On the Sample Complexity of Inverse Reinforcement Learning

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

Alberto Maria Metelli
DEIB - Politecnico di Milano
Milan, Italy
albertomaria.metelli@polimi.it

Marcello Restelli
DEIB - Politecnico di Milano
Milan, Italy
marcello.restelli@polimi.it

Abstract

Inverse reinforcement learning (IRL) denotes a powerful family of algorithms for recovering a reward function justifying the behavior demonstrated by an expert agent. A well-known limitation of IRL is the *ambiguity* in the choice of the reward function, due to the existence of multiple rewards that explain the observed behavior. This limitation has been recently circumvented by formulating IRL as the problem of estimating the *feasible reward set*, i.e., the region of the rewards compatible with the expert's behavior. In this paper, we make a step towards closing the theory gap of IRL in the case of finite-horizon problems with a generative model. We start by formally introducing the problem of estimating the feasible reward set, the corresponding PAC requirement, and discussing the properties of particular classes of rewards. Then, we provide the first minimax lower bound on the sample complexity for the problem of estimating the feasible reward set of order $\Omega\left(\frac{H^3SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)\right)$, being S and A the number of states and actions respectively, H the horizon, ϵ the desired accuracy, and δ the confidence. We analyze the sample complexity of a uniform sampling strategy (`US-IRL`), proving a matching upper bound up to logarithmic factors. Finally, we outline several open questions in IRL and propose future research directions.

1 Introduction

Inverse reinforcement learning (IRL) aims at efficiently learning a desired behavior by observing an *expert* agent and inferring their intent encoded in a *reward function* (refer to [26, 3, 2] for recent surveys on IRL). This abstract setting, which diverges from standard *reinforcement learning* [RL, 34], as the reward function has to be learned, arises in a large variety of real-world tasks. In particular, in a *human-in-the-loop* [38] scenario, when the expert is represented by a human solving a task, an explicit specification of the reward function representing the human's goal is often unavailable. Experience suggests that humans are uncomfortable when asked to describe their intent and, thus, the underlying reward; while they are much more comfortable providing demonstrations of what is believed to be the right behavior. Indeed, human behavior is usually the product of many, possibly conflicting, objectives.[1] Succeeding in retrieving a representation of the expert's reward has notable

[1] In RL, the Sutton's hypothesis [34] conjectures that a scalar reward is an adequate notion of goal.

16th European Workshop on Reinforcement Learning (EWRL 2023).

implications [33, 40, 11, 39, 18]. First, we obtain explicit information for understanding the motivations behind the expert's choices (*interpretability*). Second, the reward can be employed in RL to train artificial agents, under shifts in the features of the underlying system (*transferability*).

Since the beginning, the community recognized that the IRL problem is, per se, *ill-posed*, as multiple reward functions are compatible with the expert's behavior [25]. This ambiguity was heterogeneously addressed by the algorithmic proposals that have followed over the years, which realized in several selection criteria, including maximum margin [30], maximum entropy [41], minimum Hessian eigenvalue [22, 23], and a balance between compatibility and learning efficiency [5]. Some of these approaches come with theoretical guarantees on the sample complexity, although according to different performance indices [e.g., 1, 35, 27].

A promising line of research that aspires to overcome the ambiguity issue has been recently investigated in [24, 19]. These works focus on estimating *all* the reward functions compatible with the expert's demonstrated behavior, namely the *feasible rewards*. Remarkably, this viewpoint which focuses on the *feasible reward set*, rather than on *one* reward obtained with a specific selection criterion, as previous works did, circumvents the ambiguity problem, postponing the reward selection and pointing to the expert's intent. Although these works provide sample complexity guarantees in different settings, a rigorous understanding of the inherent complexity of the IRL problem is currently lacking.

Contributions In this paper, we aim at taking a step toward the theoretical understanding of the IRL problem. As in [24, 19], we consider the problem of estimating the feasible reward set. We focus on a *generative model* setting, where the agent can query the environment and the expert in any state, and consider finite-horizon decision problems. The contributions of the paper can be summarized as follows.

- We propose a novel framework to evaluate the accuracy in recovering the feasible reward set, based on the *Hausdorff metric* [32]. This tool generalizes existing performance indices. Furthermore, we show that the feasible reward set enjoys a desirable Lipschitz continuity property w.r.t. the IRL problem (Section 3).
- We devise a PAC (Probability Approximately Correct) framework for estimating the feasible reward set, providing the definition of (ϵ, δ)-PAC IRL algorithm. Then, we investigate the relationships between several performance indices based on the Hausdorff metric (Section 4).
- We conceive, based on the provided PAC requirements introduced, a novel sample complexity *lower bound* of order $\Omega\left(\frac{H^3SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)\right)$. This represents the most significant contribution and, to the best of our knowledge, it is the first lower bound that values the importance of the relevant features of the IRL problem. From a technical perspective, the lower bound construction merges new proof ideas with reworks of existing techniques (Section 5).
- We analyze a uniform sampling exploration strategy (UniformSampling-IRL, `US-IRL`) showing that, in the generative model setting, it matches the lower bound up to logarithmic factors (Section 6).

The complete proofs of the results presented in the main paper are reported in Appendix B. A conference version of the present paper appeared in ICML 2023 [21].[2]

2 Preliminaries

In this section, we provide the background that will be employed in the subsequent sections.

Mathematical Background Let $a, b \in \mathbb{N}$ with $a \leqslant b$, we denote with $[\![a, b]\!] := \{a, \dots, b\}$ and with $[\![a]\!] := [\![1, a]\!]$. Let $\mathcal{X}$ be a set, we denote with $\Delta^{\mathcal{X}}$ the set of probability measures over $\mathcal{X}$. Let $\mathcal{Y}$ be a set, we denote with $\Delta^{\mathcal{X}}_{\mathcal{Y}}$ the set of functions with signature $\mathcal{Y} \rightarrow \Delta^{\mathcal{X}}$. Let $(\mathcal{X}, d)$ be a (pre)metric space, where $\mathcal{X}$ is a set and $d : \mathcal{X} \times \mathcal{X} \rightarrow [0, +\infty]$ is a (pre)metric.[3] Let $\mathcal{Y}, \mathcal{Y}' \subseteq \mathcal{X}$ be non-empty sets, we define the *Hausdorff (pre)metric* [32] $\mathcal{H}_d : 2^{\mathcal{X}} \times 2^{\mathcal{X}} \rightarrow [0, +\infty]$ between $\mathcal{Y}$ and $\mathcal{Y}'$ induced by the (pre)metric d as follows:

$$\mathcal{H}_d(\mathcal{Y}, \mathcal{Y}') := \max\left\{\sup_{y \in \mathcal{Y}} \inf_{y' \in \mathcal{Y}'} d(y, y'), \sup_{y' \in \mathcal{Y}'} \inf_{y \in \mathcal{Y}} d(y, y')\right\}. \tag{1}$$

[2] `https://proceedings.mlr.press/v202/metelli23a.html`.

[3] A *premetric* d satisfies the axioms: $d(x, x') \geqslant 0$ and $d(x, x) = 0$ for all $x, x' \in \mathcal{X}$. Any *metric* is clearly a premetric.

Markov Decision Processes without Reward A time-inhomogeneous finite-horizon *Markov decision process without reward* (MDP\R) is defined as a 4-tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, H)$ where $\mathcal{S}$ is a finite state space ($S = |\mathcal{S}|$), $\mathcal{A}$ is a finite action space ($A = |\mathcal{A}|$), $p = (p_h)_{h \in \llbracket H \rrbracket}$ is the transition model where for every stage $h \in \llbracket H \rrbracket$ we have $p_h \in \Delta_{\mathcal{S} \times \mathcal{A}}^{\mathcal{S}}$, and $H \in \mathbb{N}$ is the horizon. An MDP\R is time-homogeneous if, for every stage $h \in \llbracket H - 1 \rrbracket$, we have $p_h = p_{h+1}$ a.s.; in such a case, we denote the transition model with the symbol p only. A time-inhomogeneous reward function is defined as $r = (r_h)_{h \in \llbracket H \rrbracket}$, where for every stage $h \in \llbracket H \rrbracket$ we have $r_h : \mathcal{S} \times \mathcal{A} \rightarrow [-1, 1]$.[4] A *Markov decision process* [MDP, 28] is obtained by pairing an MDP\R $\mathcal{M}$ with a reward function r. The agent's behavior is modeled with a time-inhomogeneous policy $\pi = (\pi_h)_{h \in \llbracket H \rrbracket}$ where for every stage $h \in \llbracket H \rrbracket$, we have $\pi_h \in \Delta_{\mathcal{S}}^{\mathcal{A}}$. Let $f \in \mathbb{R}^{\mathcal{S}}$ and $g \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$, we denote with $p_h f(s, a) = \sum_{s' \in \mathcal{S}} p_h(s'|s, a) f(s')$ and with $\pi_h g(s) = \sum_{a \in \mathcal{A}} \pi_h(a|s) g(s, a)$ the expectation operators w.r.t. the transition model and the policy, respectively.

Value Functions and Optimality Given an MDP\R $\mathcal{M}$, a policy π, and a reward function r, the *Q-function* $Q^{\pi}(\cdot; r) = (Q_h^{\pi}(\cdot; r))_{h \in \llbracket H \rrbracket}$ induced by r represents the expected sum of rewards collected starting from $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$ and following policy π thereafter:

$$Q_h^{\pi}(s, a; r) := \mathop{\mathbb{E}}_{(\mathcal{M}, \pi)} \left[\sum_{l=h}^{H} r_l(s_l, a_l) | s_h = s, a_h = a \right],$$

where $\mathbb{E}_{(\mathcal{M}, \pi)}$ denotes the expectation w.r.t. $\mathcal{M}$ and π, i.e., $a_h \sim \pi_h(\cdot|s_h)$ and $s_{h+1} \sim p_h(\cdot|s_h, a_h)$ for every stage $h \in \llbracket h, H \rrbracket$. The Q-function fulfills the Bellman equations [28] for every $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$:

$$\begin{aligned} &Q_h^{\pi}(s, a; r) = r_h(s, a) + p_h V_{h+1}^{\pi}(s, a; r), \\ &V_h^{\pi}(s; r) = \pi_h Q_h^{\pi}(s; r) \quad \text{ and } \quad V_{H+1}^{\pi}(s; r) = 0, \end{aligned}$$

where $V^{\pi}(\cdot; r) = (V_h^{\pi}(\cdot; r))_{h \in \llbracket H \rrbracket}$ is the *V-function*. The *advantage function* $A_h^{\pi}(s, a; r) = Q_h^{\pi}(s, a; r) - V_h^{\pi}(s; r)$ represents the relative gain of playing action $a \in \mathcal{A}$ rather than following policy π in the state-stage pair (s, h). A policy π^* is *optimal* if it has non-positive advantage everywhere, i.e., $A_h^{\pi^*}(s, a; r) \leqslant 0$ for every $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$. The Q- and V-functions of an optimal policy are denoted with $Q_h^*(s, a; r)$ and $V_h^*(s; r)$.

Inverse Reinforcement Learning An *inverse reinforcement learning* problem [IRL, 25] is defined as a pair $(\mathcal{M}, \pi^E)$, where $\mathcal{M}$ is an MDP\R and π^E is an *expert's policy*. Informally, solving an IRL problem consists in finding a reward function $(r_h)_{h \in \llbracket H \rrbracket}$ making π^E optimal for the MDP\R $\mathcal{M}$ paired with reward function r. Any reward function fulfilling this condition is called *feasible* and the set of all such reward functions is called *feasible reward set* [24, 19], defined as:

$$\mathcal{R}_{(\mathcal{M},\pi^E)} := \left\{ (r_h)_{h \in \llbracket H \rrbracket} \middle| \forall h \in \llbracket H \rrbracket : r_h : \mathcal{S} \times \mathcal{A} \rightarrow [-1,1] \wedge \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket : A_h^{\pi^E}(s,a;r) \leqslant 0 \right\}. \quad (2)$$

We will omit the subscript $(\mathcal{M}, \pi^E)$ whenever clear from the context.

Empirical MDP and Empirical Expert's Policy Let $D = \{(s_l, a_l, h_l, s_l', a_l^E)\}_{l \in \llbracket t \rrbracket}$ be a dataset of $t \in \mathbb{N}$ tuples, where for every $l \in \llbracket t \rrbracket$, we have $s_l' \sim p_{h_l}(\cdot|s_l, a_l)$ and $a_l^E \sim \pi_{h_l}^E(\cdot|s_l)$. We introduce the counts for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times \llbracket H \rrbracket$: $n_h^t(s,a,s') := \sum_{l=1}^{t} \mathbb{1}\{(s_l, a_l, h_l, s_l') = (s,a,h,s')\}$, $n_h^t(s,a) := \sum_{s' \in \mathcal{S}} n_h^t(s,a,s')$, $n_h^t(s) := \sum_{a \in \mathcal{A}} n_h^t(s,a)$, and $n_h^{t,E}(s,a) := \sum_{l=1}^{t} \mathbb{1}\{(s_l, a_l^E) = (s,a)\}$. These quantities allow defining the *empirical transition model* $\widehat{p}^t = (\widehat{p}_h^t)_{h \in \llbracket H \rrbracket}$ and *empirical expert's policy* $\widehat{\pi}^{t,E} = (\pi_h^{t,E})_{h \in \llbracket H \rrbracket}$ as follows:

$$\widehat{p}_h^t(s'|s, a) := \begin{cases} \frac{n_h^t(s,a,s')}{n_h^t(s,a)} & \text{if } n_h^t(s, a) > 0 \\ \frac{1}{S} & \text{otherwise} \end{cases}, \qquad \widehat{\pi}_h^{E,t}(a|s) := \begin{cases} \frac{n_h^{E,t}(s,a)}{n_h^t(s)} & \text{if } n_h^t(s) > 0 \\ \frac{1}{A} & \text{otherwise} \end{cases}. \quad (3)$$

In the time-homogeneous case, we simply merge the samples collected at different stages $h \in \llbracket H \rrbracket$. We denote with $(\widehat{\mathcal{M}}^t, \widehat{\pi}^{E,t})$ the *empirical IRL* problem, where $\widehat{\mathcal{M}}^t = (\mathcal{S}, \mathcal{A}, \widehat{p}^t, H)$ the empirical MDP\R induced by $\widehat{p}^t$. Finally, we denote with $\widehat{\mathcal{R}}^t := \mathcal{R}_{(\widehat{\mathcal{M}}^t, \widehat{\pi}^{E,t})}$ the feasible reward set induced $(\widehat{\mathcal{M}}^t, \widehat{\pi}^{E,t})$. We will omit the superscript t, whenever clear from the context and write $\widehat{\mathcal{R}}$.

[4] For the sake of simplicity and w.l.o.g., we restrict to reward functions bounded by 1 in absolute value.

3 Lipschitz Framework for IRL

In this section, we analyze the regularity properties of the feasible reward set in terms of the Lipschitz continuity w.r.t. the IRL problem. To make the idea more concrete, suppose that $\mathcal{R}$ is the feasible reward set obtained from the IRL problem $(\mathcal{M}, \pi^E)$ and that $\widehat{\mathcal{R}}$ is obtained with a different IRL problem $(\widehat{\mathcal{M}}, \widehat{\pi}^E)$, which we can think to as an empirical version of $(\mathcal{M}, \pi^E)$, with an estimated transition model $\widehat{p}$ replacing the true model p. Intuitively, to have any learning guarantee, "similar" IRL problems ($p \approx \widehat{p}$ and $\pi^E \approx \widehat{\pi}^E$) should lead to "similar" feasible reward sets ($\mathcal{R} \approx \widehat{\mathcal{R}}$).[5]

To formally define a Lipschitz framework, we need to select a (pre)metric for evaluating dissimilarities between feasible reward sets and IRL problems. While we defer the presentation of the (pre)metric for the IRL problems to Section 3.1, where it will emerge naturally, for the feasible reward sets, we employ the *Hausdorff (pre)metric* $\mathcal{H}_d(\mathcal{R}, \widehat{\mathcal{R}})$ (Equation 1), induced by a (pre)metric $d(r, \widehat{r})$ used to evaluate the dissimilarity between individual reward functions $r \in \mathcal{R}$ and $\widehat{r} \in \widehat{\mathcal{R}}$. With this choice, two feasible reward sets are "similar" if every reward $r \in \mathcal{R}$ is "similar" to some reward $\widehat{r} \in \widehat{\mathcal{R}}$ in terms of the (pre)metric d. In the next sections, we employ as d the metric induced by the L_∞-norm between the reward functions $r \in \mathcal{R}$ and $\widehat{r} \in \widehat{\mathcal{R}}$:[6]

$$d^{\mathrm{G}}(r, \widehat{r}) := \max_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]} |r_h(s,a) - \widehat{r}_h(s,a)|, \tag{4}$$

where G stands for "generative". In Section 3.1, we prove that the Lipschitz continuity is fulfilled when no restrictions on the reward function are enforced (besides boundedness in $[-1, 1]$). In Appendix B.5, we show that, when further restrictions on the viable rewards are required (e.g., state-only reward), such a regularity property no longer holds.

3.1 Lipschitz Continuous Feasible Reward Sets

In order to prove the Lipschitz continuity property, we use the *explicit* form of the feasible reward sets introduced in [24] and extended by [19] for the finite-horizon case, that we report below.

Lemma 3.1 (Lemma 4 of [19]). *A reward function $r = (r_h)_{h \in [\![H]\!]}$ is feasible for the IRL problem $(\mathcal{M}, \pi^E)$ if and only if there exist two functions $(A_h, V_h)_{h \in [\![H]\!]}$ where for every $h \in [\![H]\!]$ we have $A_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_{\geqslant 0}$, $V_h : \mathcal{S} \to \mathbb{R}$, and $V_{H+1} = 0$, such that for every $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]$ it holds that:*

$$r_h(s,a) = -A_h(s,a)\mathbb{1}_{\{\pi_h^E(a|s)=0\}} + V_h(s) - p_h V_{h+1}(s,a).$$

Furthermore, if $|r_h(s,a)| \leqslant 1$, if follows that $|V_h(s)| \leqslant H - h + 1$ and $A_h(s,a) \leqslant H - h + 1$.

A form of regularity of the feasible reward set was already studied in Theorem 3.1 of [24] and in Theorem 5 of [19], providing an *error propagation* analysis. These results are based on showing the existence of a *particular* reward $\widetilde{r}$ feasible for the IRL problem $(\widehat{\mathcal{M}}, \widehat{\pi}^E)$, whose distance from the original reward function $r \in \mathcal{R}$ is bounded by a dissimilarity term between $(\mathcal{M}, \pi^E)$ and $(\widehat{\mathcal{M}}, \widehat{\pi}^E)$. Unfortunately, such a reward $\widetilde{r}$ is not guaranteed to be bounded in $[-1, 1]$ even when the original reward r is (and, thus, it might be $\widetilde{r} \notin \widehat{R}$ according to Equation 2).[7] In Lemma B.1, with a modified construction, we show the existence of another *particular* feasible reward $\widehat{r}$ bounded in $[-1, 1]$ (and, thus, $\widehat{r} \in \widehat{\mathcal{R}}$). From this, the Lipschitz continuity of the feasible reward sets follows.

Theorem 3.2 (Lipschitz Continuity). *Let $\mathcal{R}$ and $\widehat{\mathcal{R}}$ be the feasible reward sets of the IRL problems $(\mathcal{M}, \pi^E)$ and $(\widehat{\mathcal{M}}, \widehat{\pi}^E)$, as in Equation (2). Then, it holds that:*[8]

$$\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}, \widehat{\mathcal{R}}) \leqslant \frac{2\rho^{\mathrm{G}}((\mathcal{M}, \pi^E), (\widehat{\mathcal{M}}, \widehat{\pi}^E))}{1 + \rho^{\mathrm{G}}((\mathcal{M}, \pi^E), (\widehat{\mathcal{M}}, \widehat{\pi}^E))}, \tag{5}$$

[5] If not, any arbitrary accurate estimate $(\widehat{p}, \widehat{\pi}^E)$ of (p, π^E), may induce feasible sets $\widehat{\mathcal{R}}$ and $\mathcal{R}$ with finite non-zero dissimilarity.

[6] We discuss other choices of d in Section 4.

[7] We illustrate in Fact B.1 an example of this phenomenon.

[8] This implies the standard Lipschitz continuity, by simply bounding $\frac{2\rho^{\mathrm{G}}((\mathcal{M},\pi^E),(\widehat{\mathcal{M}},\widehat{\pi}^E))}{1+\rho^{\mathrm{G}}((\mathcal{M},\pi^E),(\widehat{\mathcal{M}},\widehat{\pi}^E))} \leqslant 2\rho^{\mathrm{G}}((\mathcal{M}, \pi^E), (\widehat{\mathcal{M}}, \widehat{\pi}^E))$.

where $\rho^G(\cdot,\cdot)$ is a (pre)metric between IRL problems, defined as:

$$\rho^G((\mathcal{M},\pi^E),(\widehat{\mathcal{M}},\widehat{\pi}^E)):=\max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]}(H-h+1)\left(\left|\mathbb{1}_{\{\pi_h^E(a|s)=0\}}-\mathbb{1}_{\{\widehat{\pi}_h^E(a|s)=0\}}\right|+\|p_h(\cdot|s,a)-\widehat{p}_h(\cdot|s,a)\|_1\right).$$

Some observations are in order. First, the function ρ^{G} is indeed a (pre)metric since it is non-negative and takes value 0 when the IRL problems coincide. Second, as supported by intuition, ρ^{G} is composed of two terms related to the estimation of the expert's policy and of the transition model. While for the transition model, the dissimilarity is formalized by the L_1-norm distance $\|p_h(\cdot|s,a)-\widehat{p}_h(\cdot|s,a)\|_1$, for the policy, the resulting term deserves some comments. Indeed, the dissimilarity $|\mathbb{1}_{\{\pi_h^E(a|s)=0\}}-\mathbb{1}_{\{\widehat{\pi}_h^E(a|s)=0\}}|$ highlights that what matters is *whether an action $a\in\mathcal{A}$ is played by the expert* and not the corresponding probability $\pi_h^E(a|s)$. Indeed, the expert's policy plays an action (with any non-zero probability) only if it is an optimal action.

4 PAC Framework for IRL with a Generative Model

In this section, we discuss the PAC (Probably Approximately Correct) requirements for estimating the feasible reward set with access to a *generative model* of the environment. We first provide the notion of a learning algorithm estimating the feasible reward set with a generative model (Section 4.1). Then, we formally present the PAC requirement for the Hausdorff (pre)metric $\mathcal{H}_d$ (Section 4.2). Finally, we discuss the relationships between the PAC requirements with different choices of (pre)metric d (Section 4.3).

4.1 Learning Algorithms with a Generative Model

A learning algorithm for estimating the feasible reward set is a pair $\mathfrak{A}=(\mu,\tau)$, where $\mu=(\mu_t)_{t\in\mathbb{N}}$ is a *sampling strategy* defined for every time step $t\in\mathbb{N}$ as $\mu_t\in\Delta_{\mathcal{D}_{t-1}}^{\mathcal{S}\times\mathcal{A}\times[\![H]\!]}$ with $\mathcal{D}_t=(\mathcal{S}\times\mathcal{A}\times[\![H]\!]\times\mathcal{S}\times\mathcal{A})^t$ and τ is a stopping time w.r.t. a suitably defined filtration. At every step $t\in\mathbb{N}$, the learning algorithm query the environment in a triple (s_t,a_t,h_t), selected based on the sampling strategy $\mu_t(\cdot|D_{t-1})$, where $D_{t-1}=((s_l,a_l,h_l,s'_l,a_l^E))_{l=1}^{t-1}\in\mathcal{D}_{t-1}$ is the dataset of past samples. Then, the algorithm observes the next state $s'_t\sim p_{h_t}(\cdot|s_t,a_t)$ and expert's action $a_t^E\sim\pi_{h_t}^E(\cdot|s_t)$ and updates the dataset $D_t=D_{t-1}\oplus(s_t,a_t,h_t,s'_t,a_t^E)$. Based on the collected data D_τ, the algorithm computes the empirical IRL problem $(\widehat{M}^\tau,\widehat{\pi}^{E,\tau})$, based on Equation (3) and the empirical feasible reward set $\widehat{\mathcal{R}}^\tau$.

4.2 PAC Requirement

We now introduce a general notion of a PAC requirement for estimating the feasible reward set of an IRL problem. To this end, we consider the Hausdorff (pre)metric introduced in Section 3 defined in terms of the reward (pre)metric $d(r,\widehat{r})$. We denote with d-IRL the problem of estimating the feasible reward set under the Hausdorff (pre)metric $\mathcal{H}_d$.

Definition 4.1 (PAC Algorithm for d-IRL). *Let $\epsilon\in(0,2)$ and $\delta\in(0,1)$. An algorithm $\mathfrak{A}=(\mu,\tau)$ is (ϵ,δ)-PAC for d-IRL if:*

$$\mathop{\mathbb{P}}_{(\mathcal{M},\pi^E),\mathfrak{A}}\left(\mathcal{H}_d(\mathcal{R},\widehat{\mathcal{R}}^\tau)\leqslant\epsilon\right)\geqslant 1-\delta,$$

where $\mathbb{P}_{(\mathcal{M},\pi^E),\mathfrak{A}}$ denotes the probability measure induced by executing the algorithm $\mathfrak{A}$ in the IRL problem $(\mathcal{M},\pi^E)$ and $\widehat{\mathcal{R}}^\tau$ is the feasible reward set induced by the empirical IRL problem $(\widehat{\mathcal{M}}^\tau,\widehat{\pi}^{E,\tau})$ estimated with the dataset D_τ. The sample complexity *is defined as $\tau:=|D_\tau|$.*

In the next section, we show the relationship between PAC requirements defined for notable choices of d.

4.3 Different Choices of d

So far, we have evaluated the dissimilarity between the feasible reward sets by means of the Hausdorff induced by d^{G}, i.e., the L_∞-norm of between individual reward functions. In the literature, other (pre)metrics d have been proposed [e.g., 24, 19].

d^{G}_{Q*}**-IRL** Since the recovered reward functions are often used for performing forward RL, an index of interest is the dissimilarity between optimal Q-functions obtained with the reward $r \in \mathcal{R}$ and $\widehat{r} \in \widehat{\mathcal{R}}$ in the original MDP\R:

$$d^{\mathrm{G}}_{Q*}(r, \widehat{r}) := \max_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]} \left| Q^*_h(s, a; r) - Q^*_h(s, a; \widehat{r}) \right| .$$

d^{G}_{V*}**-IRL** We are often interested in not just being accurate in estimating the optimal Q-function, but rather in the performance of an optimal policy $\widehat{\pi}^*$, learned with the recovered reward $\widehat{r} \in \widehat{\mathcal{R}}$, evaluated under the true reward $r \in \mathcal{R}$:

$$d^{\mathrm{G}}_{V*}(r, \widehat{r}) := \sup_{\widehat{\pi}^* \in \Pi^*(\widehat{r})} \max_{(s,h) \in \mathcal{S} \times [\![H]\!]} \left| V^*_h(s; r) - V^{\widehat{\pi}^*}_h(s; r) \right| ,$$

where $\Pi^*(\widehat{r}) := \{\pi : \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!] : A^{\pi}_h(s,a;\widehat{r}) \leqslant 0\}$ is the set of optimal policies under the recovered reward $\widehat{r}$.

The following result formalizes the relationships between the presented d-IRL problems.

Theorem 4.1 (Relationships between d-IRL problems)**.** *Let us introduce the graphical convention for* $c > 0$*:*

$$\begin{tikzcd} \boxed{x\text{-}IRL} \arrow[r, "c"] & \boxed{y\text{-}IRL} \end{tikzcd}$$

meaning that any (ϵ, δ)*-PAC* x*-IRL algorithm is* $(c\epsilon, \delta)$*-PAC* y*-IRL. Then, the following statements hold:*

$$\begin{tikzcd} \boxed{d^{\mathrm{G}}\text{-}IRL} \arrow[r, "H"] \arrow[rr, bend left, "2H"] & \boxed{d^{\mathrm{G}}_{Q*}\text{-}IRL} \arrow[r, "2H"] & \boxed{d^{\mathrm{G}}_{V*}\text{-}IRL} \end{tikzcd} .$$

Theorem 4.1 shows that any (ϵ, δ)-PAC guarantee on d^{G}, implies (ϵ', δ)-PAC guarantees on both d^{G}_{Q*} and d^{G}_{V*}, where $\epsilon' = \Theta(H\epsilon)$ is linear in the horizon H. This justifies why focusing on d^{G}-IRL, as in the following section where sample complexity lower bounds are derived. The lower bound analysis for d^{G}_{Q*}-IRL and d^{G}_{V*}-IRL is left to future works.

5 Lower Bounds

In this section, we establish sample complexity lower bounds for the d^{G}-IRL problem based on the PAC requirement of Definition 4.1 in the generative model setting. We start presenting the general result (Section 5.1) and, then, we comment on its form and, subsequently, provide a sketch of the construction of the hard instances for obtaining the lower bound (Section 5.2). For the sake of presentation, we assume that the expert's policy π^E is known; the extension to the case of unknown π^E is reported in Appendix C.

5.1 Main Result

In this section, we report the main result of the lower bound of the sample complexity of learning the feasible reward set.

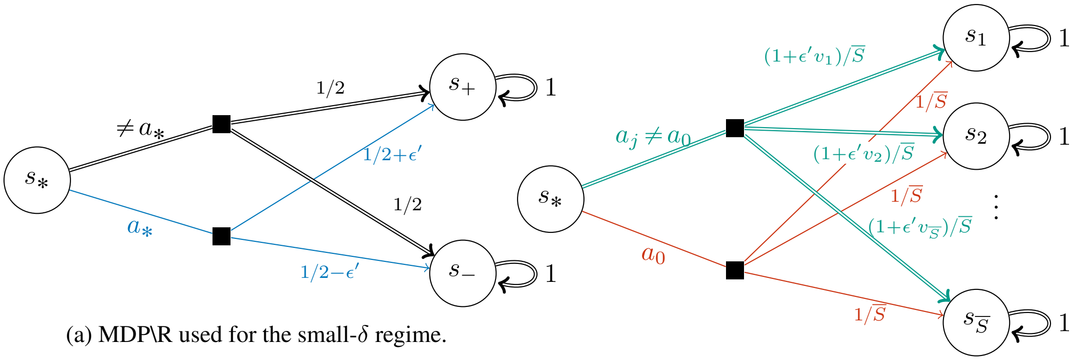
Theorem 5.1 (Lower Bound for d^{G}-IRL)**.** *Let* $\mathfrak{A} = (\mu, \tau)$ *be an* (ϵ, δ)*-PAC algorithm for* d^{G}*-IRL. Then, there exists an IRL problem* $(\mathcal{M}, \pi^E)$ *such that, if* $\epsilon \leqslant 1/64$*,* $\delta \leqslant 1/32$*,* $S \geqslant 9$*,* $A \geqslant 2$*, and* $H \geqslant 12$*, the expected sample complexity is lower bounded by:*

- *if the transition model* p *is time-inhomogeneous:*

$$\mathop{\mathbb{E}}_{(\mathcal{M}, \pi^E), \mathfrak{A}} [\tau] \geqslant \Omega\left(\frac{H^3 SA}{\epsilon^2} \left(\log\left(\frac{1}{\delta}\right) + S \right) \right);$$

- *if the transition model* p *is time-homogeneous:*

$$\mathop{\mathbb{E}}_{(\mathcal{M}, \pi^E), \mathfrak{A}} [\tau] \geqslant \Omega\left(\frac{H^2 SA}{\epsilon^2} \left(\log\left(\frac{1}{\delta}\right) + S \right) \right),$$



(a) MDP\R used for the small-δ regime.

(b) MDP\R used for the large-δ regime.

Figure 1: The MDP\R employed in the constructions of the lower bounds of Section 5. The expert's policy is $\pi^E(s) = a_0$. $\Longrightarrow$ denotes a transition executed for multiple actions.

where $\mathbb{E}_{(\mathcal{M},\pi^E),\mathfrak{A}}$ *denotes the expectation w.r.t. the probability measure* $\mathbb{P}_{(\mathcal{M},\pi^E),\mathfrak{A}}$.

Some observations are in order. First, the derived lower bound displays a linear dependence on the number of actions A and dependence on the horizon H raised to a power 2 or 3, which depends on whether the underlying transition model is time-homogeneous, as common even for forward RL [e.g., 6, 8]. Second, we identify two different regimes visible inside the parenthesis related to the dependence on the number of states S and the confidence δ. Specifically, for small values of δ (i.e., $\delta \approx 0$), the dominating part is $\log\left(\frac{1}{\delta}\right)$, leading to a sample complexity of order $\Omega\left(\frac{H^3 SA}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$. Instead, for large δ (i.e., $\delta \approx 1/32$), the most relevant part is the one corresponding to S, leading to sample complexity of order $\Omega\left(\frac{H^3 S^2 A}{\epsilon^2}\right)$ (both for the time-inhomogeneous case). An analogous two-regime behavior has been previously observed in the reward-free exploration setting [12, 14, 20].

5.2 Sketch of the Proof

In this section, we provide a sketch of the construction of the lower bounds of Theorem 5.1. The idea consists in deriving two separate bounds depending on the regime of δ, which are based on two building blocks reported in Figure 1. These instances are used to build lower bounds for a single state s_* and the extension to multiple states and stages follows standard constructions [e.g., 8].

Small-δ regime Figure 1a reports the instances employed in this regime. The expert's policy is $\pi^E(s) = a_0$. From state s_*, all actions bring the system to the absorbing states s_+ and s_- with equal probability, except for action $a_* \neq a_0$ that increases by $\epsilon' > 0$ the probability of reaching state s_+. The learner, in order to recover a correct feasible reward set, has to identify which is the action behaving like a_* (among the A available ones) to force action a_0 to be optimal. Considering $\Theta(A)$ instances, in which action a_* changes, an application of *Bretagnolle-Huber inequality* [17, Theorem 14.2] allows deriving a sample complexity lower bounded by $\Omega\left(\frac{AH^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$.

Large-δ regime Figure 1b depicts the instances used in this regime. The expert's policy is again $\pi^E(s) = a_0$. The system, instead, is made of $\overline{S} = \Theta(S)$ next states reachable with equal probability by playing action a_0. All other actions $a_j \neq a_0$ alter the probability distribution of the next state. Specifically, by playing the action $a_j \neq a_0$, the probability of reaching the next state s'_k is given by $(1+\epsilon' v_k^{(j)})/\overline{S}$, where $v^{(j)} \in \{-1,1\}^{\overline{S}}$ is a vector such that $\sum_{k=1}^{\overline{S}} v_k^{(j)} = 0$. By varying v_j in a suitable set, defined by means of novel packing argument based on Hamming coding (Lemma D.6), we obtain $\Theta(2^{\overline{S}})$ instances each one separated by a finite dissimilarity, depending on ϵ'. We obtain the lower bound by means of an application of the *Fano's inequality* [9, Proposition 4] which results in order $\Omega\left(\frac{((1-\delta)-\log 2)S^2 AH^2}{\epsilon^2}\right)$.

Extension to Multiple States and Stages At the beginning, the system randomly chooses a problem between Figure 1a and Figure 1b. Then, it transitions to the state in which the system may randomly

```
Input: significance $\delta \in (0,1)$, $\epsilon$ target accuracy
$t \leftarrow 0$, $\epsilon_0 \leftarrow +\infty$
while $\epsilon_t > \epsilon$ do
    $t \leftarrow t + SAH$
    Collect one sample from each $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]$
    Update $\widehat{p}^t$ according with (3)
    Update $\epsilon_t = \max_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]} \mathcal{C}_h^t(s,a)$ (resp. $\widetilde{\mathcal{C}}_h^t(s,a)$)
end while
```

Algorithm 1: UniformSampling-IRL (US-IRL) for time-inhomogeneous (resp. time-homogeneous) transition models.

remain for $\overline{H} < H$ stages after which it transitions with uniform probability to any of the $\Theta(S)$ states. Our approach allows employing a single construction for both the time-inhomogeneous and time-homogeneous settings, depending on the value of $\overline{H}$. Specifically, we select $\overline{H} = \Theta(H)$ for the time-inhomogeneous case and $\overline{H} = O(1)$ for the time-homogeneous case. In any state s_* and stage h_*, the agent can face the problems shown in Figure 1. By varying s_* and h_* among its possible HS (resp. S) values, we get the bounds in Theorem 5.1.

Remark 5.1 (Generative vs Forward models)**.** *This construction suffices for obtaining a bound for the generative model, but it can be easily extended to work with the* forward model *of the environment (in which the agent interacts via trajectories only) by means of a standard* tree-based construction *[12, 8]. In such a case, the resulting PAC guarantee would no longer be expressed via the* $L\infty$*-norm distance* d^G *between reward, but* worst-case *over the visitation distributions induced by the policies:* $d^F(r,\widehat{r}) := \sup_\pi \mathbb{E}_{\mathcal{M},\pi}[|r_h(s,a) - \widehat{r}_h(s,a)|]$.

6 Algorithm

In this section, we analyze the sample complexity of a uniform sampling strategy (UniformSampling-IRL, US-IRL) for the d^G-IRL problem (Algorithm 1). We start presenting the sample complexity analysis (Section 6.1) and, then, we provide a sketch of the proof (Section 6.2).

6.1 Main Result

The US-IRL algorithm was presented in [24, 19] but analyzed for different IRL formulations (see Section A). We revise it since it matches our sample complexity lower bounds, provided that more sophisticated concentration tools w.r.t. those employed in [24, 19]. For the sake of presentation, we assume that the expert's policy π^E is known; the extension to unknown π^E is reported in Appendix C. At each iteration, the algorithm collects a sample from every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]$ and, for time-inhomogeneous models, computes the confidence function:

$$\mathcal{C}_h^t(s,a) := 2\sqrt{2}(H-h+1)\sqrt{\frac{2\beta\big(n_h^t(s,a),\delta\big)}{n_h^t(s,a)}}, \tag{6}$$

where $\beta\big(n,\delta\big) := \log(SAH/\delta) + (S-1)\log\big(e(1+n/(S-1)\big)$.[9] The algorithm stops as soon as all confidence functions fall below the threshold ϵ. The following theorem provides the sample complexity of US-IRL.

Theorem 6.1 (Sample Complexity of US-IRL)**.** *Let* $\epsilon > 0$ *and* $\delta \in (0,1)$*, US-IRL is* (ϵ,δ)*-PAC for* d^G*-IRL and with probability at least* $1-\delta$ *it stops after* τ *samples with:*

[9] In the time-homogeneous case, the algorithm merges the samples collected at different $h \in [\![H]\!]$ for the estimation of the transition model and replaces the confidence function with:

$$\widetilde{\mathcal{C}}_h^t(s,a) := 2\sqrt{2}(H-h+1)\sqrt{\frac{2\widetilde{\beta}\big(n^t(s,a),\delta\big)}{n^t(s,a)}}, \tag{7}$$

where $\widetilde{\beta}\big(n,\delta\big) := \log(SA/\delta) + (S-1)\log\big(e(1+n/(S-1)\big)$ and $n^t(s,a) = \sum_{h=1}^H n_h^t(s,a)$.

- *if the transition model* p *is time-inhomogeneous:*

$$\tau \leqslant \frac{8H^3SA}{\epsilon^2}\left(\log\left(\frac{SAH}{\delta}\right)+(S-1)C\right),$$

where $C = 1+\log(1+(64H^4)/(\epsilon^4(S-1))\times\big(\log((SAH)/\delta)+\sqrt{e}(S-1+\sqrt{S-1}))^2\big)$*;*

- *if the transition model* p *is time-homogeneous:*

$$\tau \leqslant \frac{8H^2SA}{\epsilon^2}\left(\log\left(\frac{SA}{\delta}\right)+(S-1)\widetilde{C}\right),$$

where $\widetilde{C} = 1+\log(1+(64H^4)/(\epsilon^4(S-1))\times\big(\log((SA)/\delta)+\sqrt{e}(S-1+\sqrt{S-1}))^2\big)$.

Thus, time-inhomogeneous (resp. time-homogeneous) transition models, `US-IRL` suffers a sample complexity bound of order $\widetilde{O}\left(\frac{H^3SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)\right)$ (resp. $\widetilde{O}\left(\frac{H^2SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)\right)$) matching the lower bounds of Theorem 5.1 up to logarithmic factors for both regimes of δ.

6.2 Sketch of the Proof

The idea of the proof is to exploit Theorem 3.2 to reduce the Hausdorff distance to the L_1-norm between the transition model $\|\widehat{p}_h^t(\cdot|s,a)-p_h(\cdot|s,a)\|_1$. It is worth noting this term replaces $|(\widehat{p}_h^t - p_h)V_h|$ appearing in previous works [24, 19] that was comfortably bounded using Höeffding's inequality. In our case, the L_1-norm is unavoidable due to the Hausdorff distance that implies a worst-case choice of the reward function and, thus, of V_h. This term has to be carefully bounded using the stronger KL-divergence concentration result of [13, Proposition 1] to get the $O(\log(1/\delta)+S)$ rate.[10]

7 Conclusions and Open Questions

In this paper, we provided contributions to the understanding of the complexity of the IRL problem. We conceived a lower bound of order $\Omega\left(\frac{H^3SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)\right)$ on the number samples collected with a generative model in the finite-horizon setting. This result is of relevant interest since it sets, for the first time, the complexity of the IRL problem, defined as the problem of estimating the feasible reward set. Furthermore, we showed that a uniform sampling strategy matches the lower bound up to logarithmic factors. Nevertheless, the IRL problem is far from being closed. In the following, we outline a road map of open questions, hoping to inspire researchers to work in this appealing area.

Forward Model The most straightforward extension of our findings is moving to the *forward model* setting, in which the agent can interact with the environment through trajectories only. As we already noted, our lower bounds can be comfortably extended to this setting. However, in this case, the PAC requirement has to be relaxed since controlling the L_∞-norm between rewards is no longer a viable option (e.g., for the possible presence of almost unreachable states). Which distance notion should be used for this setting? Will the Lipschitz regularity of Section 3 still hold?

Problem-Dependent Analysis Our analysis is *worst-case* in the class of IRL problems. Would it be possible to obtain a *problem-dependent* complexity results? Previous problem-dependent analyses provided results tightly connected to the properties of the specific reward selection procedure [24, 19]. Clearly, a currently open question, in all settings in which reward is missing, including reward-free exploration [12] and IRL, is how to define a problem-dependent quantity in replacement of the suboptimality gaps.

Reward Selection Our PAC guarantees concern with the complete feasible reward set. However, algorithmic solutions to IRL implement a specific criterion for selecting a reward (e.g., maximum entropy, maximum margin). How the PAC guarantee based on the Hausdorff distance relates to guarantees on a single reward selected with a *specific criterion* within $\mathcal{R}$?

Acknowledgements

This paper is supported by PNRR-PE-AI FAIR project funded by the NextGeneration EU program.

[10] A more naïve application of the L_1-concentration of [37] would lead to the worse $O(S\log(1/\delta))$ rate.

References

[1] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-first International Conference on Machine Learning (ICML)*, volume 69 of *ACM International Conference Proceeding Series*. ACM, 2004.

[2] Stephen C. Adams, Tyler Cody, and Peter A. Beling. A survey of inverse reinforcement learning. *Artif. Intell. Rev.*, 55(6):4307–4346, 2022.

[3] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artif. Intell.*, 297:103500, 2021.

[4] Gérard D. Cohen and Peter Frankl. Good coverings of hamming spaces with spheres. *Discret. Math.*, 56(2-3):125–131, 1985.

[5] Angelo Damiani, Giorgio Manganini, Alberto Maria Metelli, and Marcello Restelli. Balancing sample efficiency and suboptimality in inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 4618–4629. PMLR, 2022.

[6] Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 2818–2826, 2015.

[7] 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.

[8] Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory (ALT)*, volume 132 of *Proceedings of Machine Learning Research*, pages 578–598. PMLR, 2021.

[9] Sebastien Gerchinovitz, Pierre Ménard, and Gilles Stoltz. Fano's inequality for random variables. *Statistical Science*, 35(2):178–201, 2020.

[10] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer series in statistics. Springer, 2002.

[11] Mahdi Imani and Ulisses M. Braga-Neto. Control of gene regulatory networks using bayesian inverse reinforcement learning. *IEEE ACM Trans. Comput. Biol. Bioinform.*, 16(4):1250–1261, 2019.

[12] Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 4870–4879. PMLR, 2020.

[13] Anders Jonsson, Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Edouard Leurent, and Michal Valko. Planning in markov decision processes with gap-dependent sample complexity. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[14] Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. In *Algorithmic Learning Theory (ALT)*, volume 132 of *Proceedings of Machine Learning Research*, pages 865–891. PMLR, 2021.

[15] Abi Komanduru and Jean Honorio. On the correctness and sample complexity of inverse reinforcement learning. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 7110–7119, 2019.

[16] Abi Komanduru and Jean Honorio. A lower bound for the sample complexity of inverse reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 5676–5685. PMLR, 2021.

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

[18] Amarildo Likmeta, Alberto Maria Metelli, Giorgia Ramponi, Andrea Tirinzoni, Matteo Giuliani, and Marcello Restelli. Dealing with multiple experts and non-stationarity in inverse reinforcement learning: an application to real-life problems. *Mach. Learn.*, 110(9):2541–2576, 2021.

[19] David Lindner, Andreas Krause, and Giorgia Ramponi. Active exploration for inverse reinforcement learning. *CoRR*, abs/2207.08645, 2022.

[20] Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 7599–7608. PMLR, 2021.

[21] Alberto Maria Metelli, Filippo Lazzati, and Marcello Restelli. Towards theoretical understanding of inverse reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 24555–24591. PMLR, 23–29 Jul 2023.

[22] Alberto Maria Metelli, Matteo Pirotta, and Marcello Restelli. Compatible reward inverse reinforcement learning. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 2050–2059, 2017.

[23] Alberto Maria Metelli, Matteo Pirotta, and Marcello Restelli. On the use of the policy gradient and hessian in inverse reinforcement learning. *Intelligenza Artificiale*, 14(1):117–150, 2020.

[24] Alberto Maria Metelli, Giorgia Ramponi, Alessandro Concetti, and Marcello Restelli. Provably efficient learning of transferable rewards. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 7665–7676. PMLR, 2021.

[25] Andrew Y. Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pages 663–670. Morgan Kaufmann, 2000.

[26] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J. Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *Found. Trends Robotics*, 7(1-2):1–179, 2018.

[27] Matteo Pirotta and Marcello Restelli. Inverse reinforcement learning through policy gradient minimization. In *Proceedings of the Thirtieth Conference on Artificial Intelligence (AAAI)*, pages 1993–1999. AAAI Press, 2016.

[28] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1994.

[29] Giorgia Ramponi, Amarildo Likmeta, Alberto Maria Metelli, Andrea Tirinzoni, and Marcello Restelli. Truly batch model-free inverse reinforcement learning about multiple intentions. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 2359–2369. PMLR, 2020.

[30] Nathan D. Ratliff, J. Andrew Bagnell, and Martin Zinkevich. Maximum margin planning. In *Proceedings of the Twenty-Third International Conference on Machine Learning (ICML)*, volume 148 of *ACM International Conference Proceeding Series*, pages 729–736. ACM, 2006.

[31] Alex Ravsky. How to prove upper bound for partial sum of binomial coefficients (version: 2018-12-17). Mathematics Stack Exchange, 2018.

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

[33] Sahand Sharifzadeh, Ioannis Chiotellis, Rudolph Triebel, and Daniel Cremers. Learning to drive using inverse reinforcement learning and deep q-networks. *CoRR*, abs/1612.03653, 2016.

[34] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[35] Umar Syed and Robert E. Schapire. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems 20 (NeurIPS)*, pages 1449–1456. Curran Associates, Inc., 2007.

[36] Monica C Vroman. *Maximum likelihood inverse reinforcement learning*. Rutgers The State University of New Jersey-New Brunswick, 2014.

[37] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

[38] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. A survey of human-in-the-loop for machine learning. *Future Gener. Comput. Syst.*, 135:364–381, 2022.

[39] Zheng Wu, Liting Sun, Wei Zhan, Chenyu Yang, and Masayoshi Tomizuka. Efficient sampling-based maximum entropy inverse reinforcement learning with application to autonomous driving. *IEEE Robotics Autom. Lett.*, 5(4):5355–5362, 2020.

[40] Chao Yu, Guoqi Ren, and Jiming Liu. Deep inverse reinforcement learning for sepsis treatment. In *2019 IEEE International Conference on Healthcare Informatics, ICHI 2019, Xi'an, China, June 10-13, 2019*, pages 1–3. IEEE, 2019.

[41] Siliang Zeng, Chenliang Li, Alfredo Garcia, and Mingyi Hong. Maximum-likelihood inverse reinforcement learning with finite-time guarantees. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[42] Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the Twenty-Third Conference on Artificial Intelligence (AAAI)*, pages 1433–1438. AAAI Press, 2008.

Appendix

A Related Works

In this appendix, we discuss the related works about sample complexity analysis, lower bounds for IRL, sample complexity analysis for specific IRL algorithms, and reward-free exploration.

Sample Complexity for Estimating the Feasible Reward Set The notion of feasible reward set $\mathcal{R}$ was introduced in [25] in an *implicit* form in the infinite-horizon discounted case as a *linear feasibility* problem and, subsequently, adapted to the finite-horizon case in [19]. Furthermore, in [24, 19] an *explicit* form of the reward functions belonging to the feasible region $\mathcal{R}$ was provided. In these works, the problem of estimating the feasible reward set is studied for the first time considering a "reference" pair of rewards $(\overline{r},\breve{r})\in\mathcal{R}\times\widehat{\mathcal{R}}$ against which to compare the rewards inside the recovered sets, leading to the (pre)metric:

$$\widetilde{\mathcal{H}}_d(\mathcal{R},\mathcal{R},\overline{r},\breve{r}) := \max\left\{\inf_{\widehat{r}\in\widehat{\mathcal{R}}} d(\overline{r},\widehat{r}), \inf_{r\in\mathcal{R}} d(r,\breve{r})\right\}. \tag{8}$$

Compared to the Hausdorff (pre)metric (Equation 1), in Equation (8) there is no maximization over the choice of $(\overline{r},\breve{r})$, leading to a simpler problem.[11] In [24], a uniform sampling approach (similar to Algorithm 1) is proved to achieve a sample complexity of order $\widetilde{O}\left(\frac{\gamma^2 SA}{(1-\gamma)^4\epsilon^2}\right)$ for the index of Equation (8) with $d=d^{\mathrm{G}}_{Q*}$ in the discounted setting with generative model. For the forward model case, the `AceIRL` algorithm [19] suffers a sample complexity of order $\widetilde{O}\left(\frac{H^5 SA}{\epsilon^2}\right)$ for the index of Equation (8) with $d=d^{\mathrm{F}}_{V*}$, in the finite-horizon case.[12] Unfortunately, the reward recovered by `AceIRL` reward function is not guaranteed to be bounded by a predetermined constant (e.g., $[-1,1]$). Modified versions of these algorithms allow embedding problem-dependent features under a specific choice of a reward within the set.

Sample Complexity Lower Bounds in IRL To the best of our knowledge, the only work that proposes a sample complexity lower bound for IRL is [16]. The authors consider a finite state and action MDP\R and the IRL algorithm of [25] for β-strict separable IRL problems (i.e., with suboptimality gap at least β) with state-only rewards in the discounted setting. When only two actions are available ($A=2$) and the samples are collected starting in each state with equal probability, by means of a geometric construction and Fano's inequality, the authors derive an $\Omega(S\log S)$ lower bound on the number of trajectories needed to identify a reward function. Note that this analysis limits to the *identification* of a reward function within a finite set, rather than evaluating the accuracy of recovering the feasible reward set.

Sample Complexity of IRL Algorithms Differently from forward RL, the theoretical understanding of the IRL problem is largely less established and the sample complexity analysis proposed in the literature often limit to specific algorithms. In the class of *feature expectation* approaches, the seminal work [1] propose IRL algorithms guaranteed to output an ϵ-optimal policy (made of a mixture of Markov policies) after $\widetilde{O}\left(\frac{k}{\epsilon^2(1-\gamma)^2}\log\left(\frac{1}{\delta}\right)\right)$ trajectories (ideally of infinite length). The result holds in a discounted setting (being γ the discount factor) under the assumption that the true reward function $r(s)=w^T\phi(s)$ is state-only and linear in some *known* features ϕ of dimensionality k. In [35], a game-theoretic approach to IRL, named `MWAL`, is proposed improving [1] in terms of computational complexity and allowing the absence of an expert, preserving similar theoretical guarantees in the same setting. `Modular IRL` [36], that integrates supervised learning capabilities in the IRL algorithm, is guaranteed to produce an ϵ-optimal policy after $\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ trajectories. This class of algorithms, however, requires, as an inner step, to compute the optimal policy $\widehat{\pi}$ for every candidate reward function $\widehat{r}$. This step (and the corresponding sample complexity) is somehow hidden in the analysis since they either assume the knowledge of the transition model and apply dynamic

[11] In this sense, a PAC guarantee according to Definition 4.1, implies a PAC guarantee defined w.r.t. (pre)metric of Equation (8).

[12] As discussed in Remark 5.1, in the forward model case, the dissimilarity is in expectation w.r.t. the worst-case policy.

programming [e.g., 36] or the access to a black-box RL algorithm [e.g., 1]. In the class of *maximum entropy* approaches [42], the `Maximum Likelihood IRL` [41] converges to a stationary solution with $\widetilde{O}(\epsilon^{-2})$ trajectories for *non-linear* reward parametrization (with bounded gradient and Lipschitz smooth), when the underlying Markov chain is ergodic. Furthermore, the authors prove that, when the reward is linear in some features, the recovered solution corresponds to `Maximum Entropy IRL` [42]. Concerning the *gradient-based* approaches, [27] and [29] prove finite-sample convergence guarantee to the expert's weight under linear parametrization as a function of the accuracy of the gradient estimation. Surprisingly, a theoretical analysis of the IRL progenitor algorithm of [25] has been proposed only recently in [15]. A β-strict separability setting is enforced in which the rewards are assumed to lead to a suboptimality gap of at least $\beta > 0$ when playing any non-optimal action. For finite MDPs, known expert's policy, under the demanding assumption that each state is reachable in one step with a minimum probability $\alpha > 0$, and focusing on state-only reward, the authors prove that the algorithm outputs a β-strict separable feasible reward in at most $\widetilde{O}\left(\frac{1+\gamma^2\Xi^2}{\alpha\beta^2(1-\gamma)^4}\log\left(\frac{1}{\delta}\right)\right)$ trajectories, where $\Xi \leqslant S$ is the number of possible successor states. Recently, an approach with theoretical guarantees has been proposed for continuous states [7].

Reward-Free Exploration Reward-free exploration [RFE, 12, 14, 20] is a setting for pure exploration in MDPs composed of two phases: exploration and planning. In the exploration phase, the agent learns an estimated transition model $\widehat{p}$ without any reward feedback. In the planning phase, the agent is faced with a reward function r and has to output an estimated optimal policy $\widehat{\pi}^*$, using $\widehat{p}$ since no further interaction with the environment is admitted. In this sense, RFE shares this two-phase procedure with our IRL problem, but, instead of the *planning* phase, we face the *computation* of the feasible reward set.[13] In RFE exploration, the sample complexity is computed against the performance of the learned policy $\widehat{\pi}^*$ under the reward r, i.e., $V^*(\cdot;r) - V^{\widehat{\pi}^*}(\cdot;r)$, whose lower bound of the sample complexity has order $\Omega\left(\frac{H^2SA}{\epsilon^2}\left(H\log\left(\frac{1}{\delta}\right)+S\right)\right)$ [12, 14]. The best known algorithm, `RF-Express`, proposed in [20] archives an almost-matching sample complexity of order $\Omega\left(\frac{H^3SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)\right)$. The relevant connection with what we present in this paper is the fact that the derivation of the lower bounds shares similarity especially in the construction of the instances. Nevertheless, in the time-inhomogeneous case, we achieve a higher lower bound of order $\Omega\left(\frac{H^3SA}{\epsilon^2}\left(\log\left(\frac{1}{\delta}\right)+S\right)\right)$. The connection between IRL and RFE should be investigated in future works, as also mentioned in [19].

B Proofs

In this appendix, we report the proofs we omitted in the main paper.

B.1 Proofs of Section 3

Lemma B.1. *Let r be feasible for the IRL problem $(\mathcal{M}, \pi^E)$ bounded in $[-1,1]$ (i.e., $r \in \mathcal{R}$) and defined according to Lemma 3.1 as $r_h(s,a) = -A_h(s,a)\mathbb{1}_{\{\pi^E_h(a|s)=0\}} + V_h(s) - p_h V_{h+1}(s,a)$. Let $(\widehat{\mathcal{M}}, \widehat{\pi}^E)$ be an IRL problem and define for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]$:*

$$\epsilon_h(s,a) := -A_h(s,a)\left(\mathbb{1}_{\{\pi^E_h(a|s)=0\}} - \mathbb{1}_{\{\widehat{\pi}^E_h(a|s)=0\}}\right) \\ + \left((p_h - \widehat{p}_h)V_{h+1}\right)(s,a).$$

Then, the reward function $\widehat{r}$ defined according to Lemma 3.1 as $\widehat{r}_h(s,a) = -\widehat{A}_h(s,a)\mathbb{1}_{\{\widehat{\pi}^E_h(a|s)=0\}} + \widehat{V}_h(s) - \widehat{p}_h\widehat{V}_{h+1}(s,a)$ for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]$ with:

$$\widehat{A}_h(s,a) = \frac{A_h(s,a)}{1+\epsilon}, \quad \widehat{V}_h(s) = \frac{V_h(s)}{1+\epsilon}, \quad \widehat{V}_{H+1}(s) = 0.$$

where $\epsilon := \max_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]} |\epsilon_h(s,a)|$, is feasible for the IRL problem $(\widehat{\mathcal{M}}, \widehat{\pi}^E)$ and bounded in $[-1,1]$ (i.e., $\widehat{r} \in \widehat{\mathcal{R}}$).

[13] As shown in previous works, the computation of the feasible reward set can be formulated with a *linear feasibility problem* [25].

Proof. Given the reward function $r_h(s,a) = -A_h(s,a)\mathbb{1}_{\{\pi_h^E(a|s)=0\}} + V_h(s) - p_h V_{h+1}(s,a)$, we define the reward function:

$$\widetilde{r}_h(s,a) = -A_h(s,a)\mathbb{1}_{\{\widehat{\pi}_h^E(a|s)=0\}} + V_h(s) - \widehat{p}_h V_{h+1}(s,a),$$

that, thanks to Lemma 3.1, makes policy $\widehat{\pi}^E$ optimal. However, it is not guaranteed that $\widetilde{r} \in \widehat{\mathcal{R}}$ since it can take values larger than 1. Thus, we define the reward:

$$\widehat{r}_h(s,a) = \frac{\widetilde{r}_h(s,a)}{1+\epsilon} = -\frac{A_h(s,a)}{1+\epsilon}\mathbb{1}_{\{\widehat{\pi}_h^E(a|s)=0\}} + \frac{V_h}{1+\epsilon}(s) - \widehat{p}_h \frac{V_{h+1}}{1+\epsilon}(s,a),$$

which simply scales $\widetilde{r}_h$ and preserves the optimality of $\widehat{\pi}^E$. We now prove that $\widehat{r}_h(s,a)$ is bounded in $[-1,1]$. To do so, we prove that $\widetilde{r}_h(s,a)$ is bounded in $[-(1+\epsilon),(1+\epsilon)]$:

$$\begin{aligned}
|\widetilde{r}_h(s,a)| &\leqslant |r_h(s,a)| + |\widetilde{r}_h(s,a) - r_h(s,a)| \\
&= 1 + \left| -A_h(s,a)\mathbb{1}_{\{\widehat{\pi}_h^E(a|s)=0\}} + \widehat{p}_h V_{h+1}(s) - \left(-A_h(s,a)\mathbb{1}_{\{\pi_h^E(a|s)=0\}} + p_h V_{h+1}(s) \right) \right| \\
&= 1 + |\epsilon_h(s,a)| \leqslant 1 + \epsilon.
\end{aligned}$$

□

Theorem 3.2 (Lipschitz Continuity)**.** *Let $\mathcal{R}$ and $\widehat{\mathcal{R}}$ be the feasible reward sets of the IRL problems $(\mathcal{M},\pi^E)$ and $(\widehat{\mathcal{M}},\widehat{\pi}^E)$, as in Equation* (2)*. Then, it holds that:*[14]

$$\mathcal{H}_{d^G}(\mathcal{R},\widehat{\mathcal{R}}) \leqslant \frac{2\rho^G((\mathcal{M},\pi^E),(\widehat{\mathcal{M}},\widehat{\pi}^E))}{1+\rho^G((\mathcal{M},\pi^E),(\widehat{\mathcal{M}},\widehat{\pi}^E))}, \tag{5}$$

where $\rho^G(\cdot,\cdot)$ is a (pre)metric between IRL problems, defined as:

$$\rho^G((\mathcal{M},\pi^E),(\widehat{\mathcal{M}},\widehat{\pi}^E)) := \max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} (H-h+1)\left(\left| \mathbb{1}_{\{\pi_h^E(a|s)=0\}} - \mathbb{1}_{\{\widehat{\pi}_h^E(a|s)=0\}} \right| + \|p_h(\cdot|s,a) - \widehat{p}_h(\cdot|s,a)\|_1 \right).$$

Proof. Let $\widehat{r}$ as defined in the proof of Lemma B.1. Then, we have:

$$\begin{aligned}
|r_h(s,a) - \widehat{r}_h(s,a)| &= \left| r_h(s,a) - \frac{\widetilde{r}_h(s,a)}{1+\epsilon} \right| \\
&\leqslant \frac{1}{1+\epsilon}\left(|r_h(s,a) - \widetilde{r}_h(s,a)| + \epsilon |r_h(s,a)| \right) \\
&\leqslant \frac{2\epsilon}{1+\epsilon}.
\end{aligned}$$

By recalling that $\frac{2\epsilon}{1+\epsilon}$ is a non-decreasing function of ϵ, we bound it by replacing ϵ with an upper bound:

$$\begin{aligned}
\epsilon &= \max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} |\epsilon_h(s,a)| \\
&\leqslant \max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} (H-h+1)\left[\left| \mathbb{1}_{\{\pi_h^E(a|s)=0\}} - \mathbb{1}_{\{\widehat{\pi}_h^E(a|s)=0\}} \right| + \|p_h(\cdot|s,a) - \widehat{p}_h(\cdot|s,a)\|_1 \right] \\
&=: \rho^{\mathrm{G}}((\mathcal{M},\pi^E),(\widehat{\mathcal{M}},\widehat{\pi}^E)),
\end{aligned}$$

where we used Hölder's inequality recalling that $|V_{h+1}(s)| \leqslant H-h$ and $|A_h(s,a)| \leqslant H-h+1$. Clearly, $\rho^{\mathrm{G}}((\mathcal{M},\pi^E),(\widehat{\mathcal{M}},\widehat{\pi}^E))$ is a (pre)metric. □

Fact B.1. *There exist two MDP\R $\mathcal{M}$ and $\widehat{\mathcal{M}}$ with transition models p and $\widehat{p}$ respectively, an expert's policy π^E and a reward function $r_h(s,a) = -A_h(s,a)\mathbb{1}_{\{\pi^E(a|s)=0\}} + V_h(s) - p_h V_{h+1}(s)$ feasible for the IRL problem $(\mathcal{M},\pi^E)$ bounded in $[-1,1]$ (i.e., $r \in \mathcal{R}$) such that the reward function $\widehat{r}_h(s,a) = -A_h(s,a)\mathbb{1}_{\{\pi^E(a|s)=0\}} + V_h(s) - \widehat{p}_h V_{h+1}(s,a)$ is feasible for the IRL problem $(\widehat{\mathcal{M}},\pi^E)$ not bounded in $[-1,1]$.*

[14] This implies the standard Lipschitz continuity, by simply bounding $\frac{2\rho^{\mathrm{G}}((\mathcal{M},\pi^E),(\widehat{\mathcal{M}},\widehat{\pi}^E))}{1+\rho^{\mathrm{G}}((\mathcal{M},\pi^E),(\widehat{\mathcal{M}},\widehat{\pi}^E))} \leqslant 2\rho^{\mathrm{G}}((\mathcal{M},\pi^E),(\widehat{\mathcal{M}},\widehat{\pi}^E))$.

Proof. We consider the MDP\R in Figure 2 with optimal policy and reward function defined for every $h \in [\![H]\!]$ and $H=10$ as:

$$\pi_h^E(s_1)=a_1,\, \pi_h^E(s_2)=a_2,$$
$$r_h(s_1,a_1)=r_h(s_2,a_1)=0,\, r_h(s_1,a_2)=-1,\, r_h(s_2,a_2)=1.$$

Simple calculations lead to the V-function and advantage function values:

$$V_h^{\pi^E}(s_1)=0,\, V_h^{\pi^E}(s_2)=H-h+1,$$
$$A_h^{\pi^E}(s_1,a_1)=0,\, A_h^{\pi^E}(s_1,a_2)=-1+(H-h)/10,\, A_h^{\pi^E}(s_2,a_1)=-1-(H-h)/10,\, A_h^{\pi^E}(s_2,a_2)=0.$$

We consider as alternative transition model $\widehat{p}=1-p$. After tedious calculations we obtain the alternative reward function:

$$\widehat{r}_h(s_1,a_1)=-(H-h),\, \widehat{r}_h(s_1,a_2)=1-(H-h),\, \widehat{r}_h(s_2,a_1)=H-h+2,\, \widehat{r}_h(s_2,a_2)=H-h+1.$$

It is simple to observe that for some (s,a,h) we have $|\widehat{r}_h(s,a)|>1$.

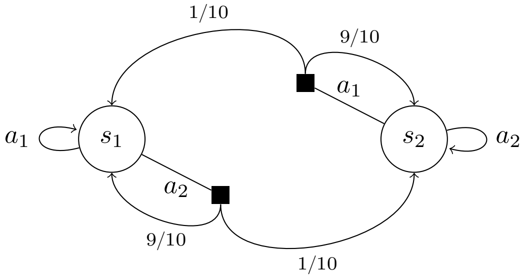


Figure 2: The MDP\R employed in Fact B.1.

□

B.2 Proofs of Section 4

Theorem 4.1 (Relationships between d-IRL problems)**.** *Let us introduce the graphical convention for $c>0$:*

$$\begin{tikzcd} \boxed{x\text{-IRL}} \arrow[r, "c"] & \boxed{y\text{-IRL}} \end{tikzcd}$$

meaning that any (ϵ,δ)-PAC x-IRL algorithm is $(c\epsilon,\delta)$-PAC y-IRL. Then, the following statements hold:

$$\begin{tikzcd} \boxed{d^{\mathrm{G}}\text{-IRL}} \arrow[r, "H"] \arrow[rr, bend left, "2H"] & \boxed{d^{\mathrm{G}}_{Q*}\text{-IRL}} \arrow[r, "2H"] & \boxed{d^{\mathrm{G}}_{V*}\text{-IRL}} \end{tikzcd}.$$

Proof. Let $\mathfrak{A}$ be an (ϵ,δ)-PAC d^{G}-IRL algorithm. This means that with probability at least $1-\delta$, we have that for any IRL problem $\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R},\widehat{\mathcal{R}}^\tau)\leqslant\epsilon$. We introduce the following visitation distributions, defined for every $s,s'\in\mathcal{S}$, $h,l\in[\![H]\!]$ with $l\geqslant h$, and $a,a'\in\mathcal{A}$:

$$\eta^{\pi}_{s,a,h,l}(s',a')=\mathop{\mathbb{P}}_{\mathcal{M},\pi}\left(s_l=s',a_l=a'|s_h=s,a_h=a\right), \qquad \eta^{\pi}_{s,h,l}(s',a')=\sum_{a\in\mathcal{A}}\pi_h(a|s)\eta^{\pi}_{s,a,h,l}(s',a').$$

d^{G}-IRL → d^{G}_{Q*}-IRL Let us consider the optimal Q-function difference and let π^* an optimal policy under the reward function r, we have:

$$Q^*_h(s,a;r)-Q^*_h(s,a;\widehat{r})\leqslant Q^{\pi^*}_h(s,a;r)-Q^{\pi^*}_h(s,a;\widehat{r})$$

$$
\begin{aligned}
&= \sum_{l=h}^{H} \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} \eta^{\pi^*}_{s,a,h,l}(s',a')(r_l(s',a') - \widehat{r}_l(s',a')) \\
&\leqslant \max_{(s,a,h')\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} |r_{h'}(s,a) - \widehat{r}_{h'}(s,a)| \underbrace{\sum_{l=h}^{H} \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} \eta^{\pi^*}_{s,a,h,l}(s',a')}_{=1} \\
&= (H-h+1) \max_{(s,a,h')\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} |r_{h'}(s,a) - \widehat{r}_{h'}(s,a)| \\
&\leqslant H \max_{(s,a,h')\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} |r_{h'}(s,a) - \widehat{r}_{h'}(s,a)|.
\end{aligned}
$$

As a consequence, we have:

$$
\mathcal{H}_{d^{\mathrm{G}}_{Q^*}}(\mathcal{R}, \widehat{\mathcal{R}}^\tau) \leqslant H \mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}, \widehat{\mathcal{R}}^\tau).
$$

$d^{\mathbf{G}}$-IRL → $d^{\mathbf{G}}_{V^*}$-IRL Let us consider the value functions and let π^* (resp. $\widehat{\pi}^*$) be an optimal policy under reward function r (resp. $\widehat{r}$), we have:

$$
\begin{aligned}
V^*_h(s;r) - V^{\widehat{\pi}^*}_h(s;r) &= V^{\pi^*}_h(s;r) - V^{\widehat{\pi}^*}_h(s;r) \pm V^{\widehat{\pi}^*}_h(s;\widehat{r}) \\
&\leqslant V^{\pi^*}_h(s;r) - V^{\pi^*}_h(s;\widehat{r}) + V^{\widehat{\pi}^*}_h(s;\widehat{r}) - V^{\widehat{\pi}^*}_h(s;r) \\
&= \sum_{l=h}^{H} \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} \eta^{\pi^*}_{s,h,l}(s',a')(r_l(s',a') - \widehat{r}_l(s',a')) \\
&\quad + \sum_{l=h}^{H} \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} \eta^{\widehat{\pi}^*}_{s,h,l}(s',a')(\widehat{r}_l(s',a') - r_l(s',a')) \\
&\leqslant \max_{(s,a,h')\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} |r_{h'}(s,a) - \widehat{r}_{h'}(s,a)| \\
&\quad \times \left(\sum_{l=h}^{H} \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} \eta^{\pi^*}_{s,h,l}(s',a') + \sum_{l=h}^{H} \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} \eta^{\widehat{\pi}^*}_{s,h,l}(s',a') \right) \\
&= 2(H-h+1) \max_{(s,a,h')\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} |r_{h'}(s,a) - \widehat{r}_{h'}(s,a)| \\
&\leqslant 2H \max_{(s,a,h')\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} |r_{h'}(s,a) - \widehat{r}_{h'}(s,a)|.
\end{aligned}
$$

Thus, it follows that:

$$
\mathcal{H}_{d^{\mathrm{G}}_{V^*}}(\mathcal{R}, \widehat{\mathcal{R}}^\tau) \leqslant 2H \mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}, \widehat{\mathcal{R}}^\tau).
$$

$d^{\mathbf{G}}_{Q^*}$-IRL → $d^{\mathbf{G}}_{V^*}$-IRL To prove this result, we need to introduce further tools. Specifically, we introduce the Bellman optimal operator and the Bellman expectation operator, defined for a reward function r, policy π, $(s,h) \in \mathcal{S} \times [\![H]\!]$ and function $f_h : \mathcal{S} \to \mathbb{R}$ defined for $h \in [\![H]\!]$ with $f_{H+1} = 0$:

$$
T^*_{r,h} f_h(s) = \max_{a\in\mathcal{A}} \{ r_h(s,a) + p_h f_{h+1}(s,a) \}, \qquad T^{\pi}_{r,h} f_h(s) = \pi_h \left(r_h(s,a) + p_h f_{h+1}(s,a) \right).
$$

We recall the fixed-point properties: $T^{\pi}_{r,h} V^{\pi}_h = V^{\pi}_h$ and $T^*_{r,h} V^*_h = V^*_h$. Let π^* (resp. $\widehat{\pi}^*$) be an optimal policy under reward r (resp. $\widehat{r}$). Let us consider the following derivation:

$$
\begin{aligned}
V^*_h(s;r) - V^{\widehat{\pi}^*}_h(s;r) &= T^*_{r,h} V^*_h(s;r) - T^{\widehat{\pi}^*}_{r,h} V^{\widehat{\pi}^*}_h(s;r) \pm T^{\pi^*}_{r,h} V^*_h(s;\widehat{r}) \pm T^{\pi^*}_{\widehat{r},h} V^*_h(s;\widehat{r}) \pm T^*_{\widehat{r},h} V^*_h(s;\widehat{r}) \pm T^{\widehat{\pi}^*}_{r,h} V^{\widehat{\pi}^*}_h(s;\widehat{r}) \\
&= T^{\pi^*}_{r,h} V^*_h(s;r) - T^{\pi^*}_{r,h} V^*_h(s;\widehat{r}) + T^{\pi^*}_{r,h} V^*_h(s,\widehat{r}) - T^{\pi^*}_{\widehat{r},h} V^*_h(s;\widehat{r}) + \underbrace{T^{\pi^*}_{\widehat{r},h} V^*_h(s;\widehat{r}) - T^*_{\widehat{r},h} V^*_h(s;\widehat{r})}_{\leqslant 0} \\
&\quad + T^{\widehat{\pi}^*}_{\widehat{r},h} V^*_h(s;\widehat{r}) - T^{\widehat{\pi}^*}_{r,h} V^*_h(s;\widehat{r}) + T^{\widehat{\pi}^*}_{r,h} V^*_h(s;\widehat{r}) - T^{\widehat{\pi}^*}_{r,h} V^{\widehat{\pi}^*}_h(s;r) \\
&\leqslant \pi^*_h p_h (V^*_{h+1}(\cdot;r) - V^*_{h+1}(\cdot;\widehat{r}))(s) + \pi^*_h (r_h - \widehat{r}_h)(s)
\end{aligned}
$$

$$
\begin{aligned}
&+\widehat{\pi}^*_h(\widehat{r}_h - r_h)(s) + \widehat{\pi}^*_h p_h(V^*_{h+1}(\cdot;\widehat{r}) - V^{\widehat{\pi}^*}_{h+1}(\cdot;r))(s)\\
&= (\pi^*_h - \widehat{\pi}^*_h)(Q^*_h(\cdot;r) - Q^*_h(\cdot;\widehat{r}))(s) + \widehat{\pi}^*_h p_h(V^*_{h+1}(\cdot;r) - V^{\widehat{\pi}^*}_{h+1}(\cdot;r))(s).
\end{aligned}
$$

Let us apply the L_∞-norm over the state space and the triangular inequality, we have:

$$
\begin{aligned}
\left\| V^*_h(\cdot;r) - V^{\widehat{\pi}^*}_h(\cdot;r) \right\|_\infty &\leqslant \| (\pi^*_h - \widehat{\pi}^*_h)(Q^*_h(\cdot;r) - Q^*_h(\cdot;\widehat{r}))(\cdot) \|_\infty + \left\| \widehat{\pi}^*_h p_h (V^*_{h+1}(\cdot;r) - V^{\widehat{\pi}^*}_{h+1}(\cdot;r))(\cdot) \right\|_\infty \\
&\leqslant 2 \| Q^*_h(\cdot;r) - Q^*_h(\cdot;\widehat{r}))(\cdot) \|_\infty + \left\| V^*_{h+1}(\cdot;r) - V^{\widehat{\pi}^*}_{h+1}(\cdot;r) \right\|_\infty .
\end{aligned}
$$

By unfolding the recursion over h, we obtain:

$$
\left\| V^*_h(\cdot;r) - V^{\widehat{\pi}^*}_h(\cdot;r) \right\|_\infty \leqslant 2 \sum_{l=h}^{H} \| Q^*_l(\cdot;r) - Q^*_l(\cdot;\widehat{r}))(\cdot) \|_\infty .
$$

Thus, we have:

$$
\max_{(s,h)\in\mathcal{S}\times[\![H]\!]} \left| V^*_h(s;r) - V^{\widehat{\pi}^*}_h(s;r) \right| \leqslant 2H \max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} \left| Q^*_h(s,a;r) - Q^*_h(s,a;\widehat{r}) \right| .
$$

Since the derivation is carried out for arbitrary $\widehat{\pi}^*$, it follows that:

$$
\mathcal{H}_{d^{\mathrm{G}}_{V^*}}(\mathcal{R}, \widehat{\mathcal{R}}^\tau) \leqslant 2H \mathcal{H}_{d^{\mathrm{G}}_{Q^*}}(\mathcal{R}, \widehat{\mathcal{R}}^\tau).
$$

□

B.3 Proofs of Section 5

Theorem 5.1 (Lower Bound for d^{G}-IRL)**.** *Let $\mathfrak{A} = (\mu, \tau)$ be an (ϵ, δ)-PAC algorithm for d^{G}-IRL. Then, there exists an IRL problem $(\mathcal{M}, \pi^E)$ such that, if $\epsilon \leqslant 1/64$, $\delta \leqslant 1/32$, $S \geqslant 9$, $A \geqslant 2$, and $H \geqslant 12$, the expected sample complexity is lower bounded by:*

- *if the transition model p is time-inhomogeneous:*

$$
\mathop{\mathbb{E}}_{(\mathcal{M},\pi^E),\mathfrak{A}}[\tau] \geqslant \Omega\left(\frac{H^3 SA}{\epsilon^2} \left(\log\left(\frac{1}{\delta}\right) + S \right) \right);
$$

- *if the transition model p is time-homogeneous:*

$$
\mathop{\mathbb{E}}_{(\mathcal{M},\pi^E),\mathfrak{A}}[\tau] \geqslant \Omega\left(\frac{H^2 SA}{\epsilon^2} \left(\log\left(\frac{1}{\delta}\right) + S \right) \right),
$$

where $\mathbb{E}_{(\mathcal{M},\pi^E),\mathfrak{A}}$ denotes the expectation w.r.t. the probability measure $\mathbb{P}_{(\mathcal{M},\pi^E),\mathfrak{A}}$.

Proof. We put together the results of Theorem B.2 and Theorem B.3, by recalling that $\max\{a,b\} \geqslant \frac{a+b}{2}$, or, equivalently, assuming to observe instances like the ones of Theorem B.2 w.p. $1/2$ as well as those of Theorem B.3. □

Theorem B.2. *Let $\mathfrak{A} = (\mu, \tau)$ be an (ϵ, δ)-PAC algorithm for d^{G}-IRL. Then, there exists an IRL problem $(\mathcal{M}, \pi^E)$ such that, if $\epsilon \leqslant 1/2$, $\delta < 1/16$, $S \geqslant 9$, $A \geqslant 2$, and $H \geqslant 12$, the expected sample complexity is lower bounded by:*

- *if the transition model p is time-inhomogeneous:*

$$
\mathop{\mathbb{E}}_{(\mathcal{M},\pi^E),\mathfrak{A}}[\tau] \geqslant \Omega\left(\frac{H^3 SA}{\epsilon^2} \log\left(\frac{1}{\delta}\right) \right);
$$

- *if the transition model p is time-homogeneous:*

$$
\mathop{\mathbb{E}}_{(\mathcal{M},\pi^E),\mathfrak{A}}[\tau] \geqslant \Omega\left(\frac{H^2 SA}{\epsilon^2} \log\left(\frac{1}{\delta}\right) \right).
$$

Proof. **Step 1: Instances Construction** The construction of the hard MDP\R instances follows similar steps as the ones presented in the constructions of lower bounds for policy learning [8] and the hard instances are reported in Figure 3 in a semi-formal way. The state space is given by $\mathcal{S}=\{s_{\text{start}}, s_{\text{root}}, s_-, s_+, s_1, \dots, s_{\overline{S}}\}$ and the action space is given by $\mathcal{A}=\{a_0, a_1, \dots, a_{\overline{A}}\}$. The transition model is described below and the horizon is $H \geqslant 3$. We introduce the constant $\overline{H} \in [\![H]\!]$, whose value will be chosen later. Let us observe, for now, that if $\overline{H}=1$, the transition model is time-homogeneous.

The agent begins in state s_{start}, where every action has the same effect. Specifically, if the stage $h < \overline{H}$, then there is probability $1/2$ to remain in s_{start} and a probability $1/2$ to transition to s_{root}. Instead, if $h \geqslant \overline{H}$, the state transitions to s_{root} deterministically. From state s_{root}, every action has the same effect and the state transitions with equal probability $1/\overline{S}$ to a state s_i with $i \in [\![\overline{S}]\!]$. In all states s_i, apart from a specific one, i.e., state s_*, all actions have the same effect, i.e., transitioning to states s_- and s_+ with equal probability $1/2$. State s_* behaves as the other ones if the stage $h \neq h_*$, where $h_* \in [\![H]\!]$ is a predefined stage. If, instead, $h = h_*$, all actions $a_j \neq a_*$ behave like in the other states, while for action a_*, we have a $1/2+\epsilon'$ probability of reaching s_+ (and consequently probability $1/2-\epsilon'$ of reaching s_-), with $\epsilon' \in [0, 1/4]$. Notice that, having fixed $\overline{H}$, the possible values of h_* are $\{3, \dots, 2+\overline{H}\}$. States s_+ and s_- are absorbing states. The expert's policy always plays action a_0.

Let us consider the base instance $\mathcal{M}_0$ in which there is no state behaving like s_*. Additionally, by varying the triple $\ell := (s_*, a_*, h_*) \in \{s_1, \dots, s_{\overline{S}}\} \times \{a_1, \dots, a_{\overline{A}}\} \times [\![3, \overline{H}+2]\!] =: \mathcal{I}$, we can construct the class of instances denoted by $\mathbb{M} = \{\mathcal{M}_\ell : \ell \in \{0\} \cup \mathcal{I}\}$.

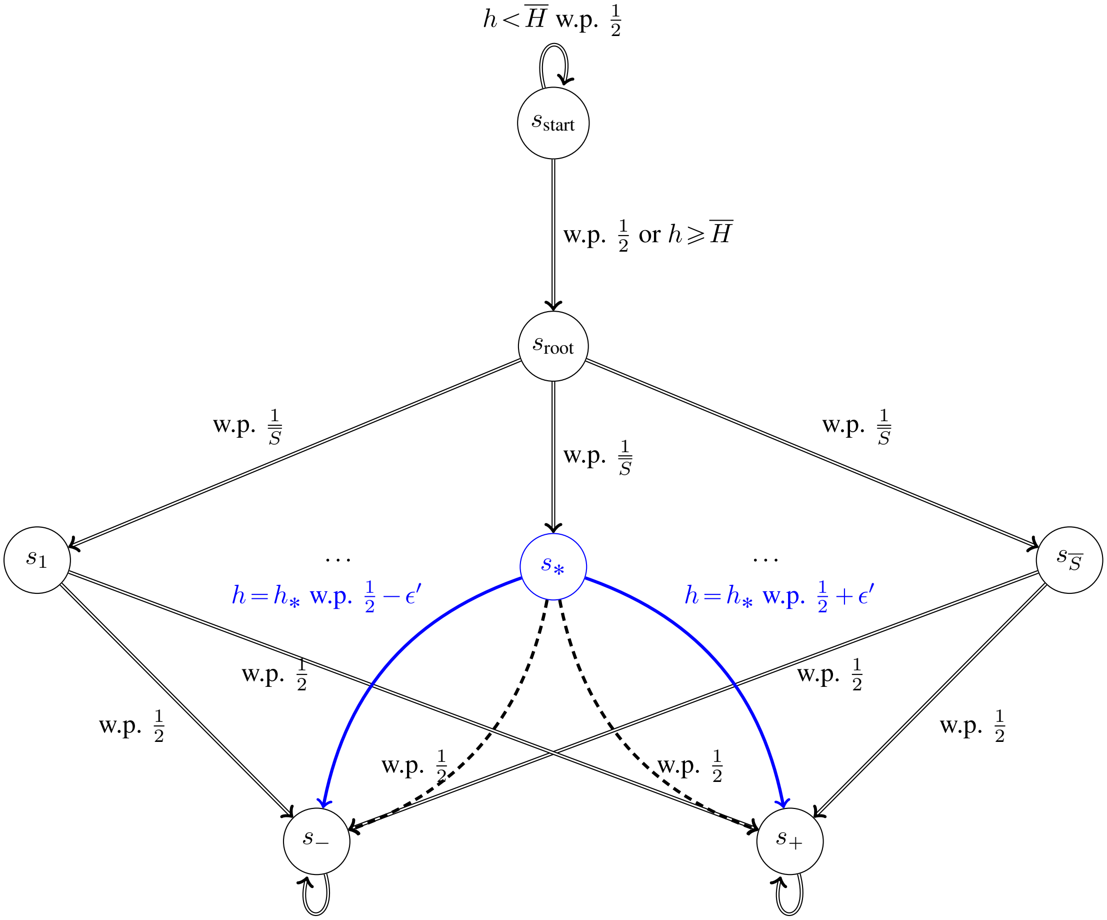

$h < \overline{H}$ w.p. $\frac{1}{2}$
s_{start}
w.p. $\frac{1}{2}$ or $h \geqslant \overline{H}$
s_{root}
w.p. $\frac{1}{\overline{S}}$
w.p. $\frac{1}{\overline{S}}$
w.p. $\frac{1}{\overline{S}}$
s_1
$\dots$
s_*
$\dots$
$s_{\overline{S}}$
$h = h_*$ w.p. $\frac{1}{2} - \epsilon'$
$h = h_*$ w.p. $\frac{1}{2} + \epsilon'$
w.p. $\frac{1}{2}$
w.p. $\frac{1}{2}$
w.p. $\frac{1}{2}$
w.p. $\frac{1}{2}$
w.p. $\frac{1}{2}$
w.p. $\frac{1}{2}$
s_-
s_+

Figure 3: Semi-formal representation of the the hard instances MDP\R used in the proof of Theorem B.2.

Step 2: Feasible Set Computation Let us consider an instance $\mathcal{M}_\ell \in \mathbb{M}$, we now seek to provide a lower bound to the Hausdorff distance $\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}_{\mathcal{M}_0}, \mathcal{R}_{\mathcal{M}_\ell})$. To this end, we focus on the triple $\ell = (s_*, a_*, h_*)$ and we enforce the convenience of action a_0 over action a_*. For the base MDP\R $\mathcal{M}_0$, let $r^0 \in \mathcal{R}_{\mathcal{M}_0}$, we have:

$$\begin{aligned}
&r^0_{h_*}(s_*, a_0) + \frac{1}{2}\sum_{l=h_*+1}^{H}\left(r^0_l(s_-) + r^0_l(s_+)\right) \geqslant r^0_{h_*}(s_*, a_*) + \frac{1}{2}\sum_{l=h+1}^{H}\left(r^0_l(s_-) + r^0_l(s_+)\right)\\
&\quad \implies r^0_{h_*}(s_*, a_0) \geqslant r^0_{h_*}(s_*, a_*),
\end{aligned}$$

For the alternative MDP\R $\mathcal{M}_\ell$, let $r^\ell \in \mathcal{R}_{\mathcal{M}_\ell}$, we have:

$$\begin{aligned}
&r^\ell_{h_*}(s_*, a_0) + \frac{1}{2}\sum_{l=h_*+1}^{H}\left(r^\ell_l(s_-) + r^\ell_l(s_+)\right) \geqslant r^\ell_{h_*}(s_*, a_*) + \sum_{l=h_*+1}^{H}\left(\left(\frac{1}{2} - \epsilon'\right) r^\ell_l(s_-) + \left(\frac{1}{2} + \epsilon'\right) r^\ell_l(s_+)\right)\\
&\quad \implies r^\ell_{h_*}(s_*, a_0) \geqslant r^\ell_{h_*}(s_*, a_*) - \epsilon' \sum_{l=h_*+1}^{H}\left(r^\ell_l(s_-) - r^\ell_l(s_+)\right).
\end{aligned}$$

In order to lower bound the Hausdorff distance $\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}_{\mathcal{M}_0}, \mathcal{R}_{\mathcal{M}_\ell})$, we proceed as follows:

$$\begin{aligned}
\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}_{\mathcal{M}_0}, \mathcal{R}_{\mathcal{M}_\ell}) &= \max\left\{\sup_{r^0 \in \mathcal{R}_{\mathcal{M}_0}} \inf_{r^\ell \in \mathcal{R}_{\mathcal{M}_\ell}} d^{\mathrm{G}}(r^0, r^\ell), \sup_{r^\ell \in \mathcal{R}_{\mathcal{M}_\ell}} \inf_{r^0 \in \mathcal{R}_{\mathcal{M}_0}} d^{\mathrm{G}}(r^\ell, r^0)\right\}\\
&\geqslant \sup_{r^\ell \in \mathcal{R}_{\mathcal{M}_\ell}} \inf_{r^0 \in \mathcal{R}_{\mathcal{M}_0}} d^{\mathrm{G}}(r^\ell, r^0)\\
&\geqslant \inf_{r^0 \in \mathcal{R}_{\mathcal{M}_0}} d^{\mathrm{G}}(r^\ell, r^0),
\end{aligned}$$

for a specific choice of the reward function r^ℓ for $\mathcal{M}_\ell$ defined as:

$$r^\ell_l(s_-) = -r^\ell_l(s_+) = 1,\, r^\ell_{h_*}(s_*, a_*) = 1,\, r^\ell_{h_*}(s_*, a_0) = 1 - 2\epsilon'(H - h_*),$$

where we enforce $\epsilon' \leqslant \min_{h_* \in [\![3, \overline{H}+2]\!]} 1/(H - h_*) = 1/(H-3) \leqslant 1/4$ (which is guaranteed for $H \geqslant 7$) to ensure $r^\ell_{h_*}(s_*, a_0) \geqslant -1$. Then, for notational convenience, for the MDP\R $\mathcal{M}_0$, we set $y := r^0_{h_*}(s_*, a_0)$ and $x := r^0_{h_*}(s_*, a_*)$:

$$\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}_{\mathcal{M}_0}, \mathcal{R}_{\mathcal{M}_\ell}) \geqslant \min_{\substack{x, y \in [-1,1]\\ y \geqslant x}} \max\left\{|x - 1|, \left|y - 1 + 2\epsilon'(H - h_*)\right|\right\} = \epsilon'(H - h_*).$$

We enforce the following constraint on this quantity:

$$\forall h_* \in [\![3, \overline{H}+2]\!] : (H - h_*)\epsilon' \geqslant 2\epsilon \implies \epsilon' \geqslant \max_{h_* \in [\![3, \overline{H}+2]\!]} \frac{2\epsilon}{(H - h_*)} = \frac{2\epsilon}{(H - \overline{H} - 2)}. \tag{9}$$

Notice that $\epsilon' \leqslant 1/4$ whenever $H \geqslant \overline{H} + 10$. This latter condition, together with $\epsilon' \leqslant 1/(H-3)$, implies $\epsilon \leqslant \frac{H - \overline{H} - 2}{2(H-3)}$ that is satisfied for $\epsilon \leqslant 1/2$.

Step 3: Lower bounding Probability Let us consider an (ϵ, δ)-correct algorithm $\mathfrak{A}$ that outputs the estimated feasible set $\widehat{\mathcal{R}}$. Thus, for every $\imath \in \mathcal{I}$, we can lower bound the error probability:

$$\begin{aligned}
\delta &\geqslant \sup_{\text{all } \mathcal{M} \text{ MDP\R and expert policies } \pi} \mathbb{P}_{(\mathcal{M}, \pi), \mathfrak{A}}\left(\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\mathcal{M}}, \widehat{\mathcal{R}}\right) \geqslant \epsilon\right)\\
&\geqslant \sup_{\mathcal{M} \in \mathbb{M}} \mathbb{P}_{(\mathcal{M}, \pi), \mathfrak{A}}\left(\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\mathcal{M}}, \widehat{\mathcal{R}}\right) \geqslant \epsilon\right)\\
&\geqslant \max_{\ell \in \{0, \imath\}} \mathbb{P}_{(\mathcal{M}_\ell, \pi), \mathfrak{A}}\left(\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\mathcal{M}_\ell}, \widehat{\mathcal{R}}\right) \geqslant \epsilon\right).
\end{aligned}$$

For every $\imath \in \mathcal{I}$, let us define the *identification function* (whose dependence on the estimated feasible reward set $\widehat{\mathcal{R}}$ is omitted to avoid a too heavy notation):

$$\Psi_\imath := \operatorname*{arg\,min}_{\ell \in \{0, \imath\}} \mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\mathcal{M}_\ell}, \widehat{\mathcal{R}}\right).$$

Let $\jmath \in \{0,\imath\}$. If $\Psi_\imath = \jmath$, then, $\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}_{\mathcal{M}_{\Psi_\imath}}, \mathcal{R}_{\mathcal{M}_\jmath}) = 0$. Otherwise, if $\Psi_\imath \neq \jmath$, we have:

$$\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\mathcal{M}_{\Psi_\imath}}, \mathcal{R}_{\mathcal{M}_\jmath}\right) \leqslant \mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\mathcal{M}_{\Psi_\imath}}, \widehat{\mathcal{R}}\right) + \mathcal{H}_{d^{\mathrm{G}}}\left(\widehat{\mathcal{R}}, \mathcal{R}_{\mathcal{M}_\jmath}\right) \leqslant 2\mathcal{H}_{d^{\mathrm{G}}}\left(\widehat{\mathcal{R}}, \mathcal{R}_{\mathcal{M}_\jmath}\right),$$

where the first inequality follows from triangular inequality and the second one from the definition of identification function $\Psi_\imath$. From Equation (9), we have that $\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\mathcal{M}_{\Psi_\imath}}, \mathcal{R}_{\mathcal{M}_\jmath}\right) \geqslant 2\epsilon$. Thus, it follows that $\mathcal{H}_{d^{\mathrm{G}}}\left(\widehat{\mathcal{R}}, \mathcal{R}_{\mathcal{M}_\jmath}\right) \geqslant \epsilon$. This implies the following inclusion of events for $\jmath \in \{0,\imath\}$:

$$\left\{\mathcal{H}_{d^{\mathrm{G}}}\left(\widehat{\mathcal{R}}, \mathcal{R}_{\mathcal{M}_\jmath}\right) \geqslant \epsilon\right\} \supseteq \{\Psi_\imath \neq \jmath\}.$$

Thus, we can proceed by lower bounding the probability:

$$\begin{aligned}\max_{\ell \in \{0,\imath\}} \mathop{\mathbb{P}}_{(\mathcal{M}_\ell,\pi),\mathfrak{A}} \left(\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\mathcal{M}_\ell}, \widehat{\mathcal{R}}\right) \geqslant \epsilon\right) &\geqslant \max_{\ell \in \{0,\imath\}} \mathop{\mathbb{P}}_{(\mathcal{M}_\ell,\pi),\mathfrak{A}} (\Psi_\imath \neq \ell) \\ &\geqslant \frac{1}{2}\left[\mathop{\mathbb{P}}_{(\mathcal{M}_0,\pi),\mathfrak{A}} (\Psi_\imath \neq 0) + \mathop{\mathbb{P}}_{(\mathcal{M}_\imath,\pi),\mathfrak{A}} (\Psi_\imath \neq \imath)\right] \\ &= \frac{1}{2}\left[\mathop{\mathbb{P}}_{(\mathcal{M}_0,\pi),\mathfrak{A}} (\Psi_\imath \neq 0) + \mathop{\mathbb{P}}_{(\mathcal{M}_\imath,\pi),\mathfrak{A}} (\Psi_\imath = 0)\right],\end{aligned}$$

where the second inequality follows from the observation that $\max\{a,b\} \geqslant \frac{1}{2}(a+b)$ and the equality from observing that $\Psi_\imath \in \{0,\imath\}$. The intuition behind this derivation is that we lower bound the probability of making a mistake $\geqslant \epsilon$ with the probability of failing in identifying the true underlying problem. We can now apply the Bretagnolle-Huber inequality [17, Theorem 14.2] (also reported in Theorem D.1 for completeness) with $\mathbb{P} = \mathbb{P}_{(\mathcal{M}_0,\pi),\mathfrak{A}}$, $\mathbb{Q} = \mathbb{P}_{(\mathcal{M}_0,\pi)\mathfrak{A}}$, and $\mathcal{A} = \{\Psi_\imath \neq 0\}$:

$$\mathop{\mathbb{P}}_{(\mathcal{M}_0,\pi),\mathfrak{A}} (\Psi_\imath \neq 0) + \mathop{\mathbb{P}}_{(\mathcal{M}_\imath,\pi),\mathfrak{A}} (\Psi_\imath = 0) \geqslant \frac{1}{2}\exp\left(-D_{\mathrm{KL}}\left(\mathop{\mathbb{P}}_{(\mathcal{M}_0,\pi),\mathfrak{A}}, \mathop{\mathbb{P}}_{(\mathcal{M}_\imath,\pi),\mathfrak{A}}\right)\right).$$

Step 4: KL-divergence Computation Let $\mathcal{M} \in \mathbb{M}$, we denote with $\mathbb{P}_{(\mathcal{M},\pi),\mathfrak{A}}$ the joint probability distribution of all events realized by the execution of the algorithm in the MDP\R (the presence of π is irrelevant as we assume it known):

$$\mathop{\mathbb{P}}_{(\mathcal{M},\pi),\mathfrak{A}} = \prod_{t=1}^{\tau} \rho_t(s_t, a_t, h_t | H_{t-1}) p_{h_t}(s'_t | s_t, a_t).$$

where ρ_t is the sampling distribution induced by the algorithm $\mathfrak{A}$ and $H_{t-1} = (s_1, a_1, h_1, s'_1, \ldots, s_{t-1}, a_{t-1}, h_{t-1}, s'_{t-1})$ is the history. Let $\imath \in \mathcal{I}$ and denote with p^0 and $p^\imath$ the transition models associated with $\mathcal{M}_0$ and $\mathcal{M}_\imath$. Let us now move to the KL-divergence:

$$\begin{aligned}D_{\mathrm{KL}}\left(\mathbb{P}_{(\mathcal{M}_0,\pi),\mathfrak{A}}, \mathbb{P}_{(\mathcal{M}_\imath,\pi),\mathfrak{A}}\right) &= \mathop{\mathbb{E}}_{(\mathcal{M}_0,\pi),\mathfrak{A}}\left[\sum_{t=1}^{\tau} \log \frac{p^0_{h_t}(s'_t|s_t,a_t)}{p^\imath_{h_t}(s'_t|s_t,a_t)}\right] \\ &= \mathop{\mathbb{E}}_{(\mathcal{M}_0,\pi),\mathfrak{A}}\left[\sum_{t=1}^{\tau} D_{\mathrm{KL}}\left(p^0_{h_t}(\cdot|s_t,a_t), p^\imath_{h_t}(\cdot|s_t,a_t)\right)\right] \\ &\leqslant \mathop{\mathbb{E}}_{(\mathcal{M}_0,\pi),\mathfrak{A}}\left[N^\tau_{h_*}(s_*,a_*)\right] D_{\mathrm{KL}}\left(p^0_{h_*}(\cdot|s_*,a_*), p^\imath_{h_*}(\cdot|s_*,a_*)\right) \\ &\leqslant 8(\epsilon')^2 \mathop{\mathbb{E}}_{(\mathcal{M}_0,\pi),\mathfrak{A}}\left[N^\tau_{h_*}(s_*,a_*)\right].\end{aligned}$$

having observed that the transition models differ in $\imath = (s_*, a_*, h_*)$ and defined $N^\tau_{h_*}(s_*,a_*) = \sum_{t=1}^{\tau} \mathbb{1}\{(s_t,a_t,h_t) = (s_*,a_*,h_*)\}$ and the last passage is obtained by Lemma D.4 with $D=2$ (and $\epsilon = 2\epsilon'$). Putting all together, we have:

$$\delta \geqslant \frac{1}{4}\exp\left(-8 \mathop{\mathbb{E}}_{(\mathcal{M}_0,\pi),\mathfrak{A}}\left[N^\tau_{h_*}(s_*,a_*)\right](\epsilon')^2\right) \implies \mathop{\mathbb{E}}_{(\mathcal{M}_0,\pi),\mathfrak{A}}\left[N^\tau_{h_*}(s_*,a_*)\right] \geqslant \frac{\log\frac{1}{4\delta}}{8(\epsilon')^2} = \frac{(H - \overline{H} - 2)^2 \log\frac{1}{4\delta}}{32\epsilon^2}.$$

Thus, since we have lower bounded the sample complexity considering the pair of MDPs $\{\mathcal{M}_0, \mathcal{M}_\imath\}$, we can proceed at summing over $(s_*, a_*, h_*) \in \mathcal{I}$ to obtain:

$$\mathop{\mathbb{E}}_{(\mathcal{M}_0,\pi),\mathfrak{A}}[\tau] \geqslant \sum_{(s_*,a_*,h_*) \in \mathcal{I}} \mathop{\mathbb{E}}_{(\mathcal{M}_0,\pi),\mathfrak{A}}\left[N^\tau_{h_*}(s_*,a_*)\right]$$

$$
\begin{aligned}
&= \sum_{(s_*,a_*,h_*)\in\mathcal{I}} \frac{(H-\overline{H}-2)^2\log\frac{1}{4\delta}}{32\epsilon^2}\\
&= \frac{\overline{SAH}(H-\overline{H}-2)^2}{32\epsilon^2}\log\frac{1}{4\delta}.
\end{aligned}
$$

The number of states is given by $S=|\mathcal{S}|=\overline{S}+4$, the number of actions is given by $A=|\mathcal{A}|=\overline{A}+1$. Let us first consider the time-homogeneous case, i.e., $\overline{H}=1$:

$$
\mathop{\mathbb{E}}_{(\mathcal{M}_0,\pi),\mathfrak{A}}[\tau]\geqslant\frac{(S-4)(A-1)(H-3)^2}{32\epsilon^2}\log\frac{1}{4\delta}.
$$

For $\delta<1/16$, $S\geqslant 9$, $A\geqslant 2$, $H\geqslant 10$, we obtain:

$$
\mathop{\mathbb{E}}_{(\mathcal{M}_0,\pi),\mathfrak{A}}[\tau]\geqslant\Omega\left(\frac{SAH^2}{\epsilon^2}\log\frac{1}{\delta}\right).
$$

For the time-inhomogeneous case, instead, we select $\overline{H}=H/2$, to get:

$$
\mathop{\mathbb{E}}_{(\mathcal{M}_0,\pi),\mathfrak{A}}[\tau]\geqslant\frac{(S-4)(A-1)(H/2)(H-H/2-2)^2}{\epsilon^2}\log\frac{1}{4\delta}.
$$

For $\delta<1/16$, $S\geqslant 9$, $A\geqslant 2$, $H\geqslant 12$, we obtain:

$$
\mathop{\mathbb{E}}_{(\mathcal{M}_0,\pi),\mathfrak{A}}[\tau]\geqslant\Omega\left(\frac{SAH^3}{\epsilon^2}\log\frac{1}{\delta}\right).
$$

□

Theorem B.3. *Let $\mathfrak{A}=(\mu,\tau)$ be an (ϵ,δ)-PAC algorithm for d^G-IRL. Then, there exists an IRL problem $(\mathcal{M},\pi^E)$ such that, if $\epsilon\leqslant 1/64$, $\delta\leqslant 1/2$, $S\geqslant 16$, $A\geqslant 2$, $H\geqslant 131$, the expected sample complexity is lower bounded by:*

- *if the transition model p is time-inhomogeneous:*

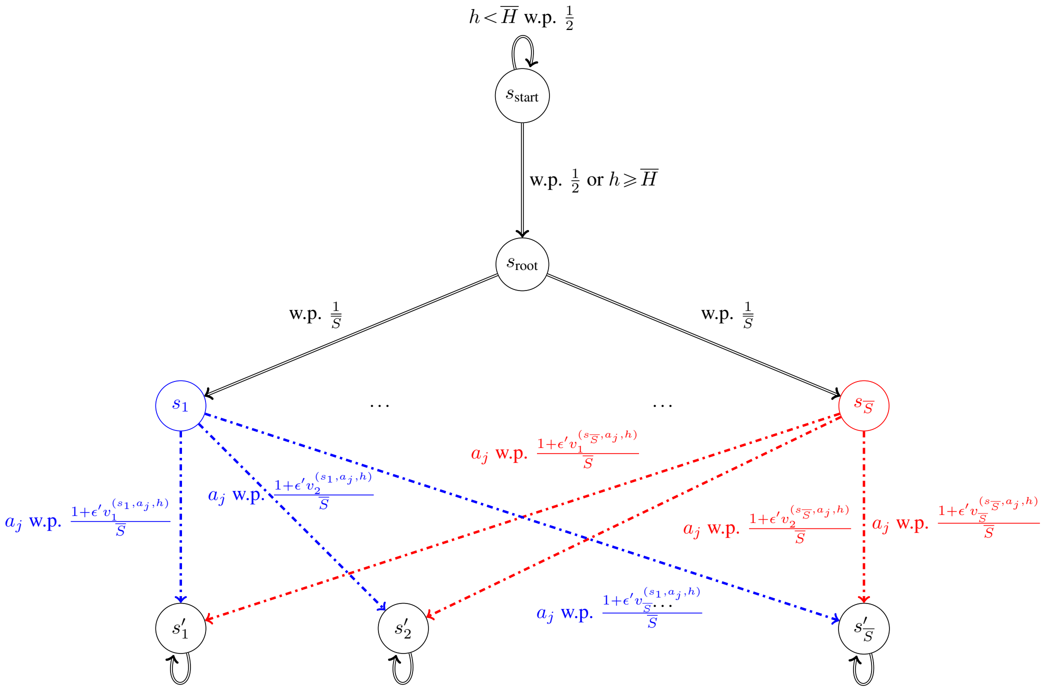
$$
\mathop{\mathbb{E}}_{(\mathcal{M},\pi^E),\mathfrak{A}}[\tau]\geqslant\frac{1}{5120}\frac{S^2AH^3}{\epsilon^2};
$$

- *if the transition model p is time-homogeneous:*

$$
\mathop{\mathbb{E}}_{(\mathcal{M},\pi^E),\mathfrak{A}}[\tau]\geqslant\frac{1}{2560}\frac{S^2AH^2}{\epsilon^2}.
$$

Proof. **Step 1: Instances Construction** The construction of the hard MDP\R instances for this second bound follows steps similar to those of reward free exploration [12] and the instances are reported in Figure 4 in a semi-formal way. The state space is given by $\mathcal{S}=\{s_{\text{start}},s_{\text{root}},s_1,\dots,s_{\overline{S}},s'_1,\dots,s'_{\overline{S}}\}$ and the action space is given by $\mathcal{A}=\{a_0,a_1,\dots,a_{\overline{A}}\}$. We assume $\overline{S}$ to be divisible by 16. The transition model is described below and the horizon is $H\geqslant 3$.

The agent begins in state s_{start}, where every action has the same effect. Specifically, if the stage $h<\overline{H}$ ($\overline{H}\in[\![H]\!]$, whose value will be chosen later), then there is probability $1/2$ to remain in s_{start} and a probability $1/2$ to transition to s_{root}. Instead, if $h\geqslant\overline{H}$, the state transitions to s_{root} deterministically. From state s_{root}, every action has the same effect and the state transitions with equal probability $1/\overline{S}$ to a state s_i with $i\in[\![\overline{S}]\!]$. In every state s_i and every stage h, action a_0 allows reaching states $s'_1,\dots,s'_{\overline{S}}$ with equal probability $1/\overline{S}$. Instead, by playing the other actions a_j with $j\geqslant 1$ at stage h, the probability distribution of the next state is given by $p_h(s'_k|s_i,a_j)=(1+\epsilon' v_k^{(s_i,a_j,h)})/\overline{S}$ where the vector $v^{(s_i,a_j,h)}=(v_1^{(s_i,a_j,h)},\dots,v_{\overline{S}}^{(s_i,a_j,h)})\in\mathcal{V}$, where $\mathcal{V}:=\{\{-1,1\}^{\overline{S}}:\sum_{j=1}^{\overline{S}}v_j=0\}$ and $\epsilon'\in[0,1/2]$. Notice that, having fixed $\overline{H}$, the possible values of h are $\{3,\dots,2+\overline{H}\}$. States $s'_1,\dots,s'_{\overline{S}}$ are absorbing states. The expert's policy always plays action a_0.



$h < \overline{H}$ w.p. $\frac{1}{2}$
s_{start}
w.p. $\frac{1}{2}$ or $h \geqslant \overline{H}$
s_{root}
w.p. $\frac{1}{\overline{S}}$
w.p. $\frac{1}{\overline{S}}$
s_1
$\dots$
$\dots$
$s_{\overline{S}}$
a_j w.p. $\frac{1+\epsilon' v_1^{(s_{\overline{S}},a_j,h)}}{\overline{S}}$
a_j w.p. $\frac{1+\epsilon' v_1^{(s_1,a_j,h)}}{\overline{S}}$
a_j w.p. $\frac{1+\epsilon' v_2^{(s_1,a_j,h)}}{\overline{S}}$
a_j w.p. $\frac{1+\epsilon' v_2^{(s_{\overline{S}},a_j,h)}}{\overline{S}}$
a_j w.p. $\frac{1+\epsilon' v_{\overline{S}}^{(s_{\overline{S}},a_j,h)}}{\overline{S}}$
a_j w.p. $\frac{1+\epsilon' v_{\overline{S}}^{(s_1,a_j,h)}}{\overline{S}}$
$\dots$
s'_1
s'_2
$s'_{\overline{S}}$

Figure 4: Semi-formal representation of the the hard instances MDP\R used in the proof of Theorem B.3.

Let us introduce the set $\mathcal{I} := \{s_1, \dots, s_{\overline{S}}\} \times \{a_1, \dots, a_{\overline{A}}\} \times [\![3, \overline{H}+2]\!]$. Let $\boldsymbol{v} = (v^\imath)_{\imath \in \mathcal{I}} \in \mathcal{V}^{\mathcal{I}}$ which is the set of vectors having as components the elements $v^\imath$ determining the probability distribution of the next state starting from the triple $\imath \in \mathcal{I}$. We denote with $\mathcal{M}_{\boldsymbol{v}}$ the MDP\R induced by $\boldsymbol{v}$. We can construct the class of instances denoted by $\mathbb{M} = \{\mathcal{M}_{\boldsymbol{v}} : \boldsymbol{v} \in \mathcal{V}^{\mathcal{I}}\}$. Moreover, we denote with $\mathcal{M}_{\boldsymbol{v} \overset{\imath}{\leftarrow} w}$ the instance in which we replace the $\imath$ component of $\boldsymbol{v}$, i.e., $v^\imath$, with $w \in \mathcal{V}$ and $\mathcal{M}_{\boldsymbol{v} \overset{\imath}{\leftarrow} 0}$ the instance in which we replace the $\imath$ component of $\boldsymbol{v}$, i.e., $v^\imath$, with the zero vector.

Step 2: Feasible Set Computation Thanks to Lemma D.6, we know that there exists a subset $\overline{\mathcal{V}} \subset \mathcal{V}$ of cardinality at least $|\overline{\mathcal{V}}| \geqslant 2^{\overline{S}/5}$ such that for every $v, w \in \overline{\mathcal{V}}$ with $v \neq w$ we have $\sum_{j=1}^{\overline{S}} |v_j - w_j| \geqslant \overline{S}/16$. Thus, we consider the set $\overline{\mathcal{V}}^{\mathcal{I}} \subset \mathcal{V}^{\mathcal{I}}$. The instances will be defined in terms of a vector $\boldsymbol{v} \in \overline{\mathcal{V}}^{\mathcal{I}}$ and we will use $v, w \in \overline{\mathcal{V}}$ with $v \neq w$ to build the alternative instances. Let $\imath \in \mathcal{I}$, the induced instances are denoted by $\mathcal{M}_{\boldsymbol{v} \overset{\imath}{\leftarrow} v}, \mathcal{M}_{\boldsymbol{v} \overset{\imath}{\leftarrow} w} \in \mathbb{M}$.

To lower bound the Hausdorff distance, we focus on the triple $\imath = (s_*, a_*, h_*)$ and we enforce the convenience of action a_0 over action a_*. For both MDP\R $\mathcal{M}_{\boldsymbol{v} \overset{\imath}{\leftarrow} v}$ and $\mathcal{M}_{\boldsymbol{v} \overset{\imath}{\leftarrow} w}$, let $r^v \in \mathcal{R}_{\mathcal{M}_{\boldsymbol{v} \overset{\imath}{\leftarrow} v}}$ and $r^w \in \mathcal{R}_{\mathcal{M}_{\boldsymbol{v} \overset{\imath}{\leftarrow} w}}$, we have:

$$r^v_{h_*}(s_*, a_0) + \frac{1}{\overline{S}} \sum_{l=h_*+1}^{H} \sum_{j=1}^{\overline{S}} r^v_l(s'_j) \geqslant r^v_{h_*}(s_*, a_*) + \sum_{l=h_*+1}^{H} \sum_{j=1}^{\overline{S}} \frac{1+\epsilon' v_j}{\overline{S}} r^v_l(s'_j)$$

$$\implies r^v_{h_*}(s_*, a_0) \geqslant r^v_{h_*}(s_*, a_*) + \frac{\epsilon'}{\overline{S}} \sum_{j=1}^{\overline{S}} v_j \sum_{l=h_*+1}^{H} r^v_l(s'_j).$$

$$r^w_{h_*}(s_*, a_0) + \frac{1}{\overline{S}} \sum_{l=h_*+1}^{H} \sum_{j=1}^{\overline{S}} r^w_l(s'_j) \geqslant r^w_{h_*}(s_*, a_*) + \sum_{l=h_*+1}^{H} \sum_{j=1}^{\overline{S}} \frac{1+\epsilon' w_j}{\overline{S}} r^w_l(s'_j)$$

$$\implies r^w_{h_*}(s_*, a_0) \geqslant r^w_{h_*}(s_*, a_*) + \frac{\epsilon'}{\overline{S}} \sum_{j=1}^{\overline{S}} w_j \sum_{l=h_*+1}^{H} r^w_l(s'_j). \tag{10}$$

In order to lower bound the Hausdorff distance $\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}v},\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w}\right)$, we proceed as in the proof of Theorem B.2 and we set for $\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}v}$:

$$r_l^v(s_j')=-v_j,\, r_{h_*}^v(s_*,a_*)=1,\, r_{h_*}^v(s_*,a_0)=1-\epsilon'(H-h_*),$$

where we enforce $\epsilon'\leqslant\min_{h_*\in[\![3,\overline{H}+2]\!]}1/(H-h_*)=1/(H-3)\leqslant 1/4$ for $H\geqslant 7$. We now want to find the closest reward function r^w for the instance $\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w}$, recalling that there are at least $\overline{S}/16$ components of the vectors v and w that are different. Clearly, we can set $r_l^w(s_j')=r_l^v(s_j')=-v_j$ for all $j\in[\![\overline{S}]\!]$ in which $v_j=w_j$ since this will not increase the Hausdorff distance and will make the constraint in Equation (10) less restrictive. For symmetry reasons, we can limit our reasoning to the case in which $v_j=-1$ and $w_j=1$ for the terms j in which they are different. This way, we have $r_l^v(s_j')=1$ and the constraint becomes:

$$\begin{aligned}\underbrace{r_{h_*}^w(s_*,a_0)}_{=:y}\geqslant\underbrace{r_{h_*}^w(s_*,a_*)}_{=:x}-\frac{N_{v,w}}{\overline{S}}\epsilon'(H-h_*)&\\
+\left(1-\frac{N_{v,w}}{\overline{S}}\right)\epsilon'(H-h_*)&\underbrace{\frac{1}{\overline{S}(H-h_*)\left(1-\frac{N_{v,w}}{\overline{S}}\right)}\sum_{j:v_j\neq w_j}^{\overline{S}}\sum_{l=h_*+1}^{H}r_l^w(s_j')}_{=:z},\end{aligned}$$

where $N_{v,w}=\sum_{j=1}^{\overline{S}}\mathbb{1}\{v_j=w_j\}$. Notice that $z\in[-1,1]$. Let $\alpha=\frac{N_{v,w}}{\overline{S}}$, the Hausdorff distance can be lower bounded by:

$$\begin{aligned}\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}v},\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w}\right)\geqslant&\min_{\substack{x,y,z\in[-1,1]\\ y\geqslant x-\alpha\epsilon'(H-h_*)+(1-\alpha)\epsilon'(H-h_*)z}}\max\left\{|x-1|,|y-(1-\epsilon'(H-h_*))|,|z-1|\right\}\\
\geqslant&\min_{\substack{x,y\in[-1,1]\\ y\geqslant x-\alpha\epsilon'(H-h_*)}}\max\left\{|x-1|,|y-(1-\epsilon'(H-h_*))|\right\}\\
=&\frac{1}{2}(1-\alpha)\epsilon'(H-h_*)\geqslant\frac{\epsilon'}{32}(H-h_*),\end{aligned}$$

where the first inequality derives from considering the aggregate term z instead of the individual rewards $r_l^w(s_j')$ (observing that in the base instance $\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}v}$ the corresponding z term takes value 1), the second inequality follows from the fact that to have a Hausdorff distance smaller than 1, we must take $z>0$ at least and, consequently, we ignore the term $|z-1|$ in the maximum and we take $z=0$ as the less restrictive case in the constraint involving x and y (being $(1-\alpha)\epsilon'(H-h_*)\geqslant 0$), and the third inequality is obtained by recalling that $1-\alpha\geqslant\frac{1}{16}$ for the packing argument.

We enforce the following constraint on this quantity:

$$\forall h_*\in[\![3,\overline{H}+2]\!]:\frac{\epsilon'}{32}(H-h_*)\geqslant 2\epsilon\implies\epsilon'\geqslant\max_{h_*\in[\![3,\overline{H}+2]\!]}\frac{64\epsilon}{H-h_*}=\frac{64\epsilon}{H-\overline{H}-2}.\tag{11}$$

Notice that $\epsilon'\leqslant 1/2$ whenever $H\geqslant\overline{H}+130$. This latter condition, together with $\epsilon'\leqslant 1/(H-3)$, implies $\epsilon\leqslant\frac{H-\overline{H}-2}{64(H-3)}$ that is satisfied for $\epsilon\leqslant 1/64$.

Step 3: Lower bounding Probability Let us consider an (ϵ,δ)-correct algorithm $\mathfrak{A}$ that outputs the estimated feasible set $\widehat{\mathcal{R}}$. Thus, consider $\imath\in\mathcal{I}$ and $\boldsymbol{v}\in\overline{\mathcal{V}}^{\mathcal{I}}$, we can lower bound the error probability:

$$\begin{aligned}\delta\geqslant&\sup_{\text{all }\mathcal{M}\text{ MDP\R and expert policies }\pi}\mathbb{P}_{(\mathcal{M},\pi),\mathfrak{A}}\left(\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\mathcal{M}},\widehat{\mathcal{R}}\right)\geqslant\epsilon\right)\\
\geqslant&\sup_{\mathcal{M}\in\mathbb{M}}\mathbb{P}_{(\mathcal{M},\pi),\mathfrak{A}}\left(\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\mathcal{M}},\widehat{\mathcal{R}}\right)\geqslant\epsilon\right)\\
\geqslant&\max_{w\in\overline{\mathcal{V}}}\mathbb{P}_{(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w},\pi),\mathfrak{A}}\left(\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w}},\widehat{\mathcal{R}}\right)\geqslant\epsilon\right).\end{aligned}$$

For every $\imath\in\mathcal{I}$ and $\boldsymbol{v}\in\overline{\mathcal{V}}^{\mathcal{I}}$, let us define the *identification function* (whose dependence on the estimated feasible reward set $\widehat{\mathcal{R}}$ is omitted to avoid a too heavy notation):

$$\Psi_{\imath,\boldsymbol{v}}:=\operatorname*{arg\,min}_{w\in\overline{\mathcal{V}}}\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w}},\widehat{\mathcal{R}}\right).$$

Let $w\in\overline{\mathcal{V}}$. If $\Psi_{\imath,\boldsymbol{v}}=w$, then, $\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}_{\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}\Psi_{\imath,\boldsymbol{v}}}},\mathcal{R}_{\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w}})=0$. Otherwise, if $\Psi_{\imath,\boldsymbol{v}}\neq w$, we have:

$$\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}_{\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}\Psi_{\imath,\boldsymbol{v}}}},\mathcal{R}_{\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w}})\leqslant\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}_{\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}\Psi_{\imath,\boldsymbol{v}}}},\widehat{\mathcal{R}})+\mathcal{H}_{d^{\mathrm{G}}}(\widehat{\mathcal{R}},\mathcal{R}_{\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w}})\leqslant 2\mathcal{H}_{d^{\mathrm{G}}}(\widehat{\mathcal{R}},\mathcal{R}_{\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w}}),$$

where the first inequality follows from triangular inequality and the second one from the definition of identification function $\Psi_{\imath,\boldsymbol{v}}$. From Equation (11), we have that $\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}_{\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}\Psi_{\imath,\boldsymbol{v}}}},\mathcal{R}_{\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w}})\geqslant 2\epsilon$. Thus, it follows that $\mathcal{H}_{d^{\mathrm{G}}}(\widehat{\mathcal{R}},\mathcal{R}_{\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w}})\geqslant\epsilon$. This implies the following inclusion of events for $w\in\overline{\mathcal{V}}$:

$$\left\{\mathcal{H}_{d^{\mathrm{G}}}(\widehat{\mathcal{R}},\mathcal{R}_{\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w}})\geqslant\epsilon\right\}\supseteq\{\Psi_{\imath,\boldsymbol{v}}\neq w\}.$$

Thus, we can proceed by lower bounding the probability:

$$\begin{aligned}\max_{w\in\overline{\mathcal{V}}}\mathop{\mathbb{P}}_{(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w},\pi),\mathfrak{A}}\left(\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w}},\widehat{\mathcal{R}}\right)\geqslant\epsilon\right)&\geqslant\max_{w\in\overline{\mathcal{V}}}\mathop{\mathbb{P}}_{(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w},\pi),\mathfrak{A}}(\Psi_{\imath,\boldsymbol{v}}\neq w)\\&\geqslant\frac{1}{|\overline{\mathcal{V}}|}\sum_{w\in\overline{\mathcal{V}}}\mathop{\mathbb{P}}_{(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w},\pi),\mathfrak{A}}(\Psi_{\imath,\boldsymbol{v}}\neq w),\end{aligned}$$

where the second inequality follows from bounding the maximum of probability with the average. We can now apply the Fano's inequality (Theorem D.2) with reference probability $\mathbb{P}_0=\mathbb{P}_{(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}0},\pi),\mathfrak{A}}$, $\mathbb{P}_w=\mathbb{P}_{(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w},\pi),\mathfrak{A}}$, and $\mathcal{A}_w=\{\Psi_{\imath,\boldsymbol{v}}\neq w\}$:

$$\frac{1}{|\overline{\mathcal{V}}|}\sum_{w\in\overline{\mathcal{V}}}\mathop{\mathbb{P}}_{(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w},\pi),\mathfrak{A}}(\Psi_{\imath,\boldsymbol{v}}\neq w)\geqslant 1-\frac{1}{\log|\overline{\mathcal{V}}|}\left(\frac{1}{|\overline{\mathcal{V}}|}\sum_{w\in\overline{\mathcal{V}}}D_{\mathrm{KL}}\left(\mathop{\mathbb{P}}_{(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w},\pi),\mathfrak{A}},\mathop{\mathbb{P}}_{(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}0},\pi),\mathfrak{A}}\right)-\log 2\right).\tag{12}$$

Step 4: KL-divergence Computation Let $\mathcal{M}$ be an instance, we denote with $\mathbb{P}_{(\mathcal{M},\pi),\mathfrak{A}}$ the joint probability distribution of all events realized by the execution of the algorithm in the MDP\R (the presence of π is irrelevant as we assume it known):

$$\mathop{\mathbb{P}}_{(\mathcal{M},\pi),\mathfrak{A}}=\prod_{t=1}^{\tau}\rho_t(s_t,a_t,h_t|H_{t-1})p_{h_t}(s_t'|s_t,a_t).$$

where ρ_t is the sampling distribution induced by the algorithm $\mathfrak{A}$ and $H_{t-1}=(s_1,a_1,h_1,s_1',\dots,s_{t-1},a_{t-1},h_{t-1},s_{t-1}')$ is the history up to time $t-1$. Let $\imath\in\mathcal{I}$ and $v\in\overline{\mathcal{V}}$ and denote with $p^{\boldsymbol{v}\overset{\imath}{\leftarrow}0}$ and $p^{\boldsymbol{v}\overset{\imath}{\leftarrow}w}$ the transition models associated with $\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}0}$ and $\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w}$. Let us now move to the KL-divergence and denoting $\imath=(s_*,a_*,h_*)$: Thus, we have:

$$\begin{aligned}D_{\mathrm{KL}}\left(\mathop{\mathbb{P}}_{(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w},\pi),\mathfrak{A}},\mathop{\mathbb{P}}_{(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}0},\pi),\mathfrak{A}}\right)&=\mathop{\mathbb{E}}_{(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w},\pi),\mathfrak{A}}\left[\sum_{t=1}^{\tau}D_{\mathrm{KL}}\left(p_{h_t}^{\boldsymbol{v}\overset{\imath}{\leftarrow}w}(\cdot|s_t,a_t),p_{h_t}^{\boldsymbol{v}\overset{\imath}{\leftarrow}0}(\cdot|s_t,a_t)\right)\right]\\&\leqslant\mathop{\mathbb{E}}_{(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w},\pi),\mathfrak{A}}\left[N_{h_*}^{\tau}(s_*,a_*)\right]D_{\mathrm{KL}}\left(p_{h_*}^{\boldsymbol{v}\overset{\imath}{\leftarrow}w}(\cdot|s_*,a_*),p_{h_*}^{\boldsymbol{v}\overset{\imath}{\leftarrow}0}(\cdot|s_*,a_*)\right)\\&\leqslant 2(\epsilon')^2\mathop{\mathbb{E}}_{(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w},\pi),\mathfrak{A}}\left[N_{h_*}^{\tau}(s_*,a_*)\right],\end{aligned}$$

having observed that the transition models differ in $\imath=(s_*,a_*,h_*)$ and defined $N_{h_*}^{\tau}(s_*,a_*)=\sum_{t=1}^{\tau}\mathbb{1}\{(s_t,a_t,h_t)=(s_*,a_*,h_*)\}$ and the last passage is obtained by Lemma D.4 with $D=\overline{S}$. Plugging into Equation (12), we obtain:

$$\delta\geqslant\frac{1}{|\overline{\mathcal{V}}|}\sum_{w\in\overline{\mathcal{V}}}\mathop{\mathbb{P}}_{(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w},\pi),\mathfrak{A}}(\Psi_{\imath,\boldsymbol{v}}\neq w)\implies\frac{1}{|\overline{\mathcal{V}}|}\sum_{w\in\overline{\mathcal{V}}}\mathop{\mathbb{E}}_{(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w},\pi),\mathfrak{A}}\left[N_{h_*}^{\tau}(s_*,a_*)\right]\geqslant\frac{(1-\delta)\log|\overline{\mathcal{V}}|-\log 2}{2(\epsilon')^2}.$$

Since the derivation is carried out for every $\imath\in\mathcal{I}$ and $\boldsymbol{v}\in\overline{\mathcal{V}}^{\mathcal{I}}$, we can perform the summation over $\imath$ and the average over $\boldsymbol{v}$:

$$\sum_{\imath\in\mathcal{I}}\frac{1}{|\overline{\mathcal{V}}|^{|\mathcal{I}|}}\sum_{\boldsymbol{v}\in\overline{\mathcal{V}}^{\mathcal{I}}}\frac{1}{|\overline{\mathcal{V}}|}\sum_{w\in\overline{\mathcal{V}}}\mathop{\mathbb{E}}_{(\mathcal{M}_{\boldsymbol{v}\overset{\imath}{\leftarrow}w},\pi),\mathfrak{A}}\left[N_{h_*}^{\tau}(s_*,a_*)\right]=\frac{1}{|\overline{\mathcal{V}}|^{|\mathcal{I}|}}\sum_{\boldsymbol{v}\in\overline{\mathcal{V}}^{\mathcal{I}}}\sum_{\imath\in\mathcal{I}}\mathop{\mathbb{E}}_{(\mathcal{M}_{\boldsymbol{v}},\pi),\mathfrak{A}}\left[N_{h_*}^{\tau}(s_*,a_*)\right]$$

$$\geqslant \overline{SAH}\frac{(1-\delta)\log|\overline{\mathcal{V}}|-\log 2}{2(\epsilon')^2}.$$

Notice that we get a guarantee on a mean under the uniform distribution of the instances of the sample complexity. Thus, there must exist one $\boldsymbol{v}^{\text{hard}}\in\overline{\mathcal{V}}$ such that:

$$\mathop{\mathbb{E}}_{(\mathcal{M}_{\boldsymbol{v}^{\text{hard}},\pi}),\mathfrak{A}}[\tau]\geqslant\sum_{\imath\in\mathcal{I}}\mathop{\mathbb{E}}_{(\mathcal{M}_{\boldsymbol{v}^{\text{hard}},\pi}),\mathfrak{A}}\left[N^{\tau}_{h_*}(s_*,a_*)\right]\geqslant\overline{SAH}\frac{(1-\delta)\log|\overline{\mathcal{V}}|-\log 2}{2(\epsilon')^2}.$$

Then, we select $\delta\leqslant 1/2$, recall that $|\overline{\mathcal{V}}|\geqslant 2^{\overline{S}/5}$, we get:

$$\mathop{\mathbb{E}}_{(\mathcal{M}_{\boldsymbol{v}^{\text{hard}},\pi}),\mathfrak{A}}[\tau]\geqslant\overline{SAH}\frac{\overline{S}/10-\log 2}{2(\epsilon')^2}=\overline{SAH}\frac{(H-\overline{H}-2)^2(\overline{S}/10-\log 2)}{8192\epsilon^2}$$

The number of states is given by $S=|\mathcal{S}|=2\overline{S}+2$, the number of actions is given by $A=|\mathcal{A}|=\overline{A}+1$. Let us first consider the time-homogeneous case, i.e., $\overline{H}=1$, for $S\geqslant 16$, $A\geqslant 2$, $H\geqslant 130$, we have:

$$\mathop{\mathbb{E}}_{(\mathcal{M}_{\boldsymbol{v}^{\text{hard}},\pi}),\mathfrak{A}}[\tau]\geqslant\Omega\left(\frac{S^2AH^2}{\epsilon^2}\right).$$

For the time inhomogeneous case, we select $\overline{H}=H/2$, to get, under the same conditions:

$$\mathop{\mathbb{E}}_{(\mathcal{M}_{\boldsymbol{v}^{\text{hard}},\pi}),\mathfrak{A}}[\tau]\geqslant\Omega\left(\frac{S^2AH^3}{\epsilon^2}\right).$$

□

B.4 Proofs of Section 6

Theorem 6.1 (Sample Complexity of `US-IRL`)**.** *Let $\epsilon>0$ and $\delta\in(0,1)$, `US-IRL` is (ϵ,δ)-PAC for d^{G}-IRL and with probability at least $1-\delta$ it stops after τ samples with:*

- *if the transition model p is time-inhomogeneous:*

$$\tau\leqslant\frac{8H^3SA}{\epsilon^2}\left(\log\left(\frac{SAH}{\delta}\right)+(S-1)C\right),$$

where $C=1+\log(1+(64H^4)/(\epsilon^4(S-1))\times\left(\log((SAH)/\delta)+\sqrt{e}(S-1+\sqrt{S-1})\right)^2\big)$;

- *if the transition model p is time-homogeneous:*

$$\tau\leqslant\frac{8H^2SA}{\epsilon^2}\left(\log\left(\frac{SA}{\delta}\right)+(S-1)\widetilde{C}\right),$$

where $\widetilde{C}=1+\log(1+(64H^4)/(\epsilon^4(S-1))\times\left(\log((SA)/\delta)+\sqrt{e}(S-1+\sqrt{S-1})\right)^2\big)$.

Proof. We start with the case in which the transition model is time-inhomogeneous. In this case, we introduce the following good event:

$$\mathcal{E}:=\left\{\forall t\in\mathbb{N},\,\forall(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]:D_{\mathrm{KL}}\Big(\widehat{p}^{\,t}_h(\cdot|s,a),p_h(\cdot|s,a)\Big)\leqslant\frac{\beta\big(n^t_h(s,a),\delta\big)}{n^t_h(s,a)}\right\},$$

where p_h is the true transition model and $\widehat{p}^{\,t}_h$ is its estimate via Equation (3) at time t. Thanks to Lemma B.4, we have that $\mathbb{P}_{(\mathcal{M},\pi^E),\mathfrak{A}}(\mathcal{E})\geqslant 1-\delta$. Thus, under the good event $\mathcal{E}$, we apply Theorem 3.2:

$$\begin{aligned}\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R},\widehat{\mathcal{R}}^{\tau})&\leqslant\frac{2\rho^{\mathrm{G}}((\mathcal{M},\pi^E),(\widehat{\mathcal{M}}^t,\widehat{\pi}^{E,t}))}{1+\rho^{\mathrm{G}}((\mathcal{M},\pi^E),(\widehat{\mathcal{M}}^t,\widehat{\pi}^{E,t}))}\\&\leqslant 2\rho^{\mathrm{G}}((\mathcal{M},\pi^E),(\widehat{\mathcal{M}}^t,\widehat{\pi}^{E,t}))\\&\leqslant 2\max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]}(H-h+1)\left(\left|\mathbb{1}_{\{\pi^E_h(a|s)=0\}}-\mathbb{1}_{\{\widehat{\pi}^{E,t}_h(a|s)=0\}}\right|+\left\|p_h(\cdot|s,a)-\widehat{p}^{\,t}_h(\cdot|s,a)\right\|_1\right)\end{aligned}$$

$$\leqslant 2 \max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} (H-h+1)\left\|p_h(\cdot|s,a)-\widehat{p}_h^t(\cdot|s,a)\right\|_1$$

$$\leqslant 2\sqrt{2} \max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} (H-h+1)\sqrt{D_{\text{KL}}\left(\widehat{p}_h^t(\cdot|s,a),p_h(\cdot|s,a)\right)} = \max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} \mathcal{C}_h^t(s,a),$$

where we exploited the fact that the expert's policy is known in the last but one passage and used Pinsker's inequality in the last passage. When US-IRL stops we have that $\max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} \mathcal{C}_h^t(s,a) \leqslant \epsilon$ and, consequently, for all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]$ we have:

$$\max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} \mathcal{C}_h^t(s,a) = \max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} 2\sqrt{2}(H-h+1)\sqrt{\frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)}} \leqslant \epsilon.$$

Thus, the algorithm stops at the smallest t such that:

$$\begin{aligned}\implies n_h^t(s,a) &\geqslant \frac{8(H-h+1)^2\beta\left(n_h^t(s,a),\delta\right)}{\epsilon^2} \\ &= \frac{8(H-h+1)^2}{\epsilon^2}\left(\log(SAH/\delta)+(S-1)\log(e(1+n_h^t(s,a)/(S-1)))\right).\end{aligned}$$

Thus, by applying Lemma 15 of [14], we obtain:

$$\begin{aligned}n_h^\tau(s,a) \leqslant{}& \frac{8(H-h+1)^2}{\epsilon^2}\left(\log\left(\frac{SAH}{\delta}\right)+(S-1)\right.\\ &\left.\times\left(1+\log\left(1+\frac{64(H-h+1)^4}{\epsilon^4(S-1)}\left(\log\left(\frac{SAH}{\delta}\right)+\sqrt{e}(S-1+\sqrt{S-1})\right)^2\right)\right)\right).\end{aligned}$$

By recalling that $\tau = SAHn_h^\tau(s,a)$, and bounding $H-h+1\leqslant H$, we obtain:

$$\begin{aligned}\tau \leqslant{}& \frac{8H^3SA}{\epsilon^2}\left(\log\left(\frac{SAH}{\delta}\right)+(S-1)\right.\\ &\left.\times\left(1+\log\left(1+\frac{64H^4}{\epsilon^4(S-1)}\left(\log\left(\frac{SAH}{\delta}\right)+\sqrt{e}(S-1+\sqrt{S-1})\right)^2\right)\right)\right).\end{aligned}$$

If the transition model is time-homogeneous, we suppress the subscript h and the algorithm US-IRL will merge together all the samples collected at different stages h. Let us define $n^t(s,a) = \sum_{h=1}^H n_h^t(s,a)$ and $n^t(s,a,s') = \sum_{h=1}^H n_h^t(s,a,s')$. Now the transition model will be estimated straightforwardly as follows:

$$\widehat{p}^t(s'|s,a) := \begin{cases} \frac{n^t(s,a,s')}{n^t(s,a)} & \text{if } n^t(s,a)>0 \\ \frac{1}{S} & \text{otherwise}\end{cases}.$$

Let us consider now the following good event:

$$\widetilde{\mathcal{E}} := \left\{\forall t\in\mathbb{N},\, \forall(s,a)\in\mathcal{S}\times\mathcal{A}\colon D_{\text{KL}}\left(\widehat{p}^t(\cdot|s,a),p(\cdot|s,a)\right)\leqslant\frac{\widetilde{\beta}\left(n^t(s,a),\delta\right)}{n^t(s,a)}\right\}.$$

Thanks to Lemma B.4, we have that $\mathbb{P}_{(\mathcal{M},\pi^E),\mathfrak{A}}(\widetilde{\mathcal{E}}) \geqslant 1-\delta$. Thus, in such a case, thanks to Theorem 3.2, we have:

$$\mathcal{H}_{d^{\text{G}}}(\mathcal{R},\widehat{\mathcal{R}}^\tau) \leqslant 2\sqrt{2} \max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} (H-h+1)\sqrt{D_{\text{KL}}\left(\widehat{p}^t(\cdot|s,a),p(\cdot|s,a)\right)} = \max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} \widetilde{\mathcal{C}}_h^t(s,a).$$

The algorithm, therefore, stops as soon as:

$$\begin{aligned}\max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} \widetilde{\mathcal{C}}_h^t(s,a) &= \max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} 2\sqrt{2}(H-h+1)\sqrt{\frac{\widetilde{\beta}\left(n^t(s,a),\delta\right)}{n^t(s,a)}} \\ &= \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} 2\sqrt{2}H\sqrt{\frac{\widetilde{\beta}\left(n^t(s,a),\delta\right)}{n^t(s,a)}} \leqslant \epsilon.\end{aligned}$$

This allows us to compute the maximum value of $n^\tau(s,a)$:

$$n^\tau(s,a) \leqslant \leqslant \frac{8H^2}{\epsilon^2}\left(\log\left(\frac{SA}{\delta}\right)+(S-1)\right.$$
$$\left.\times\left(1+\log\left(1+\frac{64H^4}{\epsilon^4(S-1)}\left(\log\left(\frac{SA}{\delta}\right)+\sqrt{e}(S-1+\sqrt{S-1})\right)^2\right)\right)\right).$$

Recalling that $\tau = SAn^\tau(s,a)$, we obtain:

$$\tau \leqslant \frac{8H^2SA}{\epsilon^2}\left(\log\left(\frac{SA}{\delta}\right)+(S-1)\right.$$
$$\left.\times\left(1+\log\left(1+\frac{64H^4}{\epsilon^4(S-1)}\left(\log\left(\frac{SA}{\delta}\right)+\sqrt{e}(S-1+\sqrt{S-1})\right)^2\right)\right)\right).$$

□

Lemma B.4. *The following statements hold:*

- *for $\beta(n,\delta) = \log(SAH/\delta) + (S-1)\log\big(e(1+n/(S-1)\big)$, we have that $\mathbb{P}(\mathcal{E}) \geqslant 1-\delta$;*
- *for $\widetilde{\beta}(n,\delta) = \log(SA/\delta) + (S-1)\log\big(e(1+n/(S-1)\big)$, we have that $\mathbb{P}(\widetilde{\mathcal{E}}) \geqslant 1-\delta$.*

Proof. Let us start with the first statement. Similarly to Lemma 10 of [14], we apply first a union bound and, then, technical Proposition 1 of [13] (also reported as Lemma D.3 for completeness) to concentrate the KL-divergence:

$$\mathbb{P}(\mathcal{E}^c) = \mathbb{P}\left(\exists t \in \mathbb{N}, \exists (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[\![H]\!] : D_{\text{KL}}\left(\widehat{p}_h^t(\cdot|s,a), p_h(\cdot|s,a)\right) \geqslant \frac{\beta\big(n_h^t(s,a),\delta\big)}{n_h^t(s,a)}\right)$$
$$\leqslant \sum_{h\in[\![H]\!]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbb{P}\left(\exists t \in \mathbb{N} : D_{\text{KL}}\left(\widehat{p}_h^t(\cdot|s,a), p_h(\cdot|s,a)\right) \geqslant \frac{\beta\big(n_h^t(s,a),\delta\big)}{n_h^t(s,a)}\right)$$
$$\leqslant \sum_{h\in[\![H]\!]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\delta}{SAH} = \delta.$$

The proof of the second statement is analogous having simply observed that the union bound has to be performed over $\mathcal{S}\times\mathcal{A}$ only. □

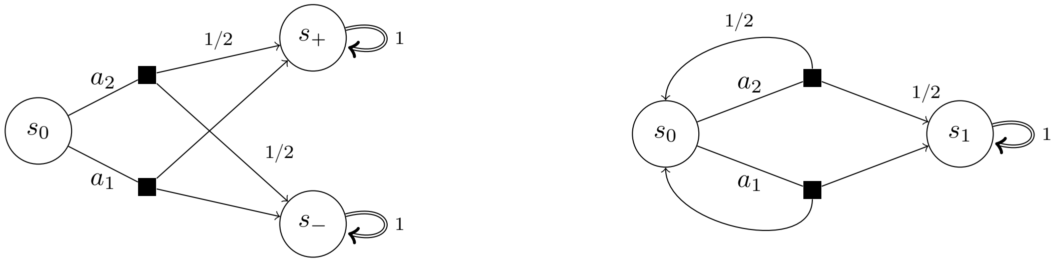
B.5 Non-Lipschitz Continuous Restricted Feasible Reward Sets

In this section, we illustrate three cases *restricted* feasible reward sets that turn out not to fulfill the thesis of Theorem 3.2. These examples, representing *strict subsets* of the feasible reward functions of Equation (2), are obtained by enforcing common conditions: state-only reward function $r_h(s)$ (Example B.1), time-homogeneous reward function $r(s,a)$ (Example B.2), and β-margin reward function (Example B.3). We present counter-examples in which in front of ϵ-close transition models, the induced feasible sets are far apart by a constant independent of ϵ. For space reasons, we report the complete derivation in Appendix **??**.

Example B.1 (State-only reward $r_h(s)$)**.** *State-only reward functions have been widely considered in many IRL approaches [e.g., 25, 1, 35, 15]. We formalize the state-only feasible reward set as follows:*

$$\mathcal{R}_{\textit{state}} = \mathcal{R} \cap \{\forall (s,a,a',h) : r_h(s,a) = r_h(s,a')\}.$$

Consider the MDP\R of Figure 5a with $H=2$, $\pi_h^E(s_0) = \widehat{\pi}_h^E(s_0) = a_1$ with $h \in \{1,2\}$. Set $p_1(s_+|s_0,a_1) = 1/2+\epsilon/4$ and $\widehat{p}_1(s_+|s_0,a_1) = 1/2-\epsilon/4$ and, thus, $\|p_1(\cdot|s_0,a_1) - \widehat{p}_1(\cdot|s_0,a_1)\|_1 = \epsilon$. Let us set $r_2(s_+) = 1$ and $r_2(s_-) = -1$, which makes π^E optimal under p. We observe that $\widehat{\mathcal{R}}$ is defined by $\widehat{r}_2(s_-) \leqslant \widehat{r}_2(s_+)$. Recalling that the rewards are bounded in $[-1,1]$, we have $\mathcal{H}_{d^G}(\mathcal{R}_{\textit{state}}, \widehat{\mathcal{R}}_{\textit{state}}) \geqslant 1$.



(a)

(b)

Figure 5: The MDP\R employed in the examples of Section B.5. $\Longrightarrow$ denotes a transition executed for multiple actions.

Proof. For the MDP\R $\mathcal{M}$, in order to make $\pi_1^E(s_0)=a_1$ optimal, we have to enforce:

$$\begin{aligned}&r_1(s_0)+\frac{2+\epsilon}{4}r_2(s_+)+\frac{2-\epsilon}{4}r_2(s_-)\geqslant r_1(s_0)+\frac{1}{2}r_2(s_+)+\frac{1}{2}r_2(s_-)\\&\implies r_2(s_+)\geqslant r_2(s_-).\end{aligned}$$

Similarly, to make $\widehat{\pi}_1^E(s_0)=a_1$, we have for $\widehat{\mathcal{M}}$:

$$\begin{aligned}&\widehat{r}_1(s_0)+\frac{2-\epsilon}{4}\widehat{r}_2(s_+)+\frac{2+\epsilon}{4}\widehat{r}_2(s_-)\geqslant \widehat{r}_1(s_0)+\frac{1}{2}\widehat{r}_2(s_+)+\frac{1}{2}\widehat{r}_2(s_-)\\&\implies \widehat{r}_2(s_+)\leqslant \widehat{r}_2(s_-).\end{aligned}$$

Thus, if we set $r_2(s_-)=1$ and $r_2(s_+)=-1$, we have:

$$\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}_{\text{state}},\widehat{\mathcal{R}}_{\text{state}})\geqslant \min_{\substack{\widehat{r}_2(s_-),\widehat{r}_2(s_+)\in[-1,1]\\ \widehat{r}_2(s_+)\leqslant \widehat{r}_2(s_-)}} \max\{|1-\widehat{r}_2(s_-)|,|-1-\widehat{r}_2(s_+)|\}=1,$$

by setting $\widehat{r}_2(s_-)=\widehat{r}_2(s_+)=0$. □

Example B.2 (Time-homogeneous reward $r(s,a)$)**.** *Time-homogeneous reward functions have been employed in several RL [e.g., 6] and IRL settings [e.g., 19]. We formalize the time-homogeneous feasible reward set as follows:*

$$\mathcal{R}_{hom}=\mathcal{R}\cap\{\forall(s,a,h,h'): r_h(s,a)=r_{h'}(s,a)\}.$$

Consider the MDP\R of Figure 5b with $H=2$, $\pi_1^E(s_0)=\widehat{\pi}_1^E(s_0)=a_1$ and $\pi_2^E(s_0)=\widehat{\pi}_2^E(s_0)=a_2$. For $h\in\{1,2\}$, we set $p_h(s_0|s_0,a_1)=1/2+\epsilon/4$ and $\widehat{p}_h(s_0|s_0,a_1)=1/2-\epsilon/4$, thus, $\|p_h(\cdot|s_0,a_1)-\widehat{p}_h(\cdot|s_0,a_1)\|_1=\epsilon$. We set $r(s_0,a_1)=1$, $r(s_0,a_2)=1-\epsilon/6$, and $r(s_1,a_1)=r(s_1,a_2)=1/2$ making π^E optimal. We can prove that $\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}_{hom},\widehat{\mathcal{R}}_{hom})\geqslant 1/4$.

Proof. Consider the MDP\R $\mathcal{M}$ and we set $r(s_0,a_1)=1$, $r(s_0,a_2)=1-\epsilon/12$, and $r(s_1,a)=1/2$ for $a\in\{a_1,a_2\}$. We immediately observe that π^E is optimal since for $h=2$, $r(s_0,a_1)\geqslant r(s_0,a_2)$ and for $h=1$:

$$\begin{aligned}&r(s_0,a_2)+\frac{2+\epsilon}{4}r(s_0,a_1)+\frac{2-\epsilon}{4}r(s_1,a)\geqslant r(s_0,a_1)+\frac{1}{2}r(s_0,a_1)+\frac{1}{2}r(s_1,a)\\&\iff r(s_0,a_2)+\left(\frac{\epsilon}{4}-1\right)r(s_0,a_1)-\frac{\epsilon}{4}r(s_1,a)\geqslant 0\\&\iff 1-\frac{\epsilon}{12}+\frac{\epsilon}{4}-1-\frac{\epsilon}{8}\geqslant 0.\end{aligned}$$

Consider now the alternative MDP\R $\widehat{\mathcal{M}}$, we have to enforce the following two conditions:

$$\widehat{r}(s_0,a_1)\geqslant\widehat{r}(s_0,a_2), \tag{13}$$

$$\begin{aligned}&\widehat{r}(s_0,a_2)+\frac{2-\epsilon}{4}\widehat{r}(s_0,a_1)+\frac{2+\epsilon}{4}\widehat{r}(s_1,a)\geqslant\widehat{r}(s_0,a_1)+\frac{1}{2}\widehat{r}(s_0,a_1)+\frac{1}{2}\widehat{r}(s_1,a)\\&\iff \widehat{r}(s_0,a_2)-\left(\frac{\epsilon}{4}+1\right)\widehat{r}(s_0,a_1)+\frac{\epsilon}{4}\widehat{r}(s_1,a)\geqslant 0.\end{aligned}\tag{14}$$

The way of enforcing Equation (13) that is less constraining for Equation (14) is setting $\widehat{r}(s_0,a_1)=\widehat{r}(s_0,a_2)$, to get:

$$-\frac{\epsilon}{4}\widehat{r}(s_0,a_1)+\frac{\epsilon}{4}\widehat{r}(s_1,a)\geqslant 0 \Longleftrightarrow \widehat{r}(s_1,a)\geqslant\widehat{r}(s_0,a_1).$$

This implies:

$$\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}_{\mathrm{hom}},\widehat{\mathcal{R}}_{\mathrm{hom}})\geqslant \min_{\substack{\widehat{r}(s_1,a),\widehat{r}(s_0,a_1)\in[-1,1]\\ \widehat{r}(s_1,a)\geqslant\widehat{r}(s_0,a_1)}} \max\left\{|1-\widehat{r}(s_0,a_1)|,\left|\frac{1}{2}-\widehat{r}(s_1,a)\right|\right\}\geqslant\frac{1}{4},$$

by setting $\widehat{r}(s_0,a_1)=\widehat{r}(s_1,a)=1/4$. □

Example B.3 (β-margin reward)**.** *A β-margin reward enforces a suboptimality gap of at least $\beta>0$ [25, 15]. We formalize it in the finite-horizon case with a sequence $\beta=(\beta_h)_{h\in[\![H]\!]}$, possibly different for every stage:*

$$\mathcal{R}_{\beta\text{-}mar}=\mathcal{R}\cap\{\forall(s,a,h): A_h^{\pi^E}(s,a;r)\in\{0\}\cup(-\infty,-\beta_h]\}.$$

Consider the MDP\R in Figure 5a with $\pi_h^E(s_0)=\widehat{\pi}_h^E(s_0)=a_1$ for $h\in\{1,2\}$. We set $p_1(s_+|s_0,a_1)=1/2+\epsilon$ and $\widehat{p}_1(s_+|s_0,a_1)=1/2-\epsilon$. We set for MDP\R $\mathcal{M}$ the reward function as $r_1(s_0,a)=0$ and $r_h(s_+,a)=-r_h(s_-,a)=1$ for $a\in\{a_1,a_2\}$ and $h\in[\![2,H]\!]$. In $(s_0,1)$ the suboptimality gap is $\beta_1=2+2\epsilon(H-1)$. By selecting $H\geqslant 1+1/\epsilon$, the feasible set $\widehat{\mathcal{R}}_{\beta\text{-}mar}$ is empty.

Proof. Concerning the MDP\R $\mathcal{M}$, we observe that by setting $r_1(s_0,a_1)=1$, $r_1(s_0,a_2)=-1$, and $r_h(s_+,a)=-r_h(s_-,a)=1$ for $a\in\{a_1,a_2\}$ and $h\in[\![2,H]\!]$, the policy π^E is optimal. In particular, in state-stage pair $(s_0,1)$ the suboptimality gap is given by $\beta_1=2+2\epsilon(H-1)$. To enforce the optimality of $\widehat{\pi}^E=\pi^E$ in the MDP\R $\widehat{\mathcal{M}}$, we have:

$$\widehat{r}_1(s_0,a_1)+\sum_{h=2}^{H}\frac{1}{2}\widehat{r}_h(s_+,a_1)+\frac{1}{2}\widehat{r}_h(s_-,a_1)\geqslant\widehat{r}_1(s_0,a_2)+\sum_{h=2}^{H}\frac{1}{2}\widehat{r}_h(s_+,a_1)+\frac{1}{2}\widehat{r}_h(s_-,a_1)+\beta_1$$
$$\Longleftrightarrow \widehat{r}_1(s_0,a_1)-\widehat{r}_1(s_0,a_2)\geqslant\beta_1.$$

Thus, if $\beta_1\geqslant 2$, we have that the feasible set $\widehat{\mathcal{R}}_{\beta\text{-sep}}$ is empty. Thus, we select $H\geqslant 1+1/\epsilon$ to have $\beta_1\geqslant 4$. □

These examples show that some common restrictions of the feasible reward set are not Lipschitz continuous w.r.t. the transition model and, more in general, w.r.t. the IRL problem. If the Lipschitz condition is violated, we argue that recovering the restricted feasible reward set efficiently by estimating the transition model is not possible. This is because as shown in the examples, arbitrary close transition models lead to restricted feasible reward sets with a finite non-zero distance. This suggests that the Lipschitz framework captures a *structural property* of the problem, being tightly connected to the possibility of learning the feasible reward set under certain restrictions.[15] The generalization of these examples to more abstract conditions for guaranteeing the Lipschitz continuity of the restricting feasible reward set is beyond the scope of the paper.

C Unknown Expert's Policy π^E

In this appendix, we extend the lower bounds and the algorithm for the case in which the expert's policy is unknown. Clearly, if the expert's policy is deterministic, under the generative model setting, its estimation is trivial as it suffices to query every state and stage (resp. state) exactly once for time-inhomogeneous (resp. time-homogeneous) policies, leading to $\mathbb{E}_{(\mathcal{M},\pi^E),\mathfrak{A}}[\tau]=HS$ (resp. $\mathbb{E}_{(\mathcal{M},\pi^E),\mathfrak{A}}[\tau]=S$). Thus, we consider a more general setting in which the expert's policy can be stochastic (still being optimal). Specifically, we consider the following assumption.

[15] We remark that this phenomenon can be interpreted as a limitation of the formulation of the IRL problem as recovering the feasible reward set by estimating the transition model and does not imply that, for instance, state-only rewards are not learnable in general.

Assumption C.1. *There exists a known constant* $\pi_{\min} \in (0,1]$ *such that every action played by the expert's policy* π^E *is played with at least probability* $\pi_{\min}$*:*

$$\forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!] : \pi_h^E(a|s) \in \{0\} \cup [\pi_{\min}, 1].$$

Intuitively, Assumption C.1 formalizes a form of identifiability for the policy. As already mentioned in Section 3, what matters for learning the feasible reward set is whether an action is played by the agent (not the corresponding probability). Assumption C.1 enforces that every optimal action must be played with a minimum (known) non-null probability $\pi_{\min}$. We shall show that if this assumption is violated, the problem becomes non-learnable.

C.1 Lower Bound

The following result provides a lower bound for learning the feasible reward set according to the PAC requirement of Definition 4.1 when the expert's policy is unknown, but the transition model is known. Clearly, one can combine this result with the ones of Section 5 to address the setting in which both the expert's policy and the transition model are unknown.

Theorem C.1. *Let* $\mathfrak{A} = (\mu, \tau)$ *be an* (ϵ, δ)*-PAC algorithm for* d^G*-IRL. Then, there exists an IRL problem* $(\mathcal{M}, \pi^E)$ *where* π^E *fulfills Assumption C.1 such that, if* $\epsilon \leqslant 1/2$*,* $\delta < 1/16$*,* $S \geqslant 7$*,* $A \geqslant 2$*, and* $H \geqslant 3$*, the number of samples* τ *is lower bounded in expectation by:*

- *if the expert's policy* π^E *is time-inhomogeneous:*

$$\mathop{\mathbb{E}}_{(\mathcal{M}, \pi^E), \mathfrak{A}} [\tau] \geqslant \frac{SH}{8 \log \frac{1}{1-\pi_{\min}}} \log\left(\frac{1}{\delta}\right).$$

- *if the expert's policy* π^E *is time-homogeneous:*

$$\mathop{\mathbb{E}}_{(\mathcal{M}, \pi^E), \mathfrak{A}} [\tau] \geqslant \frac{S}{4 \log \frac{1}{1-\pi_{\min}}} \log\left(\frac{1}{\delta}\right);$$

Before presenting the proof, let us comment the result. We observe that when Assumption C.1 is violated, i.e., $\pi_{\min} \to 0$, the sample complexity lower bound degenerates to infinity, proving that the problem becomes non-learnable.

Proof. **Step 1: Instances Construction** The hard MDP\R instances are depicted in Figure 6 in a semi-formal way. The state space is given by $\mathcal{S} = \{s_{\text{start}}, s_{\text{root}}, s_1, \dots, s_{\overline{S}}, s_{\text{sink}}\}$ and the action space is given by $\mathcal{A} = \{a_0, a_1, \dots, a_{\overline{A}}\}$. The transition model is described below and the horizon is $H \geqslant 3$. We introduce the constant $\overline{H} \in [\![H]\!]$, whose value will be chosen later. Let us observe, for now, that if $\overline{H} = 1$, the transition model is time-homogeneous.

The agent begins in state s_{start}, where every action has the same effect. Specifically, if the stage $h < \overline{H}$, then there is probability $1/2$ to remain in s_{start} and a probability $1/2$ to transition to s_{root}. Instead, if $h \geqslant \overline{H}$, the state transitions to s_{root} deterministically. From state s_{root}, every action has the same effect and the state transitions with equal probability $1/\overline{S}$ to a state s_i with $i \in [\![\overline{S}]\!]$. In all states s_i, apart from a specific one, i.e., state s_*, the expert's policy plays action a_0 deterministically, i.e., $\pi_h^E(a_0|s_i) = 1$ and the state transitions deterministically to s_{sink}. In state s_* the expert's policy plays a_0 as the other ones if the stage $h \neq h_*$, where $h_* \in [\![H]\!]$ is a predefined stage. If, instead, $h = h_*$, the expert's action plays a_0 w.p. $1 - \pi_{\min}$ and a specific action a_* w.p. $\pi_{\min} \in [0, 1/2]$. Then, the transition is deterministic to state s_{sink}. Notice that, having fixed $\overline{H}$, the possible values of h_* are $\{3, \dots, 2 + \overline{H}\}$. State s_{sink} is an absorbing state.

Let us consider the base instance π_0 in which the expert's policy always plays action a_0 deterministically.[16] Additionally, by varying the pair $\ell := (s_*, h_*) \in \{s_1, \dots, s_{\overline{S}}\} \times [\![3, \overline{H} + 2]\!] =: \mathcal{J}$, we can construct the class of instances denoted by $\mathbb{M} = \{\pi_\ell : \ell \in \{0\} \cup \mathcal{J}\}$.

Step 2: Feasible Set Computation Let us consider an instance $\pi_\ell \in \mathbb{M}$, we now seek to provide a lower bound to the Hausdorff distance $\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}_{\pi_0}, \mathcal{R}_{\pi_\ell})$. To this end, we focus on the pair $\ell = (s_*, h_*)$

[16] In this construction, the MDP\R does not change across the instances, but what changes is the expert's policy. Thus, we parametrize the instances through the policy rather than the MDP\R.

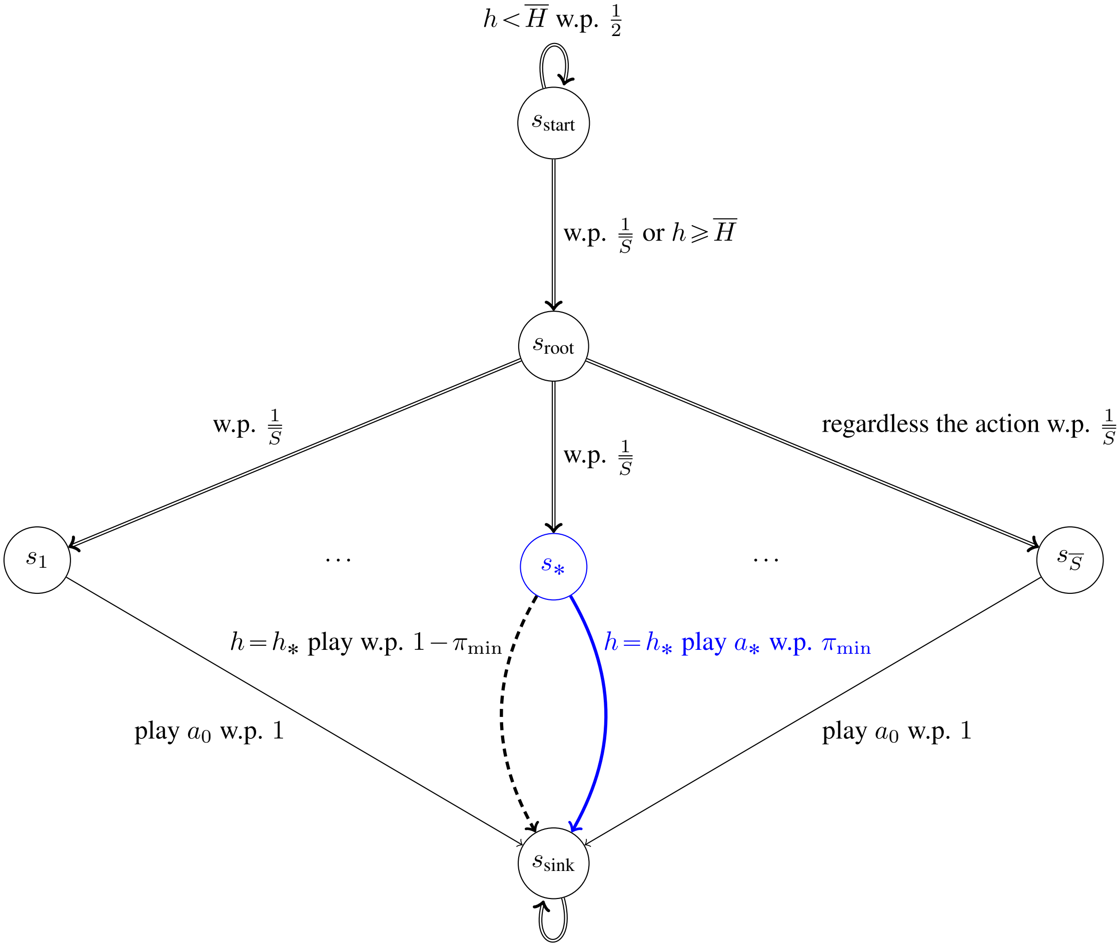

$h < \overline{H}$ w.p. $\frac{1}{2}$
s_{start}
w.p. $\frac{1}{S}$ or $h \geqslant \overline{H}$
s_{root}
w.p. $\frac{1}{S}$
w.p. $\frac{1}{S}$
regardless the action w.p. $\frac{1}{S}$
s_1
$\cdots$
s_*
$\cdots$
$s_{\overline{S}}$
$h = h_*$ play w.p. $1 - \pi_{\min}$
$h = h_*$ play a_* w.p. $\pi_{\min}$
play a_0 w.p. 1
play a_0 w.p. 1
s_{sink}

Figure 6: Semi-formal representation of the the hard instances MDP\R used in the proof of Theorem C.1.

and we enforce the convenience of both actions a_0 and a_* over the other actions. Since both actions are played with non-zero probability by the expert's policy, their value function must be the same. Let us denote with $r^\ell \in \mathcal{R}_{\pi_\ell}$, we must have for all $a_j \notin \{a_0, a_*\}$:

$$\begin{aligned}
& r^\ell_{h_*}(s_*, a_0) + \sum_{l=h_*+1}^{H} r^\ell_l(s_{\text{sink}}) \geqslant r^\ell_{h_*}(s_*, a_j) + \sum_{l=h_*+1}^{H} r^\ell_l(s_{\text{sink}}) \\
& \implies r^\ell_{h_*}(s_*, a_0) \geqslant r^\ell_{h_*}(s_*, a_j), \\
& r^\ell_{h_*}(s_*, a_0) + \sum_{l=h_*+1}^{H} r^\ell_l(s_{\text{sink}}) = r^\ell_{h_*}(s_*, a_*) + \sum_{l=h_*+1}^{H} r^\ell_l(s_{\text{sink}}) \\
& \implies r^\ell_{h_*}(s_*, a_0) = r^\ell_{h_*}(s_*, a_*).
\end{aligned}$$

Consider now the base instance π_0 and denote with $r^0 \in \mathcal{R}_{\pi_0}$. Here we have to enforce the convenience of action a_0 over all the others, including a_*:

$$\begin{aligned}
& r^0_{h_*}(s_*, a_0) + \sum_{l=h_*+1}^{H} r^\ell_l(s_{\text{sink}}) \geqslant r^0_{h_*}(s_*, a_j) + \sum_{l=h_*+1}^{H} r^\ell_l(s_{\text{sink}}) \\
& \implies r^0_{h_*}(s_*, a_0) \geqslant r^0_{h_*}(s_*, a_j),
\end{aligned}$$

$$r^0_{h_*}(s_*,a_0) + \sum_{l=h_*+1}^{H} r^0_l(s_{\text{sink}}) \geqslant r^0_{h_*}(s_*,a_*) + \sum_{l=h_*+1}^{H} r^0_l(s_{\text{sink}})$$
$$\implies r^0_{h_*}(s_*,a_0) \geqslant r^0_{h_*}(s_*,a_*).$$

In order to lower bound the Hausdorff distance, we perform a valid assignment of the rewards for the base instance:

$$r^0_{h_*}(s_*,a_0) = 1,\, r^0_{h_*}(s_*,a_*) = -1,\, r^0_{h_*}(s_*,a_\jmath) = -1.$$

Thus, the Hausdorff distance can be bounded as follows, having renamed, for convenience $x = r^\ell_{h_*}(s_*,a_0)$ and $y = r^\ell_{h_*}(s_*,a_*)$:

$$\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}_{\pi_0},\mathcal{R}_{\pi_\ell}) \geqslant \min_{\substack{x,y\in[-1,1]\\ x=y}} \max\{|x-1|,|y+1|\} = 1.$$

Step 3: Lower bounding Probability Let us consider an (ϵ,δ)-correct algorithm $\mathfrak{A}$ that outputs the estimated feasible set $\widehat{\mathcal{R}}$. Thus, for every $\imath \in \mathcal{J}$, we can lower bound the error probability:

$$\begin{aligned}
\delta &\geqslant \sup_{\text{all } \mathcal{M} \text{ MDP\textbackslash R and expert policies } \pi} \mathop{\mathbb{P}}_{(\mathcal{M},\pi),\mathfrak{A}} \left(\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_\pi,\widehat{\mathcal{R}}\right) \geqslant \frac{1}{2}\right) \\
&\geqslant \sup_{\pi\in\mathbb{M}} \mathop{\mathbb{P}}_{(\mathcal{M},\pi),\mathfrak{A}} \left(\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_\pi,\widehat{\mathcal{R}}\right) \geqslant \frac{1}{2}\right) \\
&\geqslant \max_{\ell\in\{0,\imath\}} \mathop{\mathbb{P}}_{(\mathcal{M},\pi_\ell),\mathfrak{A}} \left(\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\pi_\ell},\widehat{\mathcal{R}}\right) \geqslant \frac{1}{2}\right).
\end{aligned}$$

For every $\imath \in \mathcal{J}$, let us define the *identification function* (whose dependence on the estimated feasible reward set $\widehat{\mathcal{R}}$ is omitted to avoid a too heavy notation):

$$\Psi_\imath := \operatorname*{arg\,min}_{\ell\in\{0,\imath\}} \mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\pi_\ell},\widehat{\mathcal{R}}\right).$$

Let $\jmath \in \{0,\imath\}$. If $\Psi_\imath = \jmath$, then, $\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}_{\pi_{\Psi_\imath}},\mathcal{R}_{\pi_\jmath}) = 0$. Otherwise, if $\Psi_\imath \neq \jmath$, we have:

$$\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\pi_{\Psi_\imath}},\mathcal{R}_{\pi_\jmath}\right) \leqslant \mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\pi_{\Psi_\imath}},\widehat{\mathcal{R}}\right) + \mathcal{H}_{d^{\mathrm{G}}}\left(\widehat{\mathcal{R}},\mathcal{R}_{\pi_\jmath}\right) \leqslant 2\mathcal{H}_{d^{\mathrm{G}}}\left(\widehat{\mathcal{R}},\mathcal{R}_{\pi_\jmath}\right),$$

where the first inequality follows from triangular inequality and the second one from the definition of identification function $\Psi_\imath$. From Equation (11), we have that $\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\pi_{\Psi_\imath}},\mathcal{R}_{\pi_\jmath}\right) \geqslant 1$. Thus, it follows that $\mathcal{H}_{d^{\mathrm{G}}}\left(\widehat{\mathcal{R}},\mathcal{R}_{\pi_\jmath}\right) \geqslant \frac{1}{2}$. This implies the following inclusion of events for $\jmath \in \{0,\imath\}$:

$$\left\{\mathcal{H}_{d^{\mathrm{G}}}\left(\widehat{\mathcal{R}},\mathcal{R}_{\pi_\jmath}\right) \geqslant \frac{1}{2}\right\} \supseteq \{\Psi_\imath \neq \jmath\}.$$

Thus, we can proceed by lower bounding the probability:

$$\begin{aligned}
\max_{\ell\in\{0,\imath\}} \mathop{\mathbb{P}}_{(\mathcal{M}_\ell,\pi),\mathfrak{A}} \left(\mathcal{H}_{d^{\mathrm{G}}}\left(\mathcal{R}_{\pi_\ell},\widehat{\mathcal{R}}\right) \geqslant \frac{1}{2}\right) &\geqslant \max_{\ell\in\{0,\imath\}} \mathop{\mathbb{P}}_{(\mathcal{M}_\ell,\pi),\mathfrak{A}} (\Psi_\imath \neq \ell) \\
&\geqslant \frac{1}{2}\left[\mathop{\mathbb{P}}_{(\mathcal{M}_0,\pi),\mathfrak{A}} (\Psi_\imath \neq 0) + \mathop{\mathbb{P}}_{(\mathcal{M}_\imath,\pi),\mathfrak{A}} (\Psi_\imath \neq \imath) \right] \\
&= \frac{1}{2}\left[\mathop{\mathbb{P}}_{(\mathcal{M}_0,\pi),\mathfrak{A}} (\Psi_\imath \neq 0) + \mathop{\mathbb{P}}_{(\mathcal{M}_\imath,\pi),\mathfrak{A}} (\Psi_\imath = 0) \right],
\end{aligned}$$

where the second inequality follows from the observation that $\max\{a,b\} \geqslant \frac{1}{2}(a+b)$ and the equality from observing that $\Psi_\imath \in \{0,\imath\}$. We can now apply the Bretagnolle-Huber inequality [17, Theorem 14.2] (also reported in Theorem D.1 for completeness) with $\mathbb{P} = \mathbb{P}_{(\mathcal{M}_0,\pi),\mathfrak{A}}$, $\mathbb{Q} = \mathbb{P}_{(\mathcal{M}_0,\pi),\mathfrak{A}}$, and $\mathcal{A} = \{\Psi_\imath \neq 0\}$:

$$\mathop{\mathbb{P}}_{(\mathcal{M}_0,\pi),\mathfrak{A}} (\Psi_\imath \neq 0) + \mathop{\mathbb{P}}_{(\mathcal{M}_\imath,\pi),\mathfrak{A}} (\Psi_\imath = 0) \geqslant \frac{1}{2}\exp\left(-D_{\mathrm{KL}}\left(\mathop{\mathbb{P}}_{(\mathcal{M}_0,\pi),\mathfrak{A}}, \mathop{\mathbb{P}}_{(\mathcal{M}_\imath,\pi),\mathfrak{A}} \right)\right).$$

Step 4: KL-divergence Computation Let $\mathcal{M} \in \mathbb{M}$, we denote with $\mathbb{P}_{(\mathcal{M},\pi),\mathfrak{A}}$ the joint probability distribution of all events realized by the execution of the algorithm in the MDP\R (the presence of p is irrelevant as it does not change across the different instances):

$$\underset{(\mathcal{M},\pi),\mathfrak{A}}{\mathbb{P}} = \prod_{t=1}^{\tau} \rho_t(s_t, a_t, h_t | H_{t-1}) p_{h_t}(s'_t | s_t, a_t) \pi^E_{h_t}(a^E_t | s_t),$$

where ρ_t is the sampling distribution induced by the algorithm $\mathfrak{A}$ and $H_{t-1} = (s_1, a_1, h_1, s'_1, a^E_1, \dots, s_{t-1}, a_{t-1}, h_{t-1}, s'_{t-1}, a^E_{t-1})$ is the history. Let $\imath \in \mathcal{I}$. Let us now move to the KL-divergence between the instances π_0 and $\pi_\imath$ for some $\imath = (s_*, h_*) \in \mathcal{J}$:

$$\begin{aligned} D_{\text{KL}}\left(\mathbb{P}_{(\mathcal{M}_0,\pi),\mathfrak{A}}, \mathbb{P}_{(\mathcal{M}_\imath,\pi),\mathfrak{A}}\right) &= \underset{(\mathcal{M}_0,\pi),\mathfrak{A}}{\mathbb{E}}\left[\sum_{t=1}^{\tau} D_{\text{KL}}\left(\pi^0_{h_t}(\cdot|s_t), \pi^\imath_{h_t}(\cdot|s_t)\right)\right] \\ &\leqslant \underset{(\mathcal{M}_0,\pi),\mathfrak{A}}{\mathbb{E}}\left[N^\tau_{h_*}(s_*)\right] D_{\text{KL}}\left(\pi^0_{h_*}(\cdot|s_*), \pi^\imath_{h_*}(\cdot|s_*)\right) \\ &\leqslant \log\frac{1}{1-\pi_{\min}} \underset{(\mathcal{M}_0,\pi),\mathfrak{A}}{\mathbb{E}}\left[N^\tau_{h_*}(s_*, a_*)\right], \end{aligned}$$

having observed that the transition models differ in $\imath = (s_*, h_*)$ and defined $N^\tau_{h_*}(s_*) = \sum_{t=1}^{\tau} \mathbb{1}\{(s_t, h_t) = (s_*, h_*)\}$ and the last passage is obtained by explicitly computing the KL-divergence:

$$D_{\text{KL}}\left(\pi^0_{h_*}(\cdot|s_*), \pi^\imath_{h_*}(\cdot|s_*)\right) = \sum_{a \in \mathcal{A}} \pi^0_{h_*}(a|s_*) \log\left(\frac{\pi^0_{h_*}(a|s_*)}{\pi^\imath_{h_*}(a|s_*)}\right) = \pi^0_{h_*}(a_0|s_*) \log\left(\frac{\pi^0_{h_*}(a_0|s_*)}{\pi^\imath_{h_*}(a_0|s_*)}\right) = \log\frac{1}{1-\pi_{\min}}.$$

Putting all together, we have:

$$\delta \geqslant \frac{1}{4}\exp\left(-\log\frac{1}{1-\pi_{\min}} \underset{(\mathcal{M}_0,\pi),\mathfrak{A}}{\mathbb{E}}\left[N^\tau_{h_*}(s_*)\right]\right) \implies \underset{(\mathcal{M}_0,\pi),\mathfrak{A}}{\mathbb{E}}\left[N^\tau_{h_*}(s_*)\right] \geqslant \frac{\log\frac{1}{4\delta}}{\log\frac{1}{1-\pi_{\min}}}.$$

Thus, summing over $(s_*, a_*) \in \mathcal{J}$, we have:

$$\begin{aligned} \underset{(\mathcal{M}_0,\pi),\mathfrak{A}}{\mathbb{E}}[\tau] &\geqslant \sum_{(s_*, a_*) \in \mathcal{J}} \underset{(\mathcal{M}_0,\pi),\mathfrak{A}}{\mathbb{E}}\left[N^\tau_{h_*}(s_*, a_*)\right] \\ &= \sum_{(s_*, a_*, h_*) \in \mathcal{I}} \frac{(H - \overline{H} - 2)^2 \log\frac{1}{4\delta}}{2\epsilon^2} \\ &= \overline{SH}\frac{\log\frac{1}{4\delta}}{\log\frac{1}{1-\pi_{\min}}}. \end{aligned}$$

The number of states is given by $S = |\mathcal{S}| = \overline{S} + 3$. Let us first consider the time-homogeneous case, i.e., $\overline{H} = 1$:

$$\underset{(\mathcal{M}_0,\pi),\mathfrak{A}}{\mathbb{E}}[\tau] \geqslant (S-3)\frac{\log\frac{1}{4\delta}}{\log\frac{1}{1-\pi_{\min}}}.$$

For $\delta < 1/16$, $S \geqslant 7$, $A \geqslant 2$, $H \geqslant 2$, we obtain:

$$\underset{(\mathcal{M}_0,\pi),\mathfrak{A}}{\mathbb{E}}[\tau] \geqslant \frac{S}{4\log\frac{1}{1-\pi_{\min}}}\log\frac{1}{\delta}.$$

For the time-inhomogeneous case, instead, we select $\overline{H} = H/2$, to get:

$$\underset{(\mathcal{M}_0,\pi),\mathfrak{A}}{\mathbb{E}}[\tau] \geqslant \frac{(S-3)(H/2)}{\epsilon^2}\frac{\log\frac{1}{4\delta}}{\log\frac{1}{1-\pi_{\min}}}.$$

For $\delta < 1/16$, $S \geqslant 7$, $A \geqslant 2$, $H \geqslant 2$, we obtain:

$$\underset{(\mathcal{M}_0,\pi),\mathfrak{A}}{\mathbb{E}}[\tau] \geqslant \frac{SH}{8\log\frac{1}{1-\pi_{\min}}}\log\frac{1}{\delta}.$$

□

```
Input: significance $\delta \in (0,1)$, $\epsilon$ target accuracy
$t \leftarrow 0$, $\epsilon_0 \leftarrow +\infty$
while $\epsilon_t > \epsilon$ do
    $t \leftarrow t + SAH$
    Collect one sample from each $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]$
    Update $\widehat{p}^t$ and $\widehat{\pi}^{E,t}$ according to (3)
    Update $\epsilon_t = \max_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} \overline{\mathcal{C}}^t_h(s,a)$ (resp. $\widetilde{\mathcal{C}}^t_h(s,a)$)
end while
```

Algorithm 2: UniformSampling-IRL (US-IRL) for time-inhomogeneous (resp. time-homogeneous) transition models and expert's policies.

C.2 Algorithm

In this appendix, we extend US-IRL to the expert's policy estimation under Assumption C.1. The pseudocode is reported in Algorithm 2. The interaction protocol follows the same principles of Algorithm 1, with the only difference that the confidence function, now, must account for the policy estimation, leading to the following function for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]$:[17]

$$\overline{\mathcal{C}}^t_h(s,a) := 2(H-h+1)\left(\mathbb{1}_{\left\{n^t_h(s) \geqslant \max\left\{1, \xi(n^t_h(s), \delta/2)\right\}\right\}} + \sqrt{\frac{2\beta\left(n^t_h(s,a), \delta/2\right)}{n^t_h(s,a)}}\right). \tag{16}$$

where:

$$\xi(n,\delta) := \frac{\log(2SAHn^2/\delta)}{\log(1/(1-\pi_{\min}))}.$$

It is worth noting that we have distributed the confidence δ equally between the problem estimating the policy and that of estimating the transition model. The following theorem provides the sample complexity of US-IRL.

Theorem C.2 (Sample Complexity of US-IRL). *Let $\epsilon > 0$ and $\delta \in (0,1)$, under Assumption C.1, US-IRL is (ϵ,δ)-PAC for d^G-IRL and with probability at least $1-\delta$ it stops after τ samples with:*

- *if the transition model p and the expert's policy π^E are time-inhomogeneous:*
$$\tau \leqslant \frac{8H^3SA}{\epsilon^2}\left(\log\left(\frac{2SAH}{\delta}\right) + (S-1)\overline{C}_1\right) + SH + \frac{SH}{\log(1/(1-\pi_{\min}))}\left(\log\left(\frac{4SAH}{\delta}\right) + \overline{C}_2\right),$$
where $\overline{C}_1 = 1 + \log(1 + (64H^4)/(\epsilon^4(S-1)) \times \left(\log((2SAH)/\delta) + \sqrt{e}(S-1+\sqrt{S-1})\right)^2)$ *and* $\overline{C}_2 = 4\log\left(\frac{\log(4SAH/\delta)+2}{\log(1/(1-\pi_{\min}))}\right)$.
- *if the transition model p and the expert's policy π^E are time-homogeneous:*
$$\tau \leqslant \frac{8H^2SA}{\epsilon^2}\left(\log\left(\frac{2SA}{\delta}\right) + (S-1)\widetilde{C}_1\right) + SH + \frac{S}{\log(1/(1-\pi_{\min}))}\left(\log\left(\frac{4SA}{\delta}\right) + \widetilde{C}_2\right),$$
where $\widetilde{C}_1 = 1 + \log(1 + (64H^4)/(\epsilon^4(S-1)) \times \left(\log((2SA)/\delta) + \sqrt{e}(S-1+\sqrt{S-1})\right)^2)$ *and* $\widetilde{C}_2 = 4\log\left(\frac{\log(4SA/\delta)+2}{\log(1/(1-\pi_{\min}))}\right)$.

[17] As for the transition model, one can adapt the confidence function for the case of stationary policy in straightforward way:

$$\widetilde{\mathcal{C}}^t_h(s,a) := 2(H-h+1)\left(\mathbb{1}_{\left\{n^t_h(s) \geqslant \max\left\{1, \widetilde{\xi}(n^t(s), \delta/2)\right\}\right\}} + \sqrt{\frac{2\widetilde{\beta}\left(n^t(s,a), \delta/2\right)}{n^t(s,a)}}\right), \tag{15}$$

where:

$$\widetilde{\xi}(n,\delta) := \frac{\log(2SAn^2/\delta)}{\log(1/(1-\pi_{\min}))}.$$

In principle, one can also consider the case of a time-homogeneous transition model and time-inhomogeneous expert's policy. We omit it because it adds nothing to the characteristics of the problem and of the algorithms.

Before moving to the proof, let us observe that the result matches the rate of the lower bound of Theorem C.1 up to logarithmic terms.

Proof. We make use of the notation of the proof of Theorem 6.1. We start with the case in which the transition model is time-inhomogeneous. In addition to the good event $\mathcal{E}$ related to the transition model, we introduce the following one:

$$\mathcal{E}_\pi := \left\{ \forall t \in \mathbb{N},\, \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!] : \left| \mathbb{1}_{\pi_h^E(a|s)=0} - \mathbb{1}_{\widehat{\pi}_h^{E,t}(a|s)=0} \right| \leqslant \mathbb{1}_{\left\{n_h^t(s) \geqslant \max\left\{1, \xi(n_h^t(s), \delta/2)\right\}\right\}} \right\},$$

where π_h^E is the true expert's policy and $\widehat{\pi}^{E,t}$ is its estimate via Equation (3) at time t. Thanks to Lemma B.4 and Lemma C.3, we have that $\mathbb{P}(\mathcal{E} \cap \mathcal{E}_\pi) \geqslant 1 - \delta$. Thus, under the good event $\mathcal{E} \cap \mathcal{E}_\pi$, we apply Theorem 3.2 to obtain $\mathcal{H}_{d^{\mathrm{G}}}(\mathcal{R}, \widehat{\mathcal{R}}^\tau) \leqslant \max_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]} \overline{\mathcal{C}}_h^t(s,a)$. A sufficient condition to make this term $\leqslant \epsilon$ is to request the following ones:

$$\max_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]} 2(H-h+1) \mathbb{1}_{\left\{n_h^t(s) \geqslant \max\left\{1, \xi(n_h^t(s), \delta/2)\right\}\right\}} = 0,$$

$$\max_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]} 2\sqrt{2}(H-h+1) \sqrt{\frac{\beta\left(n_h^t(s,a), \delta/2\right)}{n_h^t(s,a)}} \leqslant \epsilon.$$

For the first one, we first enforce the condition:

$$n_h^t(s) \geqslant \xi(n_h^t(s), \delta/2) = \frac{\log(4SAH(n_h^t(s))^2/\delta)}{\log(1/(1-\pi_{\min}))} = \frac{\log(4SAH/\delta)}{\log(1/(1-\pi_{\min}))} + \frac{2 \log n_h^t(s)}{\log(1/(1-\pi_{\min}))}.$$

Using Lemma 15 of [14] and enforcing $n_h^t(s) \geqslant 1$, we obtain:

$$n_h^\tau(s) \leqslant 1 + \frac{1}{\log(1/(1-\pi_{\min}))} \left(\log(4SAH/\delta) + 4 \log\left(\frac{\log(4SAH/\delta) + 2}{\log(1/(1-\pi_{\min}))} \right) \right).$$

Combining this result with that of Theorem 6.1 for what concerns the transition model, we obtain:

$$\begin{aligned} \tau \leqslant \frac{8H^3SA}{\epsilon^2} & \left(\log\left(\frac{2SAH}{\delta}\right) + (S-1) \right. \\ & \left. \times \left(1 + \log\left(1 + \frac{64H^4}{\epsilon^4(S-1)} \left(\log\left(\frac{2SAH}{\delta}\right) + \sqrt{e}(S-1+\sqrt{S-1}) \right)^2 \right) \right) \right) \\ & + SH + \frac{SH}{\log(1/(1-\pi_{\min}))} \left(\log(4SAH/\delta) + 4 \log\left(\frac{\log(4SAH/\delta) + 2}{\log(1/(1-\pi_{\min}))} \right) \right). \end{aligned}$$

Analogous derivations can be carried out for the case of time-homogenous policy using the good event:

$$\widetilde{\mathcal{E}}_\pi := \left\{ \forall t \in \mathbb{N},\, \forall (s,a) \in \mathcal{S} \times \mathcal{A} : \left| \mathbb{1}_{\pi^E(a|s)=0} - \mathbb{1}_{\widehat{\pi}^{E,t}(a|s)=0} \right| \leqslant \mathbb{1}_{\left\{n^t(s) \geqslant \max\left\{1, \widetilde{\xi}(n^t(s), \delta/2)\right\}\right\}} \right\},$$

where $\widetilde{\xi}(n,\delta) := \frac{\log(2SAn^2/\delta)}{\log(1/(1-\pi_{\min}))}$. We omit the tedious but straightforward derivation. □

Lemma C.3. *Under Assumption C.1, the following statements hold:*

- *for* $\xi(n,\delta) := \frac{\log(2SAHn^2/\delta)}{\log(1/(1-\pi_{\min}))}$, *we have that* $\mathbb{P}(\mathcal{E}_\pi) \geqslant 1 - \delta$*;*
- *for* $\widetilde{\xi}(n,\delta) := \frac{\log(2SAn^2/\delta)}{\log(1/(1-\pi_{\min}))}$, *we have that* $\mathbb{P}(\widetilde{\mathcal{E}}_\pi) \geqslant 1 - \delta$*.*

Proof. Let us start with the first statement. We apply first a union bound and, then, Lemma D.5 to perform the concentration:

$$\begin{aligned} \mathbb{P}(\mathcal{E}_\pi^c) &= \mathbb{P}\left(\exists t \in \mathbb{N},\, \exists (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!] : \left| \mathbb{1}_{\pi_h^E(a|s)=0} - \mathbb{1}_{\widehat{\pi}_h^{E,t}(a|s)=0} \right| \leqslant \mathbb{1}_{\left\{n_h^t(s) > \max\left\{1, \xi(n_h^t(s), \delta)\right\}\right\}} \right) \\ &= \mathbb{P}\left(\exists n \in \mathbb{N},\, \exists (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!] : \left| \mathbb{1}_{\pi_h^E(a|s)=0} - \mathbb{1}_{\widehat{\pi}_h^{E,[n]}(a|s)=0} \right| > \mathbb{1}_{\{n \geqslant \max\{1, \xi(n,\delta)\}\}} \right) \end{aligned}$$

$$\leqslant \sum_{h\in[\![H]\!]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{n\geqslant 0}\mathbb{P}\left(\left|\mathbb{1}_{\pi_h^E(a|s)=0}-\mathbb{1}_{\widehat{\pi}_h^{E,[n]}(a|s)=0}\right|\leqslant \mathbb{1}_{\{n>\max\{1,\xi(n,\delta)\}\}}\right)$$

$$\leqslant \sum_{h\in[\![H]\!]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{n\geqslant 1}\mathbb{P}\left(\left|\mathbb{1}_{\pi_h^E(a|s)=0}-\mathbb{1}_{\widehat{\pi}_h^{E,[n]}(a|s)=0}\right|\leqslant \mathbb{1}_{\{n>\max\{1,\xi(n,\delta)\}\}}\right)$$

$$\leqslant \sum_{h\in[\![H]\!]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\delta}{2SAHn^2}=\frac{\pi^2}{6}\frac{\delta}{2}\leqslant \delta,$$

where on the first passage we enforced the condition on the time instants in which the policy estimate changes (i.e., when (s,h) is visited) and we denoted such an estimate as $\widehat{\pi}_h^{E,[n]}$. Then, after a union bound, we apply Lemma D.5. The proof of the second statement is analogous having simply observed that the union bound has to be performed over $\mathcal{S}\times\mathcal{A}$ only. □

D Technical Lemmas

Theorem D.1. *(*Bretagnolle-Huber inequality *[17, Theorem 14.2]) Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on the same measurable space $(\Omega,\mathcal{F})$, and let $\mathcal{A}\in\mathcal{F}$ be an arbitrary event. Then,*

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

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

Theorem D.2. *(*Fano inequality *[9, Proposition 4]) Let $\mathbb{P}_0,\mathbb{P}_1,\ldots,\mathbb{P}_M$ be probability measures on the same measurable space $(\Omega,\mathcal{F})$, and let $\mathcal{A}_1,\ldots,\mathcal{A}_M\in\mathcal{F}$ be a partition of Ω. Then,*

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

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

Lemma D.3. *[13, Proposition 1] Let $\mathbb{P}=(p_1,\ldots,p_D)$ be a categorical probability measure on the support $[\![D]\!]$. Let$\mathbb{P}_n=(\widehat{p}_1,\ldots,\widehat{p}_D)$ be the maximum likelihood estimate of $\mathbb{P}$ obtained with $n\geqslant 1$ independent samples. Then, for every $\delta\in(0,1)$ it holds that:*

$$\mathbb{P}\left(\exists n\geqslant 1: nD_{KL}(\mathbb{P}_n,\mathbb{P})>\log(1/\delta)+(D-1)\log\left(e(1+n/(D-1))\right)\right)\leqslant \delta.$$

Lemma D.4. *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 $\mathbb{Q}=\left(\frac{1}{D},\frac{1}{D},\ldots,\frac{1}{D}\right)$ and $\mathbb{P}=\left(\frac{1+v_1}{D},\frac{1+v_2}{D},\ldots,\frac{1+v_D}{D}\right)$. Then, it holds that:*

$$D_{KL}(\mathbb{P},\mathbb{Q})\leqslant 2\epsilon^2 \qquad \textit{and} \qquad D_{KL}(\mathbb{Q},\mathbb{P})\leqslant 2\epsilon^2.$$

Proof. First of all we recall that since $\sum_{i=1}^D v_i=0$, we have $|\{i\in[\![D]\!]:v_i=\epsilon\}|=|\{i\in[\![D]\!]:v_i=-\epsilon\}|=D/2$. Let us compute the KL-divergence $D_{\text{KL}}(\mathbb{P},\mathbb{Q})$:

$$\begin{aligned}
D_{\text{KL}}(\mathbb{P},\mathbb{Q}) &= \sum_{i=1}^{D}\frac{1+v_i}{D}\log\frac{\frac{1+v_i}{D}}{\frac{1}{D}} \\
&= \sum_{i\in[\![D]\!]:v_i=\epsilon}\frac{1+\epsilon}{D}\log(1+\epsilon)+\sum_{i\in[\![D]\!]:v_i=-\epsilon}\frac{1-\epsilon}{D}\log(1-\epsilon) \\
&= \frac{1+\epsilon}{2}\log(1+\epsilon)+\frac{1-\epsilon}{2}\log(1-\epsilon) \\
&= \underbrace{\frac{1}{2}\log(1-\epsilon^2)}_{\leqslant 0}+\frac{\epsilon}{2}\log(1+\epsilon)-\frac{\epsilon}{2}\log(1-\epsilon) \\
&= \frac{\epsilon}{2}\log\left(1+\frac{2\epsilon}{1-\epsilon}\right)\leqslant \frac{\epsilon^2}{1-\epsilon}\leqslant 2\epsilon^2,
\end{aligned}$$

where we used the inequality $\log(1+x) \leqslant x$ and exploited that $\epsilon \leqslant \frac{1}{2}$. Let us now move to the second KL-divergence $D_{\mathrm{KL}}(\mathbb{Q},\mathbb{P})$:

$$
\begin{aligned}
D_{\mathrm{KL}}(\mathbb{Q},\mathbb{P}) &= \sum_{i=1}^{D} \frac{1}{D} \log \frac{\frac{1}{D}}{\frac{1+v_i}{D}} \\
&= \sum_{i \in [\![D]\!]: v_i = \epsilon} \frac{1}{D} \log \frac{1}{1+\epsilon} + \sum_{i \in [\![D]\!]: v_i = -\epsilon} \frac{1}{D} \log \frac{1}{1-\epsilon} \\
&= -\frac{1}{2} \log(1-\epsilon^2) \\
&\leqslant \frac{1}{2} \left(\frac{1}{1-\epsilon^2} - 1 \right) = \frac{\epsilon^2}{2(1-\epsilon^2)} \leqslant \frac{2}{3} \epsilon^2 \leqslant 2\epsilon^2,
\end{aligned}
$$

where we used the inequality $-\log(1-x) \leqslant \frac{1}{1-x} - 1$ for $0 < x < 1$ and observed that $\epsilon \leqslant \frac{1}{2}$. □

Lemma D.5. *Let $\mathbb{P} = (p_1, \ldots, p_D)$ be a categorical probability measure on the support $[\![D]\!]$. Let $\mathbb{P}_n = (\widehat{p}_1, \ldots, \widehat{p}_D)$ be the maximum likelihood estimate of $\mathbb{P}$ obtained with $n \geqslant 1$ independent samples. Then, if $p_i \in \{0\} \cup [p_{\min}, 1]$ for some $p_{\min} \in (0,1]$. Then, for every $i \in [\![D]\!]$ individually, for every $\delta \in (0,1)$, it holds that:*

$$
\left| \mathbb{1}_{\{p_i = 0\}} - \mathbb{1}_{\{\widehat{p}_i = 0\}} \right| \leqslant \mathbb{1}_{\left\{ n \geqslant \max\left\{ 1, \frac{\log\left(\frac{1}{\delta}\right)}{\log\left(\frac{1}{1-p_{\min}}\right)} \right\} \right\}}.
$$

Proof. Let $i \in [\![D]\!]$ such that $p_i > 0$ and, thus, $\mathbb{1}_{\{p_i=0\}} = 0$. By assumption, it must be that $p_i \geqslant p_{\min}$. To make a mistake, we must have that $\mathbb{1}_{\{\widehat{p}_i = 0\}} = 1$, and, thus, $\widehat{p}_i = 0$. Thus, we compute the probability that no sample i is observed among the n ones:

$$
\mathbb{P}\left(\bigcap_{j \in [\![n]\!]} X_j \neq i \right) = \prod_{j \in [\![n]\!]} \mathbb{P}(X_j \neq i) = \mathbb{P}(X_1 \neq i)^n = (1-p_i)^n \leqslant (1-p_{\min})^n,
$$

where we exploited the fact that the random variables X_j are i.i.d.. If $n = 0$ the latter expression is 1. If, instead, $n \geqslant 1$, by setting the last expression equal to δ, we get:

$$
(1-p_{\min})^n \leqslant \delta \implies n \geqslant \frac{\log\left(\frac{1}{\delta}\right)}{\log\left(\frac{1}{1-p_{\min}}\right)}.
$$

The result follows. □

Lemma D.6. *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 $2^{\frac{D}{5}}$.*

Proof. Let us denote the packing number with $M(\epsilon; \mathcal{V}, d)$ and the covering number with $N(\epsilon; \mathcal{V}, d)$. It is well known that $N(\epsilon; \mathcal{V}, d) \leqslant M(\epsilon; \mathcal{V}, d)$ [10]. Thus, a lower bound to the covering number is a lower bound to the packing number. Let us consider the (pseudo)metric $d'(v,v') = \sum_{j=1}^{D/2} |v_j - v'_j|$ that considers the first half of the components only. Clearly, we have that $d'(v,v') \leqslant d(v,v')$. Therefore, any ϵ-cover w.r.t. $d(v,v')$ is an ϵ-cover w.r.t. $d'(v,v')$ and, consequently, $N(\epsilon; \mathcal{V}, d') \leqslant N(\epsilon; \mathcal{V}, d)$. Since the (pseudo)metric d' considers only the first half of the components, constructing an ϵ-cover of $\mathcal{V}$ w.r.t. d' is equivalent to constructing an ϵ-cover of $\mathcal{V}'$ w.r.t. d', where $\mathcal{V}' = \{-1,1\}^{D/2}$. $\mathcal{V}'$ considers the first half of the components of vectors of $\mathcal{V}$, that can be freely chosen, disregarding the summation constraint.[18] Thus, $N(\epsilon; \mathcal{V}, d') = N(\epsilon; \mathcal{V}', d')$. Notice that d' is now a proper metric on $\mathcal{V}' = \{-1,1\}^{D/2}$. Now, we reduce the problem to constructing cover on the Hamming space $\mathcal{H} = \{0,1\}^{D/2}$. Indeed, we can always map an $(\epsilon/2)$-cover for the Hamming space $\mathcal{H}$ to an ϵ-cover

[18] From an algebraic perspective, $\mathcal{V}'$ can be considered the quotient set obtained from $\mathcal{V}$ by means of the equivalence relation $v \sim v' \iff v_j = v_{j'}$ for all $j \in [\![D/2]\!]$.

for the space $\mathcal{V}'$. Specifically, let $\overline{\mathcal{H}}_{\epsilon/2}$ an $(\epsilon/2)$-cover for the Hamming space, we construct the ϵ-cover of $\mathcal{V}'$, denoted by $\overline{\mathcal{V}}'_{\epsilon}$, by applying the following transformation ($\overline{v}' \in \overline{\mathcal{V}}'_{\epsilon}$):

$$\overline{v}'_j = \begin{cases} -1 & \text{if } \overline{h}_j = 0 \\ 1 & \text{if } \overline{h}_j = 1 \end{cases} \qquad \forall j \in [\![D/2]\!], \quad \forall \overline{h} \in \overline{\mathcal{H}}_{\epsilon/2},$$

or, in more convenient way, $\overline{v}' = 2\overline{h} - 1$. Let $v' \in \mathcal{V}'$:

$$\min_{\overline{v}' \in \overline{\mathcal{V}}'_{\epsilon}} d'(v', \overline{v}') = \min_{\overline{v}' \in \overline{\mathcal{V}}'_{\epsilon}} \sum_{j=1}^{D/2} |v'_j - \overline{v}'_j| = 2 \min_{\overline{h} \in \overline{\mathcal{H}}_{\epsilon/2}} \sum_{j=1}^{D/2} |h_j - \overline{h}_j| \leqslant \epsilon.$$

The covering number of a Hamming space has been lower bounded in [4] for $\epsilon \in [\![D/2]\!]$ as:

$$\log_2 N(\epsilon; \mathcal{H}, d') \geqslant \frac{D}{2} - \log_2 \sum_{k=0}^{\epsilon} \binom{D/2}{k}.$$

We take $\epsilon = D/16$, and we use the known bound $\sum_{i=0}^{k} \binom{n}{i} \leqslant \left(\frac{en}{k}\right)^k$ [31]:

$$\sum_{k=0}^{D/16} \binom{D/2}{k} \leqslant (8e)^{D/16}.$$

From, which, we get:

$$\log_2 N(\epsilon; \mathcal{H}, d') \geqslant \frac{D}{2} - \log_2 \sum_{k=0}^{\epsilon} \binom{D/2}{k} \geqslant \frac{D}{2} - \frac{D}{16} \log_2(8e) \geqslant \frac{D}{5}.$$

□