}(h, s_i, a), \quad (4)$$

where $\Delta_{(h^*, \ell^*, a^*)}(h, s_i, a) := \mathbb{1}\{(h, s_i, a) = (h^*, s_{\ell^*}, a^*)\} \cdot \epsilon'$, for some $\epsilon' \in [0, 1/2]$. For this reason, there exists a (single) leaf ℓ^* where the agent can choose an action a^* at stage h^* to increase its probability of arriving to the good state s_g , which provides higher reward. We define states s_g and s_b to be absorbing, i.e., they satisfy $p_h(s_b|s_b, a) := p_h(s_g|s_g, a) := 1$ for any action a . The reward function is state-only and is defined as

$$\forall a \in \mathcal{A}, \quad r_h(s, a) := \mathbb{1}\{s = s_g, h \geq \overline{H} + d + 1\},$$

so that even though the agent decides to stay at s_w until stage \overline{H} , it does not lose any reward. Observe that state s_E does not provide any reward, so that to estimate the (non)compatibility, any algorithm must provide a good estimate of the optimal performance.

Finally, we define a reference MDP \mathcal{M}_0 which is an MDP of the above type but for which $\Delta_0(h, s_i, a) := 0$ for all (h, s_i, a) . For certain ϵ' and \overline{H} to choose, we define the class \mathbb{M} to be the set $\mathbb{M} := \{\mathcal{M}_0\} \cup \{\mathcal{M}_i\}_{i \in \mathcal{I}}$.

Distance between problems We will prove the lower bound for instance \mathcal{M}_0 . Observe that, in \mathcal{M}_0 , the optimal utility is:

$$J_0^* = \frac{1}{2}(H - \overline{H} - d),$$

because there is no triple with additional bias towards s_g . Instead, for any other $\mathcal{M}_i \in \mathbb{M}$, the optimal utility is:

$$J_i^* = (H - \overline{H} - d) \left(\frac{1}{2} + \epsilon' \right).$$

Therefore, if we choose $\epsilon' := 2\epsilon/(H - \overline{H} - d)$, we have that, for any $i \in \mathcal{I}$:

$$|J_0^* - J_i^*| = 2\epsilon.$$

Thus, in particular, for any estimate $\hat{J} \in \mathbb{R}$ we necessarily have $|J_0^* - \hat{J}| \leq \epsilon \implies |J_i^* - \hat{J}| > \epsilon$, and vice versa, i.e., we cannot provide an estimate \hat{J} that is ϵ -close to both J_0^* and J_i^* .

Identifying the underlying problem Following [38], let us consider a generic (ϵ, δ) -correct algorithm \mathfrak{A} that outputs the estimated optimal utility \hat{J} . Then, for all $i \in \mathcal{I}$, we have:

$$\begin{aligned} \delta &\geq \sup_{\text{all problem instances } \mathcal{M}} \mathbb{P}_{\mathcal{M}, \mathfrak{A}} \left(|J_{\mathcal{M}}^* - \hat{J}| \geq \epsilon \right) \\ &\geq \sup_{\mathcal{M} \in \mathbb{M}} \mathbb{P}_{\mathcal{M}, \mathfrak{A}} \left(|J_{\mathcal{M}}^* - \hat{J}| \geq \epsilon \right) \\ &\geq \max_{\ell \in \{0, i\}} \mathbb{P}_{\mathcal{M}_\ell, \mathfrak{A}} \left(|J_\ell^* - \hat{J}| \geq \epsilon \right). \end{aligned}$$

For every $i \in \mathcal{I}$, we define the *identification function* Ψ_i as the index of the problem “recognized” by algorithm \mathfrak{A} . In symbols:

$$\Psi_i := \arg \min_{\ell \in \{0, i\}} |J_\ell^* - \hat{J}|.$$

In words, given estimate \hat{J} returned by algorithm \mathfrak{A} , the identification function Ψ_i returns the problem between \mathcal{M}_0 and \mathcal{M}_i whose optimal utility is closest to the estimate \hat{J} . For what we have seen in the previous paragraph, problems \mathcal{M}_0 and \mathcal{M}_i lie at a distance of at least 2ϵ for all $i \in \mathcal{I}$. Therefore, for $j \in \{0, i\}$, we have the following inclusion of events:

$$\{\Psi_i \neq j\} \subseteq \{|J_j^* - \hat{J}| > \epsilon\}.$$

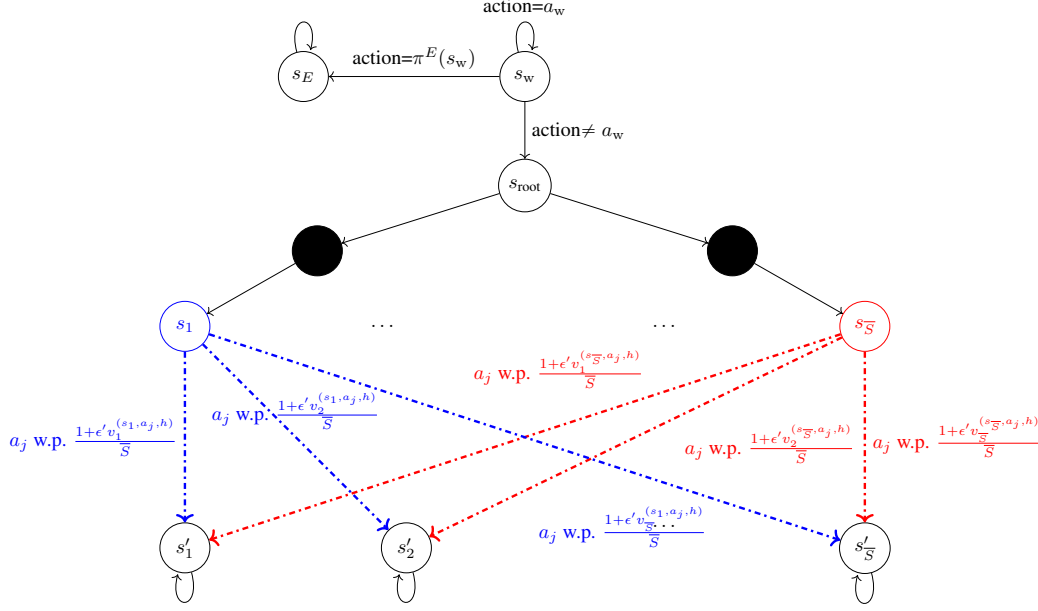


Figure 5: Hard instances.

Thanks to this fact, we can continue lower bounding the probability as:

$$\begin{aligned}
\max_{\ell \in \{0, i\}} \mathbb{P}_{\mathcal{M}_\ell, \mathfrak{A}} \left(\left| J_\ell^* - \hat{J} \right| \geq \epsilon \right) &\geq \max_{\ell \in \{0, i\}} \mathbb{P}_{\mathcal{M}_\ell, \mathfrak{A}} \{ \Psi_i \neq \ell \} \\
&\stackrel{(1)}{\geq} \frac{1}{2} \left[\mathbb{P}_{\mathcal{M}_0, \mathfrak{A}} (\Psi_i \neq 0) + \mathbb{P}_{\mathcal{M}_i, \mathfrak{A}} (\Psi_i \neq i) \right] \\
&= \frac{1}{2} \left[\mathbb{P}_{\mathcal{M}_0, \mathfrak{A}} (\Psi_i \neq 0) + \mathbb{P}_{\mathcal{M}_i, \mathfrak{A}} (\Psi_i = 0) \right] \\
&\stackrel{(2)}{\geq} \frac{1}{4} \exp^{-\text{KL}(\mathbb{P}_{\mathcal{M}_0, \mathfrak{A}}, \mathbb{P}_{\mathcal{M}_i, \mathfrak{A}})},
\end{aligned}$$

where at (1) we have lower bounded the maximum with the average, i.e., $\max\{a, b\} \geq (a + b)/2$ for all $a, b \geq 0$, and at (2) we have applied the Bretagnolle-Huber's inequality [38].

KL-divergence computation The proof can be concluded by upper bounding the KL divergence $\text{KL}(\mathbb{P}_{\mathcal{M}_0, \mathfrak{A}}, \mathbb{P}_{\mathcal{M}_i, \mathfrak{A}})$ as in the proof of Theorem 7 in [12], and then summing over all the $\Theta(SAH)$ instances to retrieve the result. \square

Lemma E.3. *Let IRL and RFE be the learning problems defined as in Appendix E.1. For each problem, if the set of reward functions \mathcal{R} in input is $\mathcal{R} = \mathfrak{R}$, then any (ϵ, δ) -PAC algorithm must collect at least the following number of exploration episodes:*

$$\tau \geq \Omega \left(\frac{H^3 S^2 A}{\epsilon^2} \right).$$

Proof. Instances description The hard instances that we use for the proof of this lemma are obtained by combining the hard instances in Lemma E.2 (i.e., the hard instances of [12]), with those in [38]. Specifically, this construction is based on the intuition described in [21] that, if we want to increase the sample complexity, we have to learn transitions also to $\Theta(S)$ states, and not just from $\Theta(S)$ states. Observe the presence of state s_E (only for IRL), which plays the same role as s_E in the proof of Lemma E.2. Any action in such state receives always reward -1 , thus it is meaningless for the estimate of the (non)compatibility, which reduces to the estimation of the optimal performance. In this manner, the expert distribution d^{p, π^E} does not provide additional information about the transition model of

other portion of the state-action space. Therefore, in the following, we will present the lower bound construction as if such state did not exist.

The hard instances are reported in Figure 5. Notice that they are exactly the same instances as those presented in the proof of Lemma E.2, with the difference that, from the \bar{S} leaves (differently from earlier, we now denote the number of leaves through \bar{S} instead of L), we do not reach just two states s_g, s_b , but we reach $\Theta(S)$ absorbing states, i.e., $s'_1, s'_2, \dots, s'_{\bar{S}}$. The transitions from the leaves to such states is the same as in [38], and we report a description below.

Let us introduce the set $\bar{\mathcal{I}} := \{s_1, \dots, s_{\bar{S}}\} \times \mathcal{A} \times \{1 + d, \dots, \bar{H} + d\}$. Let $\bar{\iota} := (s_1, a_1, 1 + d) \in \bar{\mathcal{I}}$ be a specific triple of set $\bar{\mathcal{I}}$, and denote $\mathcal{I} := \bar{\mathcal{I}} \setminus \{\bar{\iota}\}$. Let us also introduce set $\mathcal{V} := \{v \in \{-1, 1\}^{\bar{S}} : \sum_{j=1}^{\bar{S}} v_j = 0\}$. Thanks to Lemma E.6 of [38] (that we report in Lemma E.4 for simplicity), we know that there exists a subset $\bar{\mathcal{V}} \subseteq \mathcal{V}$ (of transition models) with cardinality at least $2^{\bar{S}/5}$ such that, for every pair $v, w \in \bar{\mathcal{V}}$ with $v \neq w$, we have that $\|v - w\|_1 \geq \bar{S}/16$. In other words, we know that there exists a $\bar{S}/16$ -packing of \mathcal{V} with cardinality at least $2^{\bar{S}/5}$.

Following [38], we denote by $\mathbf{v} = (v^i)_{i \in \mathcal{I}} \in \bar{\mathcal{V}}^{\mathcal{I}}$ the generic vector of $\bar{\mathcal{V}}^{\mathcal{I}}$. Now, for any $\mathbf{v} \in \bar{\mathcal{V}}^{\mathcal{I}}$, for any triple $\bar{j} \in \bar{\mathcal{I}}$, and for some parameter $\epsilon' \in [0, 1/2]$ to choose, we construct problem instance $\mathcal{M}_{\mathbf{v}, \bar{j}}$ as follows.

First of all, we define the transition model at triple $\bar{\iota}$ as:

$$p_{h_{\bar{\iota}}}(s'_i | s_{\bar{\iota}}, a_{\bar{\iota}}) = \frac{1}{\bar{S}} \quad \forall i \in [\bar{S}],$$

where observe that we use notation $\iota = (s_i, a_i, h_i) \in \bar{\mathcal{I}}$ to denote triples in $\bar{\mathcal{I}}$. Instead, for the generic triple $\iota \in \mathcal{I}$ (including triple j), the probability distribution of the next state is given by:

$$p_{h_\iota}(s'_i | s_\iota, a_\iota) = \frac{1}{\bar{S}} + \frac{\epsilon'}{\bar{S}} \mathbf{v}_i^i \quad \forall i \in [\bar{S}],$$

where \mathbf{v}_i^i represents the i -th component of the ι -th vector in \mathbf{v} . In words, the i -th component of vector $\mathbf{v}^i \in \bar{\mathcal{V}}$ creates a bias of ϵ'/\bar{S} towards the next state s'_i for all $i \in [\bar{S}]$. Since $\mathbf{v}^i \in \bar{\mathcal{V}}$, then $p_{h_\iota}(\cdot | s_\iota, a_\iota) \in \Delta^{[\bar{S}]}$ for all $\iota \in \mathcal{I}$.

We consider non-stationary reward functions. Specifically, all the rewards $r \in \mathfrak{R}$ that we consider assign reward 1 to both triples $\bar{\iota}$ and \bar{j} , i.e., $r_{h_{\bar{\iota}}}(s_{\bar{\iota}}, a_{\bar{\iota}}) = 1$ and $r_{h_{\bar{j}}}(s_{\bar{j}}, a_{\bar{j}}) = 1$. Next, for any other triple $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [\bar{H}]$ with state different from $s'_1, s'_2, \dots, s'_{\bar{S}}$, we assign reward 0. For states $s'_1, s'_2, \dots, s'_{\bar{S}}$, we consider state-only rewards whose value is always 0 in stages $[1, \bar{H} + d]$, and whose value is stationary and arbitrary afterwards. Intuitively, as in [12], forcing the reward to be 0 up $h = \bar{H} + d$ guarantees that we cannot obtain a higher expected return J by reaching the leaves states earlier (i.e., by exiting from s_w before \bar{H}).

Given the definition above, we construct the class of instances $\mathbb{M} := \{\mathcal{M}_{\mathbf{v}, \iota} : \iota \in \mathcal{I}, \mathbf{v} \in \bar{\mathcal{V}}^{\mathcal{I}}\}$. Moreover, we will use the notation $\mathcal{M}_{\mathbf{v}^{\leftarrow w}, j}$ to denote the instance in which we replace the ι component of \mathbf{v} , i.e., \mathbf{v}^i , with $w \in \mathcal{V}$ and $\mathcal{M}_{\mathbf{v}^{\leftarrow 0}, j}$ the instance in which we replace the ι component of \mathbf{v} , i.e., \mathbf{v}^i , with the zero vector. Since we will always use this notation when substituting triple j , i.e., we always use this notation in situations as $\mathcal{M}_{\mathbf{v}^{\leftarrow w}, j}$, then we omit the second parameter, and write just $\mathcal{M}_{\mathbf{v}^{\leftarrow w}} := \mathcal{M}_{\mathbf{v}^{\leftarrow w}, j}$.

Distance between problems Consider an arbitrary problem instance $\mathcal{M}_{\mathbf{v}, \iota} \in \mathbb{M}$, for certain $\iota \in \mathcal{I}$ and $\mathbf{v} \in \bar{\mathcal{V}}^{\mathcal{I}}$. Let $r \in \mathfrak{R}$ be an arbitrary reward function that satisfies the constraints described earlier. Let $\pi_{\bar{\iota}} \in \Pi$ be the deterministic policy that brings to triple $\bar{\iota}$. Then, its expected return is:

$$J^{\pi_{\bar{\iota}}}(r; \mathcal{M}_{\mathbf{v}, \iota}) = 1 + \frac{H - \bar{H} - d}{\bar{S}} \sum_{i=1}^{\bar{S}} r_i,$$

where $r_i := r_{\overline{H}+d+1}(s'_i)$ for all $i \in \llbracket \overline{S} \rrbracket$. Let policy $\pi_\iota \in \Pi$ be the deterministic policy that brings to triple ι . Then, its expected return is:

$$J^{\pi_\iota}(r; \mathcal{M}_{v,\iota}) = 1 + \frac{H - \overline{H} - d}{\overline{S}} \sum_{i=1}^{\overline{S}} r_i + \epsilon' \frac{(H - \overline{H} - d)}{\overline{S}} \sum_{i=1}^{\overline{S}} v_i^{\iota} r_i.$$

Finally, let policy $\pi_j \in \Pi$ be the deterministic policy that brings to any other triple $j \in \mathcal{T} \setminus \{\iota\}$. Then, its expected return is:

$$J^{\pi_j}(r; \mathcal{M}_{v,\iota}) = 0 + \frac{H - \overline{H} - d}{\overline{S}} \sum_{i=1}^{\overline{S}} r_i + \epsilon' \frac{(H - \overline{H} - d)}{\overline{S}} \sum_{i=1}^{\overline{S}} v_i^j r_i.$$

It should be remarked that $(v, r) = \sum_{i \in \llbracket \overline{S} \rrbracket} v_i r_i \in [-\overline{S}, \overline{S}]$ for any $r \in \mathfrak{R}$ and $v \in \overline{\mathcal{V}}$, therefore, as long as:

$$\epsilon'(H - \overline{H} - d) < 1 - \epsilon'(H - \overline{H} - d) - \epsilon \iff \epsilon' < \frac{1 - \epsilon}{2(H - \overline{H} - d)}, \quad (5)$$

then any policy π_j is cannot be ϵ -optimal in problem $\mathcal{M}_{v,\iota}$, in which, thus, the optimal policy shall be searched for between $\pi_{\overline{\iota}}$ and π_ι .

Now, consider an arbitrary pair $v, w \in \overline{\mathcal{V}}$ such that $v \neq w$, and an arbitrary triple $\iota \in \mathcal{T}$ and vector $\epsilon \in \overline{\mathcal{V}}^{\mathcal{T}}$. We now compare problem instances $\mathcal{M}_{v \stackrel{\epsilon}{\leftarrow} v}$ and $\mathcal{M}_{v \stackrel{\epsilon}{\leftarrow} w}$. Among all possible reward functions that satisfy the definition provided in the construction of the hard instances, we find reward r' such that, in every component $i \in \llbracket \overline{S} \rrbracket$, satisfies:

$$r'_i = \begin{cases} +1 & \text{if } v_i = +1 \wedge w_i = -1 \\ -1 & \text{if } v_i = -1 \wedge w_i = +1 \\ 0 & \text{if } v_i = w_i \end{cases}.$$

For what we have seen before about class $\overline{\mathcal{V}}$, we know that $\|v - w\|_1 = \sum_{i \in \llbracket \overline{S} \rrbracket} |v_i - w_i| \geq \overline{S}/16$, thus, since $v, w \in \mathcal{V}$, i.e., their components belong to $\{-1, +1\}$, we know that there are at least $\overline{S}/32$ components of v, w that differ from each other. By using reward r' , we have that:

$$\begin{aligned} \sum_{i=1}^{\overline{S}} v_i r'_i &\geq \frac{\overline{S}}{32} \geq 0, \\ \sum_{i=1}^{\overline{S}} w_i r'_i &\leq -\frac{\overline{S}}{32} \leq 0. \end{aligned}$$

As a consequence, the expected returns of policies $\pi_{\overline{\iota}}$ and π_ι in problems $\mathcal{M}_{v \stackrel{\epsilon}{\leftarrow} v}$ and $\mathcal{M}_{v \stackrel{\epsilon}{\leftarrow} w}$ are:

$$\begin{aligned} J^{\pi_{\overline{\iota}}}(r'; \mathcal{M}_{v \stackrel{\epsilon}{\leftarrow} v}) &= J^{\pi_{\overline{\iota}}}(r'; \mathcal{M}_{v \stackrel{\epsilon}{\leftarrow} w}) = 1 + \frac{H - \overline{H} - d}{\overline{S}} \sum_{i=1}^{\overline{S}} r'_i, \\ J^{\pi_\iota}(r'; \mathcal{M}_{v \stackrel{\epsilon}{\leftarrow} v}) &\geq 1 + \frac{H - \overline{H} - d}{\overline{S}} \sum_{i=1}^{\overline{S}} r'_i + \epsilon' \frac{(H - \overline{H} - d)}{32}, \\ J^{\pi_\iota}(r'; \mathcal{M}_{v \stackrel{\epsilon}{\leftarrow} w}) &\leq 1 + \frac{H - \overline{H} - d}{\overline{S}} \sum_{i=1}^{\overline{S}} r'_i - \epsilon' \frac{(H - \overline{H} - d)}{32}, \end{aligned}$$

from which we infer that:

$$J^{\pi_\iota}(r'; \mathcal{M}_{v \stackrel{\epsilon}{\leftarrow} v}) \geq J^{\pi_{\overline{\iota}}}(r'; \mathcal{M}_{v \stackrel{\epsilon}{\leftarrow} v}) = J^{\pi_{\overline{\iota}}}(r'; \mathcal{M}_{v \stackrel{\epsilon}{\leftarrow} w}) \geq J^{\pi_\iota}(r'; \mathcal{M}_{v \stackrel{\epsilon}{\leftarrow} w}).$$

Now, let us choose $\epsilon' > 64\epsilon/(H - \overline{H} - d)$. To satisfy also the constraint in Equation (5), we can roughly assume $\epsilon < 1/256$ and set $\epsilon' = 65\epsilon/(H - \overline{H} - d)$. Thanks to this choice, observe that:

$$\begin{aligned} J^{\pi_\iota}(r'; \mathcal{M}_{v \stackrel{\epsilon}{\leftarrow} v}) &> J^{\pi_{\overline{\iota}}}(r'; \mathcal{M}_{v \stackrel{\epsilon}{\leftarrow} v}) + 2\epsilon, \\ J^{\pi_{\overline{\iota}}}(r'; \mathcal{M}_{v \stackrel{\epsilon}{\leftarrow} w}) &> J^{\pi_\iota}(r'; \mathcal{M}_{v \stackrel{\epsilon}{\leftarrow} w}) + 2\epsilon. \end{aligned}$$

In words, policy π_i is optimal in problem $\mathcal{M}_{v \leftarrow v}$, and policy $\pi_{\bar{v}}$ is worse than 2ϵ -suboptimal in such problem. In addition, observe that policy $\pi_{\bar{v}}$ is optimal in problem $\mathcal{M}_{v \leftarrow w}$, and policy π_i is worse than 2ϵ -suboptimal in such problem. We stress that any stochastic policy in-between π_i and $\pi_{\bar{v}}$ cannot be ϵ -optimal for both problems.

To sum up, for the choice of ϵ' made earlier, for arbitrary pairs of problems $\mathcal{M}_{v \leftarrow v}$ and $\mathcal{M}_{v \leftarrow w}$, we have seen that there exist rewards in \mathfrak{R} for which a policy ϵ -optimal for problem $\mathcal{M}_{i,v}$ is not ϵ -optimal for problem $\mathcal{M}_{i,w}$, and vice versa.

Identifying the underlying problem: RFE. We consider first RFE, and then IRL.

Let us consider an (ϵ, δ) -correct algorithm \mathfrak{A} for RFE, that outputs, for any reward function $r \in \mathfrak{R}$, a policy $\hat{\pi}_r$. For simplicity, we consider as output of Algorithm \mathfrak{A} a function $\hat{\pi} : \mathfrak{R} \rightarrow \Pi$, that takes in input a reward and outputs a policy.

For any $i \in \mathcal{I}$ and $v \in \bar{\mathcal{V}}^{\mathcal{I}}$, we can lower bound the error probability as:

$$\begin{aligned} \delta &\geq \sup_{\text{all problem instances } \mathcal{M}} \mathbb{P}_{\mathcal{M}, \mathfrak{A}} \left(\sup_{r \in \mathfrak{R}} J_{\mathcal{M}}^*(r) - J_{\hat{\pi}_r}^{\mathcal{M}}(r) \geq \epsilon \right) \\ &\stackrel{(1)}{\geq} \sup_{\mathcal{M} \in \mathbb{M}} \mathbb{P}_{\mathcal{M}, \mathfrak{A}} \left(\sup_{r \in \mathfrak{R}} J_{\mathcal{M}}^*(r) - J_{\hat{\pi}_r}^{\mathcal{M}}(r) \geq \epsilon \right) \\ &\stackrel{(2)}{\geq} \max_{w \in \bar{\mathcal{V}}} \mathbb{P}_{\mathcal{M}_{v \leftarrow w}, \mathfrak{A}} \left(\sup_{r \in \mathfrak{R}} J_{\mathcal{M}_{v \leftarrow w}}^*(r) - J_{\hat{\pi}_r}^{\mathcal{M}_{v \leftarrow w}}(r) \geq \epsilon \right), \end{aligned}$$

where at (1) we have lower bounded by replacing all possible RFE problem instances with problem instances in \mathbb{M} , and at (2) we have lower bounded by replacing all instances in \mathbb{M} with just instances $\{\mathcal{M}_{v \leftarrow w} : w \in \bar{\mathcal{V}}\}$ for the fixed triple i and vector v .

For every $i \in \mathcal{I}$ and $v \in \bar{\mathcal{V}}^{\mathcal{I}}$, we define the *identification function* $\Psi_{i,v}$ as the index of the problem $w \in \bar{\mathcal{V}}$ “recognized” by algorithm \mathfrak{A} . In symbols:

$$\Psi_{i,v} := \arg \min_{w \in \bar{\mathcal{V}}} \sup_{r \in \mathfrak{R}} J_{\mathcal{M}_{v \leftarrow w}}^*(r) - J_{\hat{\pi}_r}^{\mathcal{M}_{v \leftarrow w}}(r).$$

In words, given estimate $\hat{\pi} : \mathfrak{R} \rightarrow \Pi$ returned by algorithm \mathfrak{A} , the identification function $\Psi_{i,v}$ returns the problem in $\{\mathcal{M}_{v \leftarrow w} : w \in \bar{\mathcal{V}}\}$ whose solution $\pi : \mathfrak{R} \rightarrow \Pi$ is closest to the estimate $\hat{\pi}$. For what we have seen in the previous paragraph, for any $v, w \in \bar{\mathcal{V}}$ with $v \neq w$, for any fixed $i \in \mathcal{I}$ and $v \in \bar{\mathcal{V}}^{\mathcal{I}}$, there exists a reward function $r' \in \mathfrak{R}$ such that no policy can have expected utility ϵ -close to the optimal expected utility of both problems $\mathcal{M}_{v \leftarrow v}$ and $\mathcal{M}_{v \leftarrow w}$. Therefore, for $w \in \bar{\mathcal{V}}$, we have the following inclusion of events:

$$\{\Psi_{i,v} \neq w\} \subseteq \left\{ \sup_{r \in \mathfrak{R}} J_{\mathcal{M}_{v \leftarrow w}}^*(r) - J_{\hat{\pi}_r}^{\mathcal{M}_{v \leftarrow w}}(r) > \epsilon \right\}.$$

We can continue to lower bound the probability as:

$$\begin{aligned} \max_{w \in \bar{\mathcal{V}}} \mathbb{P}_{\mathcal{M}_{v \leftarrow w}, \mathfrak{A}} \left(\sup_{r \in \mathfrak{R}} J_{\mathcal{M}_{v \leftarrow w}}^*(r) - J_{\hat{\pi}_r}^{\mathcal{M}_{v \leftarrow w}}(r) \geq \epsilon \right) &\stackrel{(3)}{\geq} \frac{1}{|\bar{\mathcal{V}}|} \sum_{w \in \bar{\mathcal{V}}} \mathbb{P}_{\mathcal{M}_{v \leftarrow w}, \mathfrak{A}} (\Psi_{i,v} \neq w) \\ &\stackrel{(4)}{\geq} 1 - \frac{1}{\log |\bar{\mathcal{V}}|} \left(\frac{1}{|\bar{\mathcal{V}}|} \sum_{w \in \bar{\mathcal{V}}} \text{KL}(\mathbb{P}_{\mathcal{M}_{v \leftarrow w}, \mathfrak{A}}, \mathbb{P}_{\mathcal{M}_{v \leftarrow 0}, \mathfrak{A}}) - \log 2 \right), \end{aligned}$$

where at (3) we have lower bounded the maximum over $\bar{\mathcal{V}}$ with the average, and at (4) we have applied, similarly to [38], the Fano’s inequality, reported in Theorem E.5 for simplicity.

Identifying the underlying problem: IRL. For IRL, it is possible to carry out a similar derivation. However, we remark that, now, the error is measured based on the expected utilities, and not on the policies.

Let us consider an (ϵ, δ) -correct algorithm \mathfrak{A} for IRL, that outputs, for any reward function $r \in \mathfrak{R}$, a utility \hat{J}_r . For simplicity, we consider as output of Algorithm \mathfrak{A} a function $\hat{J} : \mathfrak{R} \rightarrow \mathbb{R}$, that takes in input a reward and outputs a utility.

For any $\iota \in \mathcal{I}$ and $\mathbf{v} \in \overline{\mathcal{V}}^{\mathcal{I}}$, we can lower bound the error probability as:

$$\begin{aligned} \delta &\geq \sup_{\text{all problem instances } \mathcal{M}} \mathbb{P}_{\mathcal{M}, \mathfrak{A}} \left(\sup_{r \in \mathfrak{R}} |J_{\mathcal{M}}^*(r) - \hat{J}_r| \geq \epsilon \right) \\ &\geq \sup_{\mathcal{M} \in \mathcal{M}} \mathbb{P}_{\mathcal{M}, \mathfrak{A}} \left(\sup_{r \in \mathfrak{R}} |J_{\mathcal{M}}^*(r) - \hat{J}_r| \geq \epsilon \right) \\ &\geq \max_{w \in \overline{\mathcal{V}}} \mathbb{P}_{\mathcal{M}_{\mathbf{v}^{\perp w}}, \mathfrak{A}} \left(\sup_{r \in \mathfrak{R}} |J_{\mathcal{M}_{\mathbf{v}^{\perp w}}}^*(r) - \hat{J}_r| \geq \epsilon \right). \end{aligned}$$

For any $\iota \in \mathcal{I}$ and $\mathbf{v} \in \overline{\mathcal{V}}^{\mathcal{I}}$, we define an identification function $\Psi_{\iota, \mathbf{v}}$ as:

$$\Psi_{\iota, \mathbf{v}} := \arg \min_{w \in \overline{\mathcal{V}}} \sup_{r \in \mathfrak{R}} |J_{\mathcal{M}_{\mathbf{v}^{\perp w}}}^*(r) - \hat{J}_r|,$$

and by a reasoning analogous to that for RFE, we can continue to lower bounding as:

$$\begin{aligned} \max_{w \in \overline{\mathcal{V}}} \mathbb{P}_{\mathcal{M}_{\mathbf{v}^{\perp w}}, \mathfrak{A}} \left(\sup_{r \in \mathfrak{R}} |J_{\mathcal{M}_{\mathbf{v}^{\perp w}}}^*(r) - \hat{J}_r| \geq \epsilon \right) &\geq \frac{1}{|\overline{\mathcal{V}}|} \sum_{w \in \overline{\mathcal{V}}} \mathbb{P}_{\mathcal{M}_{\mathbf{v}^{\perp w}}, \mathfrak{A}} (\Psi_{\iota, \mathbf{v}} \neq w) \\ &\geq 1 - \frac{1}{\log |\overline{\mathcal{V}}|} \left(\frac{1}{|\overline{\mathcal{V}}|} \sum_{w \in \overline{\mathcal{V}}} \text{KL}(\mathbb{P}_{\mathcal{M}_{\mathbf{v}^{\perp w}}, \mathfrak{A}}, \mathbb{P}_{\mathcal{M}_{\mathbf{v}^{\perp 0}}, \mathfrak{A}}) - \log 2 \right), \end{aligned} \quad (6)$$

which represents the same lower bound obtained also for RFE.

KL-divergence computation The following derivation is analogous to that of [38]. To bound the KL-divergence term, for any $\iota \in \mathcal{I}$, we can write:

$$\begin{aligned} \text{KL}(\mathbb{P}_{\mathcal{M}_{\mathbf{v}^{\perp w}}, \mathfrak{A}}, \mathbb{P}_{\mathcal{M}_{\mathbf{v}^{\perp 0}}, \mathfrak{A}}) &\stackrel{(1)}{=} \mathbb{E}_{\mathcal{M}_{\mathbf{v}^{\perp w}}, \mathfrak{A}} [N_{h_{\iota}}^{\tau}(s_{\iota}, a_{\iota})] \text{KL}(p_{h_{\iota}}^{\mathcal{M}_{\mathbf{v}^{\perp w}}}(\cdot | s_{\iota}, a_{\iota}), p_{h_{\iota}}^{\mathcal{M}_{\mathbf{v}^{\perp 0}}}(\cdot | s_{\iota}, a_{\iota})) \\ &\stackrel{(2)}{\leq} 2(\epsilon')^2 \mathbb{E}_{\mathcal{M}_{\mathbf{v}^{\perp w}}, \mathfrak{A}} [N_{h_{\iota}}^{\tau}(s_{\iota}, a_{\iota})], \end{aligned}$$

where at (1) we have applied Lemma E.7, and at (2) we have applied Lemma E.6 (having observed that the transition models differ in ι and defined $N_{h_{\iota}}^{\tau}(s_{\iota}, a_{\iota}) = \sum_{t=1}^{\tau} \mathbb{1}\{(s_t, a_t, h_t) = (s_{\iota}, a_{\iota}, h_{\iota})\}$).

Plugging into Equation (6), we get:

$$\delta \geq \frac{1}{|\overline{\mathcal{V}}|} \sum_{w \in \overline{\mathcal{V}}} \mathbb{P}_{\mathcal{M}_{\mathbf{v}^{\perp w}}, \mathfrak{A}} (\Psi_{\iota, \mathbf{v}} \neq w) \implies \frac{1}{|\overline{\mathcal{V}}|} \sum_{w \in \overline{\mathcal{V}}} \mathbb{E}_{\mathcal{M}_{\mathbf{v}^{\perp w}}, \mathfrak{A}} [N_{h_{\iota}}^{\tau}(s_{\iota}, a_{\iota})] \geq \frac{(1 - \delta) \log |\overline{\mathcal{V}}| - \log 2}{2(\epsilon')^2}.$$

Notice that, since $|\overline{\mathcal{V}}| = \Theta(e^S)$ and $\epsilon' = \Theta(\epsilon/H)$, then this bound is in the order of $\Omega(\frac{H^2 S}{\epsilon^2})$. To get the additional $\Omega(SAH)$ dependence, we can make the same observation as in [38], i.e., that since the derivation is carried out for every $\iota \in \mathcal{I}$ and $\mathbf{v} \in \overline{\mathcal{V}}^{\mathcal{I}}$, we can perform the summation over ι and the average over \mathbf{v} . By noticing that we get a guarantee on a mean under the uniform distribution of the instances of the sample complexity, we realize that there must exist one $\mathbf{v}^{\text{hard}} \in \overline{\mathcal{V}}$ for which it holds the desired $\Omega(\frac{H^3 S^2 A}{\epsilon^2})$ dependency. \square

E.2.1 Technical Tools

We report here some results from other works. The notation adopted is the same as the original works.

Lemma E.4 (Lemma E.6 of [38]). *Let $\mathcal{V} = \{v \in \{-1, 1\}^D : \sum_{j=1}^D v_j = 0\}$. Then, the $\frac{D}{16}$ -packing number of \mathcal{V} w.r.t. the metric $d(v, v') = \sum_{j=1}^D |v_j - v'_j|$ is lower bounded by $2^{\frac{D}{5}}$.*

Theorem E.5. (Theorem E.2 of [38]) *Let $\mathbb{P}_0, \mathbb{P}_1, \dots, \mathbb{P}_M$ be probability measures on the same measurable space (Ω, \mathcal{F}) , and let $\mathcal{A}_1, \dots, \mathcal{A}_M \in \mathcal{F}$ be a partition of Ω . Then,*

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

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

Lemma E.6 (Lemma E.4 of [38]). *Let $\epsilon \in [0, 1/2]$ and $\mathbf{v} \in \{-\epsilon, \epsilon\}^D$ such that $\sum_{i=1}^d v_i = 0$. Consider the two categorical distributions $\mathbb{P} = (\frac{1}{D}, \frac{1}{D}, \dots, \frac{1}{D})$ and $\mathbb{Q} = (\frac{1+v_1}{D}, \frac{1+v_2}{D}, \dots, \frac{1+v_D}{D})$. Then, it holds that:*

$$D_{KL}(\mathbb{P}, \mathbb{Q}) \leq 2\epsilon^2 \quad \text{and} \quad D_{KL}(\mathbb{Q}, \mathbb{P}) \leq 2\epsilon^2.$$

Lemma E.7 (Lemma 5 of [12]). *Let \mathcal{M} and \mathcal{M}' be two MDPs that are identical except for their transition probabilities, denoted by p_h and p'_h , respectively. Assume that we have $\forall (s,a), p_h(\cdot|s,a) \ll p'_h(\cdot|s,a)$. Then, for any stopping time τ with respect to $(\mathcal{F}_H^t)_{t \geq 1}$ that satisfies $\mathbb{P}_{\mathcal{M}}\tau < \infty = 1$,*

$$KL(\mathcal{P}_{\mathcal{M}}^{I_H^\tau}, \mathcal{P}_{\mathcal{M}'}^{I_H^\tau}) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{h \in [H-1]} \mathbb{E}_{\mathcal{M}} [N_{h,s,a}^\tau] KL(p_h(\cdot|s,a), p'_h(\cdot|s,a)),$$

where $N_{h,s,a}^\tau := \sum_{t=1}^\tau \mathbb{1}\{(S_h^t, A_h^t) = (s, a)\}$ and $I_H^\tau : \Omega \rightarrow \bigcup_{t \geq 1} \mathcal{I}_H^t : \omega \mapsto I_H^{\tau(\omega)}(\omega)$ is the random vector representing the history up to episode τ .

F A Use Case for Objective-Free Exploration (OFE)

Consider the following setting. You are given a certain MDP without reward $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$, in which you do not know neither d_0 nor p . Your job is to explore the environment to collect samples that allow you to construct estimates $\hat{d}_0 \approx d_0$ and $\hat{p} \approx p$, that will be subsequently used to perform a task in a given class \mathcal{F} in an (ϵ, δ) -correct manner. Of course the number of samples should be as small as possible. How do you explore? It depends on which problems are contained in class \mathcal{F} .

A use case for OFE is the following.

Example F.1. *Assume that we are given a single fixed environment (for instance, a warehouse), in which there are many tasks to do (e.g., labelling objects, putting stuff on the shelves, bringing products from one side to the other), and assume (it is reasonable) that it is desirable to have one robot for each task. To teach these robots how to behave, we decide to use RL. Since all the robots work in the same environment (warehouse), then the (unknown) transition model is the same. For this reason, an efficient exploration (potentially through RFE) is meaningful. However, we realize that some tasks are difficult to design (i.e., the rewards of such tasks). For these tasks, we prefer to use a human expert to exhibit demonstrations, and then use ReL (in particular, IRL), to learn the reward, that will be subsequently used for AL. To perform IRL nicely, the samples collected at the beginning shall be used. To sum up, we might be interested in performing multiple RL and IRL tasks in the same unknown MDP, and, for efficiency reasons, our exploration of the environment has to be performed only once (before) being given the tasks to solve.*

NeurIPS Paper Checklist

1. Claims

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

Answer: [Yes]

Justification: The Reviewer can find the list of contributions in Section 1, and the related content in the rest of the paper.

Guidelines:

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

2. Limitations

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

Answer: [Yes]

Justification: The Reviewer can find a “Limitations” paragraph in Section 7.

Guidelines:

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

3. Theory Assumptions and Proofs

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

Answer: [Yes]

Justification: The Reviewer can find the full set of complete proofs in the supplemental material.

Guidelines:

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

4. Experimental Result Reproducibility

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

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

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

5. Open access to data and code

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

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

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

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

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

7. Experiment Statistical Significance

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

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

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

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

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

9. Code Of Ethics

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

Answer: [Yes]

Justification: The research conducted in the paper does not harm human subjects or participants, and there are no data-related concerns. In addition, we believe that there are no noteworthy potential harmful consequences of our research.

Guidelines:

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

10. Broader Impacts

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

Answer: [No]

Justification: This paper represents foundational research, and it is not tied to particular applications. In addition, we believe that there is no direct path to any negative applications.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

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

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

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

12. Licenses for existing assets

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

Answer: [NA]

Justification: The paper does not use existing assets.

Guidelines:

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

13. New Assets

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

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

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

14. Crowdsourcing and Research with Human Subjects

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

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

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

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

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

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

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