Efficient Inverse Reinforcement Learning of Transferable Rewards

Giorgia Ramponi [*1] **Alberto Maria Metelli** [*1] **Marcello Restelli** [1]

Abstract

The reward function is widely accepted as a succinct, robust, and transferable representation of a task. Typical approaches, at the basis of Inverse Reinforcement Learning (IRL), leverage on expert demonstrations to recover a reward function. In this paper, we study the theoretical properties of the class of reward functions that are compatible with the expert's behavior. We analyze how the limited knowledge of the expert's policy and the environment affects the reward reconstruction phase. Then, we examine how the error propagates to the learned policy's performance when transferring the reward function to a different environment. We employ these findings to devise a provably efficient active sampling approach, aware of the need for transferring the reward function, that can be paired with a large variety of IRL algorithms.

1. Introduction

Inverse Reinforcement Learning (IRL, Osa et al., 2018) aims at recovering a reward function by observing the behavior of an expert. One of the main challenges of IRL is that the problem itself is ill-posed, as multiple solutions are admissible (Ng & Russell, 2000). Several criteria have been proposed to address this *ambiguity* issue, based on different principles (Abbeel & Ng, 2004; Ratliff et al., 2006a; Ziebart et al., 2008; Metelli et al., 2017; Ho & Ermon, 2016). Taking a step back, in the IRL framework, typically, the transition model of the underlying Markov Decision Process (MDP, Puterman, 2014) is unknown to the algorithm as the expert's policy. In general, these elements are estimated by interacting with the environment and by querying the expert. This leads to an unavoidable error on the *feasible set* of reward functions, i.e., the ones compatible with the expert's demonstrations. Motivated by this, the first question we aim to address is:

(Q1) *How does the error on the transition model and the expert's policy propagate to the recovered reward?*

Clearly, any answer to this question will depend on the chosen IRL algorithm, i.e., on the criterion for selecting *one* reward function within the *feasible set*. To avoid the dependence on the specific IRL algorithm, we will address (Q1) while studying the feasible set properties.

From an applicative point of view, the IRL's objective is twofold: *explainability* and *transferability*. On the one hand, understanding the expert's intentions is useful for descriptive purposes and can help interpret the expert's decisions (Russell & Santos, 2019; Juozapaitis et al., 2019; Hayat et al., 2019). On the other hand, the recovered reward function can be used to learn the same task in another, possibly different, environment (Abbeel & Ng, 2004; Levine et al., 2011; Fu et al., 2017). This ability makes the IRL approach more powerful than Behavioral Cloning (BC, Osa et al., 2018) ones. These considerations motivate our second question:

(Q2) *How does the error on the recovered reward affect the performance of the policy learned in a different environment?*

Contributions In this paper, we study the error on the recovered reward due to the estimation of the transition model and of the expert's policy (Section 3). Then, we consider the reward transferring problem, i.e., when we have a target environment and we can interact only with a different source environment and its expert's policy (Section 4). Then, we study the sample complexity of learning the feasible reward set with a generative model, starting with a simple uniform sampling strategy (Section 5). Finally, we use these findings to construct a new algorithm, *Transferable Reward ActiVE irL* (TRAVEL), that adapts the sampling strategy to the problem. Finally, we derive a problem-dependent upper bound to the sample complexity for the IRL setting (Section 6).

2. Preliminaries

In this section, we introduce the background that will be employed throughout the rest of the paper.

[*] Equal contribution [1] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy. Correspondence to: Alberto Maria Metelli <albertomaria.metelli@polimi.it>.

Workshop on Reinforcement Learning Theory @ ICML 2021.

Markov Decision Processes A discounted Markov Decision Process without Reward function (MDP\R) is defined as a tuple $\mathcal{M}=(\mathcal{S},\mathcal{A},p,\gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $p\in\Delta_{\mathcal{S}\times\mathcal{A}}^{\mathcal{S}}$ is the transition model, and $\gamma\in[0,1)$ is the discount factor. We denote with $\mathcal{M}\cup r$ the MDP obtained by paring $\mathcal{M}$ and r. The agent's behavior is modeled by a policy $\pi\in\Delta_{\mathcal{S}}^{\mathcal{A}}$.

Operators For $f\in\mathbb{R}^{\mathcal{S}}$, we introduce the operator $(Ef)(s,a)=f(s)$. Given $\pi\in\Delta_{\mathcal{S}\times\mathcal{A}}^{\mathcal{S}}$ and $f\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ we denote with $(B^{\pi}f)(s,a)=f(s,a)\mathbb{1}\{\pi(a|s)>0\}$ and $(\overline{B}^{\pi}f)(s,a)=f(s,a)\mathbb{1}\{\pi(a|s)=0\}$. Finally, we denote the expectation under the discounted occupancy measure with $(I_{\mathcal{S}\times\mathcal{A}}-\gamma P\pi)^{-1}f=\sum_{t\in\mathbb{N}}(\gamma P\pi)^t f$. See Appendix A for a complete definition of the operators.

Value Functions and Optimality The *Q-function* $Q^{\pi}_{\mathcal{M}\cup r}\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ of a policy π in MDP $\mathcal{M}\cup r$ is the expected discounted sum of the rewards starting from a state-action pair and playing policy π thereafter, and defined via the Bellman equation (Sutton et al., 1998): $Q^{\pi}_{\mathcal{M}\cup r}=r+\gamma P\pi Q^{\pi}_{\mathcal{M}\cup r}$. The *V-function* is the expectation of the Q-function over the action space: $V^{\pi}_{\mathcal{M}\cup r}=\pi Q^{\pi}_{\mathcal{M}\cup r}$. The *advantage function* $A^{\pi}_{\mathcal{M}\cup r}=Q^{\pi}_{\mathcal{M}\cup r}-EV^{\pi}_{\mathcal{M}\cup r}$ provides the one-step performance gain achieved by playing an action in a state rather than following policy π. A policy $\pi^*\in\Delta_{\mathcal{S}}^{\mathcal{A}}$ is optimal if it yields non-positive advantage, i.e., $A^{\pi^*}_{\mathcal{M}\cup r}(s,a)\leqslant 0$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. We denote with $\Pi^*_{\mathcal{M}\cup r}\subseteq\Delta_{\mathcal{S}}^{\mathcal{A}}$ the set of optimal policies for the MDP $\mathcal{M}\cup r$.

3. Recovering Feasible Rewards

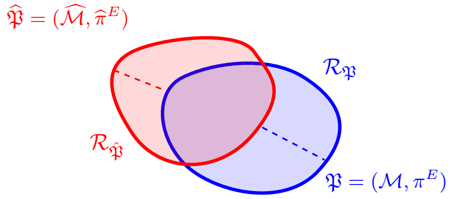
The IRL problem admits multiple solutions, i.e., it suffers from an *ambiguity* issue. Thus, an IRL algorithm $\mathscr{A}$, implementing a criterion for selecting a reward function within this set, can be seen as a *choice function* mapping an IRL problem $\mathfrak{P}$ to a feasible reward, i.e., $\mathscr{A}:\mathfrak{P}\mapsto r\in\mathcal{R}_{\mathfrak{P}}$. We primarily focus on the properties of the feasible set that can be *implicitly* characterized in the following way.

Lemma 3.1 (Feasible Reward Set Implicit (Ng & Russell, 2000)). *Let $\mathfrak{P}=(\mathcal{M},\pi^E)$ be an IRL problem. Let $r\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, then r is a feasible reward, i.e., $r\in\mathcal{R}_{\mathfrak{P}}$ if and only if for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ it holds that:*

$$\begin{aligned} &(i) \quad Q^{\pi^E}_{\mathcal{M}\cup r}(s,a)-V^{\pi^E}_{\mathcal{M}\cup r}(s)=0 \qquad \text{if } \pi^E(a|s)>0,\\ &(ii) \quad Q^{\pi^E}_{\mathcal{M}\cup r}(s,a)-V^{\pi^E}_{\mathcal{M}\cup r}(s)\leqslant 0 \qquad \text{if } \pi^E(a|s)=0. \end{aligned}$$

Furthermore, if condition (ii) holds with the strict inequality, π^E is the unique optimal policy under r, i.e., $\Pi^*_{\mathcal{M}\cup r}=\{\pi^E\}$.

Both conditions are expressed in terms of the advantage function $A^{\pi^E}_{\mathcal{M}\cup r}(s,a)=Q^{\pi^E}_{\mathcal{M}\cup r}(s,a)-V^{\pi^E}_{\mathcal{M}\cup r}(s)$. Specifically, condition (i) prescribes that the advantage function of the actions that are played by the expert's policy π^E must be null, whereas condition (ii) ensures that the actions that are not played have a non-positive advantage. In other words, Lemma 3.1 requires π^E to be an optimal policy under reward function r. From Lemma 3.1, we derive an *explicit* form of the reward functions belonging to the feasible set.



$\widehat{\mathfrak{P}} = (\widehat{\mathcal{M}}, \widehat{\pi}^E)$
$\mathcal{R}_{\mathfrak{P}}$
$\mathcal{R}_{\widehat{\mathfrak{P}}}$
$\mathfrak{P} = (\mathcal{M}, \pi^E)$

Figure 1. Feasible reward sets of two IRL problems: $\mathfrak{P}$ is an IRL problem and $\widehat{\mathfrak{P}}$ a version of $\mathfrak{P}$ estimated from samples.

Lemma 3.2 (Feasible Reward Set Explicit). *Let $\mathfrak{P}=(\mathcal{M},\pi^E)$ be an IRL problem. Let $r\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, then r is a feasible reward, i.e., $r\in\mathcal{R}_{\mathfrak{P}}$ if and only if there exist $\zeta\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\geqslant 0}$ and $V\in\mathbb{R}^{\mathcal{S}}$ such that:*

$$r=-\overline{B}^{\pi^E}\zeta+(E-\gamma P)V.$$

Thus, the reward function is the sum of two terms. The first term $-\overline{B}^{\pi^E}\zeta$ depends on the expert's policy π^E only but not on the MDP. It is zero for all actions the expert plays, i.e., those such that $\pi^E(a|s)>0$, while its value is non-positive for actions that the expert does not play, i.e., for those with $\pi^E(a|s)=0$. Requiring a strictly positive ζ allows enforcing π^E as the unique optimal policy. The second term $(E-\gamma P)V$ depends on the MDP but not on the expert's policy. It can be interpreted as a *reward-shaping* via function V, which is well-known to preserve the optimality of the expert's policy (Ng & Russell, 2000).

Error Propagation in the Feasible Reward Set We now study the *error propagation* in the feasible set, addressing question (Q1) of Section 1. Specifically, we consider two IRL problems $\mathfrak{P}=(\mathcal{M},\pi^E)$ and $\widehat{\mathfrak{P}}=(\widehat{\mathcal{M}},\widehat{\pi}^E)$. $\widehat{\mathfrak{P}}$ can be thought of as an approximate version of $\mathfrak{P}$, where the transition model $\widehat{p}$ and the expert's policy $\widehat{\pi}^E$ are estimated through samples. Intuitively, an error in estimating the transition model p and the expert's policy π^E results in an error in the estimation of the feasible sets $\mathcal{R}_{\mathfrak{P}}$. Since the IRL problem is ambiguous, we will be satisfied whenever we recover an accurate approximation of the feasible set. Informally, we will say that the estimated feasible set $\mathcal{R}_{\widehat{\mathfrak{P}}}$ is "close" to the exact one $\mathcal{R}_{\mathfrak{P}}$ if for *every* reward $r\in\mathcal{R}_{\mathfrak{P}}$ there exists *one* estimated reward $\widehat{r}\in\mathcal{R}_{\widehat{\mathfrak{P}}}$ that is "close" to r and vice versa (Figure 1).[1]

Theorem 3.1 (Error Propagation). *Let $\mathfrak{P}=(\mathcal{M},\pi^E)$ and $\widehat{\mathfrak{P}}=(\widehat{\mathcal{M}},\widehat{\pi}^E)$ be two IRL problems. Then, for any $r\in\mathcal{R}_{\mathfrak{P}}$ such that $r=-\overline{B}^{\pi^E}\zeta+(E-\gamma P)V$ and $\|r\|_{\infty}\leqslant R_{\max}$ there*

[1] This notion of "closeness" between two sets is formalized by the *Hausdorff distance* (Rockafellar & Wets, 2009).

exists $\widehat{r} \in \mathcal{R}_{\widehat{\mathfrak{P}}}$ such that element-wise it holds that:

$$|r-\widehat{r}| \leqslant \overline{B}^{\pi^E} B^{\widehat{\pi}^E}\zeta + \gamma\left|\left(P-\widehat{P}\right)V\right|.$$

Furthermore, $\|\zeta\|_\infty \leqslant \frac{R_{\max}}{1-\gamma}$ and $\|V\|_\infty \leqslant \frac{R_{\max}}{1-\gamma}$.

The result states the *existence* of a reward $\widehat{r}$ in the estimated feasible set $\mathcal{R}_{\widehat{\mathfrak{P}}}$ fulfilling the bound that consists of two components. The first one $\overline{B}^{\pi^E} B^{\widehat{\pi}^E}\zeta$ depends on the policy approximation only. Specifically, this term is non-zero in the state-action pairs such that $\pi^E(a|s) = 0$ and $\widehat{\pi}^E(a|s) > 0$ only, i.e., for the actions that are not played by the expert but are wrongly believed to be played. Thus, to zero out this term it suffices to identify for each state *one* action played by the expert. The second term $|(P-\widehat{P})V|$, instead, concerns the estimation error of the transition model. Clearly, by reversing r with $\widehat{r}$, we obtain a symmetric statement.

4. Transferring Rewards

One of the advantages of IRL over BC is the possibility of reusing the learned reward function in a different environment. Specifically, there is an expert agent playing an optimal policy π^E in a *source* MDP\R $\mathcal{M}$. We want to recover a reward function explaining the expert's policy π^E in $\mathcal{M}$, knowing that we will employ it in a different *target* MDP\R $\mathcal{M}'$ for policy learning.

4.1. Transferable Reward Assumption

It might happen that different rewards are inducing the same expert's policy π^E in the source MDP\R $\mathcal{M}$, while generating different optimal policies in the target MDP\R $\mathcal{M}'$. More formally, let r^E be the true (and unknown) reward optimized by the expert's policy π^E and let $(\pi')^E$ the policy that the expert would play in $\mathcal{M}'$ optimizing the same r^E. Suppose we are able to solve the source IRL problem $\mathfrak{P} = (\mathcal{M}, \pi^E)$ finding $r \in \mathcal{R}_{\mathfrak{P}}$, possibly different from r^E. There is no guarantee that r will make the policy $(\pi')^E$ optimal in the target MDP\R $\mathcal{M}'$. In other words, r might not be a solution to the target IRL problem $\mathfrak{P}' = (\mathcal{M}', (\pi')^E)$. In order to solve this additional ambiguity issue, we enforce the following assumption.

Assumption 4.1. *Let $\mathfrak{P} = (\mathcal{M}, \pi^E)$ and $\mathfrak{P}' = (\mathcal{M}', (\pi')^E)$ be the source and target IRL problems. The corresponding feasible sets satisfy $\mathcal{R}_{\mathfrak{P}'} \supseteq \mathcal{R}_{\mathfrak{P}}$.*

With this assumption, we guarantee that every reward that is feasible for the source MDP\R $\mathcal{M}$ is also feasible for the target MDP\R $\mathcal{M}'$. We think that this assumption is unavoidable in our setting since we have no information regarding the *optimality* of the observed expert policy π^E in the target MDP\R $\mathcal{M}'$. The intuition behind Assumption 4.1 is that, by simply observing the expert playing an action in a state, we can only conclude that the action is optimal, but we are unable to judge "how much" suboptimal are the actions that the agent does not play.

4.2. Error Propagation on the Value Function

In this section, we focus on question (Q2) presented in Section 1. We discuss, under Assumption 4.1, how an error on the reward function propagates into an error in estimating the optimal value function, when transferring the recovered reward to a possibly different MDP\R $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, p', \gamma')$. We present Lemma 4.1, which provides upper and lower bounds to the difference between the optimal Q-function under the true reward function $Q^*_{\mathcal{M}' \cup r}$ and the optimal Q-function under the estimated reward function $Q^*_{\mathcal{M}' \cup \widehat{r}}$.

Lemma 4.1 (Simulation Lemma 1)**.** *Let $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, p', \gamma')$ be an MDP\R, let $r, \widehat{r} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ be two reward functions. Then, for every $\pi^* \in \Pi^*_{\mathcal{M}'\cup r}$ and $\widehat{\pi}^* \in \Pi^*_{\mathcal{M}'\cup\widehat{r}}$ optimal policies for the MDPs $\mathcal{M}' \cup r$ and $\mathcal{M}' \cup \widehat{r}$ respectively, the following inequalities hold element-wise:*

$$Q^*_{\mathcal{M}'\cup r} - Q^*_{\mathcal{M}'\cup\widehat{r}} \leqslant \left(I_{\mathcal{S}\times\mathcal{A}} - \gamma' P' \pi^*\right)^{-1}(r-\widehat{r}),$$

$$Q^*_{\mathcal{M}'\cup r} - Q^*_{\mathcal{M}'\cup\widehat{r}} \geqslant \left(I_{\mathcal{S}\times\mathcal{A}} - \gamma' P' \widehat{\pi}^*\right)^{-1}(r-\widehat{r}).$$

In particular, it holds that:

$$\left\|Q^*_{\mathcal{M}'\cup r} - Q^*_{\mathcal{M}'\cup\widehat{r}}\right\|_\infty \leqslant \max_{\pi\in\{\pi^*,\widehat{\pi}^*\}} \left\|\left(I_{\mathcal{S}\times\mathcal{A}} - \gamma' P' \pi\right)^{-1}(r-\widehat{r})\right\|_\infty.$$

The result suggests that we need to be accurate in estimating the reward function of the state-action pairs that are highly visited by the discounted occupancy measures of the optimal policy $\pi^* \in \Pi^*_{\mathcal{M}'\cup r}$ and of the policy $\widehat{\pi}^* \in \Pi^*_{\mathcal{M}'\cup\widehat{r}}$ induced by the estimated reward function $\widehat{r}$.

5. Learning Transferable Rewards with a Generative Model

We study the problem of learning a transferable reward function with a generative model. Specifically, we consider the following setting: (i) $\mathcal{M}$ and $\mathcal{M}'$ have the same state and action spaces; (ii) we have access to the generative model of $\mathcal{M}$; (iii) we can query the expert's policy π^E in any state of $\mathcal{M}$; (iv) the expert's policy π^E is deterministic; (v) we know the transition model p' and the discount factor γ' of $\mathcal{M}'$. At each iteration $k \in [K]$, we collect at most $n_{\max} \in \mathbb{N}$ samples. When the generative model is queried about a state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, it responds with a transition triple (s,a,s'), where $s' \sim p(\cdot|s,a)$, and with an expert decision $\pi^E(s)$. The sampling strategy $\mathscr{S}$ decides, at each iteration k, how to allocate the $n_{\max}$ samples over the state-action space $\mathcal{S}\times\mathcal{A}$, with the goal of estimating the feasible set accurately. To this purpose, we introduce the following PAC requirement.

Definition 5.1. *Let $\mathscr{S}$ be a sampling strategy. Let $\mathcal{R}_{\mathfrak{P}}$ be the exact feasible set and $\mathcal{R}_{\widehat{\mathfrak{P}}}$ be the feasible set recov-*

Algorithm 1 Uniform Sampling IRL

```
**Input:** significance $\delta \in (0,1)$, $\epsilon$ target accuracy, $n_{\max}$ maximum number of samples per iteration
$k \leftarrow 0$, $\epsilon_0 = \frac{1}{1-\gamma'}$
**while** $\epsilon_k > \epsilon$ **do**
    Collect $\lceil \frac{n_{\max}}{SA} \rceil$ from each $(s,a) \in \mathcal{S} \times \mathcal{A}$
    Update accuracy $\epsilon_{k+1} = \frac{1}{1-\gamma'} \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathcal{C}_{k+1}(s,a)$
    Update $\widehat{p}_{k+1}$ and $\widehat{\pi}^E_{k+1}$, $k \leftarrow k+1$
**end while**
```

ered after observing $n \geqslant 0$ samples collected in the source MDP\R $\mathcal{M}$. Let $(\overline{r}, \check{r}) \in \mathcal{R}_{\mathfrak{P}} \times \mathcal{R}_{\widehat{\mathfrak{P}}}$ be a pair of target rewards, we say that $\mathscr{S}$ is (ϵ, δ, n)-correct for MDP\R $\mathcal{M}'$ and for the target rewards $(\overline{r}, \check{r})$ if with probability at least $1-\delta$ it holds that:

$$\begin{aligned} \inf_{\widehat{r} \in \mathcal{R}_{\widehat{\mathfrak{P}}}} \left\| Q^*_{\mathcal{M}' \cup \overline{r}} - Q^*_{\mathcal{M}' \cup \widehat{r}} \right\|_\infty \leqslant \epsilon \\ \inf_{r \in \mathcal{R}_{\mathfrak{P}}} \left\| Q^*_{\mathcal{M}' \cup \check{r}} - Q^*_{\mathcal{M}' \cup r} \right\|_\infty \leqslant \epsilon. \end{aligned} \tag{1}$$

5.1. Transition Model and Policy Estimation

For each iteration $k \in [K]$, we denote by $n_k(s,a,s')$ the number of times the triple $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ is visited in episode k and $n_k(s,a) = \sum_{s' \in \mathcal{S}} n_k(s,a,s')$. For the transition model estimation, we define the cumulative counts $N_k(s,a,s') = \sum_{j \in [k]} n_j(s,a,s')$ and $N_k(s,a) = \sum_{j \in [k]} n_j(s,a)$, which lead to the estimate: $\widehat{p}_k(s'|s,a) = N_k(s,a,s')/N_k^+(s,a)$, where $x^+ = \max\{1, x\}$. Concerning the estimated expert's policy $\widehat{\pi}^E_k$, since the expert is deterministic, the first time we sample a state $s \in \mathcal{S}$ we recover the true policy $\pi^E(s)$. Whenever a state is visited, the expert reveals an action. Thanks to the error propagation (Theorem 3.1) and the simulation lemma (Lemma 4.1), at iteration $k+1$, given a target reward $\overline{r} \in \mathcal{R}_{\mathfrak{P}}$, there exists an estimated reward $\widehat{r}_{k+1} \in \mathcal{R}_{\widehat{\mathfrak{P}}_{k+1}}$ such that $|\overline{r} - \widehat{r}_{k+1}|(s,a) \leqslant \mathcal{C}_{k+1}(s,a)$ where, $\mathcal{C}_{k+1}(s,a)$ is a suitably defined confidence interval accounting for the uncertainty in the transition model and expert's policy estimation (see Lemma B.4). Moreover, given a target reward $\check{r}_{k+1} \in \mathcal{R}_{\widehat{\mathfrak{P}}_{k+1}}$, we can guarantee that under the same good event $\mathcal{E}$, there exists an exact reward function $r \in \mathcal{R}_{\mathfrak{P}}$ such that $|r - \check{r}_{k+1}|(s,a) \leqslant \mathcal{C}_{k+1}(s,a)$ as well.

5.2. Uniform Sampling Strategy

The first algorithm we present employs a *uniform sampling* strategy to allocate samples over $\mathcal{S} \times \mathcal{A}$, until the desired accuracy $\epsilon > 0$ is reached (Algorithm 1). The stopping condition makes use of the obtained confidence intervals:

$$\epsilon_{k+1} := \frac{1}{1-\gamma'} \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathcal{C}_{k+1}(s,a) \leqslant \epsilon.$$

Thanks to Lemma 4.1, we are guaranteed that, under $\mathcal{E}$, when the stopping condition is activated, the recovered feasible set fulfills Definition 5.1, as shown below.

Algorithm 2 TRAVEL

```
**Input:** significance $\delta \in (0,1)$, $\epsilon$ target accuracy, $n_{\max}$ maximum number of samples per iteration, IRL algorithm $\mathscr{A}$
$k \leftarrow 0$, $\epsilon_0 = \frac{1}{1-\gamma'}$
**while** $\epsilon_k > \epsilon$ **do**
    Solve optimization problem in Eq (2) for $n_{k+1}$ and $\epsilon_{k+1}$
    Collect $n_{k+1}(s,a)$ samples from $(s,a) \in \mathcal{S} \times \mathcal{A}$
    Update $\widehat{p}_{k+1}$ and $\widehat{\pi}^E_{k+1}$, $k \leftarrow k+1$
**end while**
```

Theorem 5.1 (Sample Complexity of Uniform Sampling IRL). *If Algorithm 1 stops at iteration K with accuracy ϵ_K, then with probability at least $1-\delta$ it fulfills Definition 5.1, for arbitrary target reward functions $\overline{r}$ and $\check{r}$, with a number of samples upper bounded by:*

$$n \leqslant \widetilde{\mathcal{O}} \left(\frac{\gamma^2 R^2_{\max} SA}{(1-\gamma')^2 (1-\gamma)^2 \epsilon_K^2} \right).$$

6. Active Learning of Transferable Rewards

In this section, we present a novel algorithm, named *Transferable Reward ActiVE irL* (TRAVEL), that adapts the sampling strategy to the structure of the problem. In order to choose which state-action pairs to sample from, we make use of Lemma 4.1. Suppose we are at iteration $k \in [K]$ and we have to decide how to allocate the $n_{\max}$ samples of the next iteration $k+1$. We have already observed that, under the good event $\mathcal{E}$, there exists $\widehat{r}_{k+1} \in \mathcal{R}_{\widehat{\mathfrak{P}}_{k+1}}$ and $r \in \mathcal{R}_{\mathfrak{P}}$ such that $|\overline{r} - \widehat{r}_{k+1}|(s,a) \leqslant \mathcal{C}_{k+1}(s,a)$ and $|\check{r}_{k+1} - r|(s,a) \leqslant \mathcal{C}_{k+1}(s,a)$. Then, by Lemma 4.1 we have that:

$$\left\| Q^*_{\mathcal{M}' \cup \overline{r}} - Q^*_{\mathcal{M}' \cup \widehat{r}_{k+1}} \right\|_\infty \leqslant \max_{\pi \in \{\overline{\pi}^*, \widehat{\pi}^*_{k+1}\}} \left\| \left(I_{\mathcal{S} \times \mathcal{A}} - \gamma' P' \pi \right)^{-1} \mathcal{C}_{k+1} \right\|_\infty,$$

$$\left\| Q^*_{\mathcal{M}' \cup \check{r}_{k+1}} - Q^*_{\mathcal{M}' \cup r} \right\|_\infty \leqslant \max_{\pi \in \{\check{\pi}^*_{k+1}, \pi^*\}} \left\| \left(I_{\mathcal{S} \times \mathcal{A}} - \gamma' P' \pi \right)^{-1} \mathcal{C}_{k+1} \right\|_\infty,$$

where the policies are *arbitrarily* selected in the corresponding sets: $\overline{\pi}^* \in \Pi^*_{\mathcal{M}' \cup \overline{r}}$, $\widehat{\pi}^*_{k+1} \in \Pi^*_{\mathcal{M}' \cup \widehat{r}_{k+1}}$, $\check{\pi}^*_{k+1} \in \Pi^*_{\mathcal{M}' \cup \check{r}_{k+1}}$, and $\pi^* \in \Pi^*_{\mathcal{M}' \cup r}$. In principle, we could optimize the right-hand side of the previous inequalities over n_{k+1} to obtain the sample allocation. However, we have no knowledge about all the involved policies. Thus, we resort to a surrogate bound that leads to an allocation better than the uniform one. To this purpose, given an IRL algorithm $\mathscr{A}$, we follow the spirit of (Zanette et al., 2019) extending the maximization over a set of policies $\Pi^{\mathscr{A}}_k$ that, with high probability, contains the needed ones:

$$\Pi^{\mathscr{A}}_k = \left\{ \pi \in \Delta^{\mathcal{A}}_{\mathcal{S}} : \sup_{\mu_0 \in \Delta^{\mathcal{S}}} \mu_0^\intercal \left(V^*_{\mathcal{M}' \cup \mathscr{A}(\mathcal{R}_{\widehat{\mathfrak{P}}_k})} - V^\pi_{\mathcal{M}' \cup \mathscr{A}(\mathcal{R}_{\widehat{\mathfrak{P}}_k})} \right) \leqslant 4\epsilon_k \right\},$$

where the value of ϵ_k will be defined later. Here is the first point in which we actually make use of an IRL algorithm $\mathscr{A}$, whose goal is to choose a reward in the feasible reward set. The rationale in the definition of $\Pi^{\mathscr{A}}_k$ is to constrain the search for the policy to those yielding a value function at iteration k close to the estimated optimal one. We can now formulate the optimization problem:

$$\begin{aligned}\epsilon_{k+1} := \min_{n_{k+1} \in \mathbb{N}^{\mathcal{S}\times\mathcal{A}}} \max_{\substack{\mu_0 \in \Delta^{\mathcal{S}\times\mathcal{A}} \\ \pi \in \Delta_{\mathcal{S}}^{\mathcal{A}}}} & {\mu_0}^\intercal \left(I_{\mathcal{S}\times\mathcal{A}} - \gamma' P' \pi\right)^{-1} \mathcal{C}_{k+1} \\ \text{s.t. } & {\mu_0}^\intercal E\left(V^*_{\mathcal{M}' \cup \mathscr{A}(\mathcal{R}_{\widehat{\mathfrak{P}}_k})} - V^\pi_{\mathcal{M}' \cup \mathscr{A}(\mathcal{R}_{\widehat{\mathfrak{P}}_k})}\right) \leqslant 4\epsilon_k \\ & \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} n_{k+1}(s,a) \leqslant n_{\max}.\end{aligned} \tag{2}$$

The program is a minimax in which we look for the sample allocation n_{k+1} that minimizes the bound on the value function difference of Lemma 4.1, under the worst possible policy π in the set $\Pi_k^{\mathscr{A}}$ and initial state-action distribution μ_0. Therefore ϵ_k, used to define $\Pi_k^{\mathscr{A}}$, is the objective function value of the previous iteration k. It can be proved that under the good event $\mathcal{E}$, $\Pi_k^{\mathscr{A}}$ contains a specimen of all the required optimal policies, i.e., $\overline{\pi}^*$, $\widehat{\overline{\pi}}^*_{k+1}$, $\widecheck{\pi}^*_{k+1}$, and π^* (Corollary B.2). The constant $n_{\max}$ is the maximum number of samples allowed per iteration and it is a user-defined parameter. By choosing the $n_{\max}$ value, the user has to make a trade-off between time and sample efficiency. If the value of $n_{\max}$ is too high, the algorithm achieves the desired ϵ-correctness very quickly but with a possible sample inefficient behavior (close to uniform); if the value of $n_{\max}$ is too low, many iterations are needed to achieve the desired accuracy ϵ, but choosing more carefully where to sample. The pseudocode of TRAVEL is reported in Algorithm 2.

It is worth noting that we have not specified which IRL algorithm should be employed to recover a reward function. Indeed, any IRL algorithm $\mathscr{A}$ can be used for this purpose, provided that it selects a reward function within the feasible set $\mathcal{R}_{\widehat{\mathfrak{P}}}$. We stress that the main goal of this paper is not to provide a new IRL algorithm for choosing a good reward from the feasible reward set, but to explain how to recover a good approximation of this feasible set.

Sample Complexity In this section, we prove that TRAVEL fulfills the PAC-condition of Definition 5.1. In order to provide this result, we use as suboptimality gaps the negative advantage: $-A^*_{\mathcal{M}'\cup\widetilde{r}}(s,a) = V^*_{\mathcal{M}'\cup\widetilde{r}}(s) - Q^*_{\mathcal{M}'\cup\widetilde{r}}(s,a)$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, where $\widetilde{r} \in \operatorname{arginf}_{r\in\mathcal{R}_{\mathfrak{P}}} \|r - \mathscr{A}(\mathcal{R}_{\widehat{\mathfrak{P}}_K})\|_\infty$ is the reward function in the exact feasible set $\mathcal{R}_{\mathfrak{P}}$ closest to the one returned by the IRL algorithm $\mathscr{A}$ applied to the estimated feasible set $\mathcal{R}_{\widehat{\mathfrak{P}}_K}$.

Theorem 6.1 (Sample Complexity of TRAVEL)**.** *If Algorithm 2 stops at iteration K with accuracy ϵ_K and accuracy ϵ_{K-1} at the previous iteration, then with probability at least $1-\delta$ it fulfills Definition 5.1, for arbitrary target reward functions $\overline{r}$ and $\check{r}$, with a number of samples upper bounded by $n = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} N_K(s,a)$ where:*

$$N_K(s,a) \leqslant \widetilde{\mathcal{O}}\left(\min\left\{\frac{\gamma^2 R_{\max}^2}{(1-\gamma')^2(1-\gamma)^2\epsilon_K^2}, \frac{\gamma^2 R_{\max}^2 \epsilon_{K-1}^2}{(1-\gamma)^2(-A(s,a))^2\epsilon_K^2}\right\}\right).$$

The significance of this result depends on two main components: the ratio between the two objectives ϵ_{K-1} and ϵ_K and the suboptimality gaps. The latter depends, although indirectly, on the employed IRL algorithm $\mathscr{A}$. The more suboptimal the action is, the less the action will be sampled. Instead, the $\epsilon_{K-1}/\epsilon_K$ component depends on the choice of the $n_{\max}$ value: if this value is small, then the ratio will also be small. As discussed in the previous section, if the $n_{\max}$ value is too high, the algorithm tends to sample every action-state pair uniformly.

In the appendix we provide the experimental evaluation. We show the need for employing IRL over BC when our goal is to transfer knowledge to a target environment. Then, we highlight the benefits of the sampling strategy of TRAVEL over Uniform Sampling. Finally, we show how TRAVEL can be combined with different IRL algorithms.

7. Conclusions

In this paper, we have studied how to efficiently learn a transferable reward from a theoretical perspective. Using the concept of feasible reward set, introduced by Ng & Russell (2000), we have derived novel bounds on the error of the reward function, given an error on the transition model and the expert's policy. We have then obtained similar results on the performance in a target environment using the rewards recovered from a source environment, introducing new simulation lemmas. Based on these findings, we have proposed two algorithms, Uniform Sampling IRL and TRAVEL, which, given a generator model for the source MDP, decide the sampling strategy for querying the generator. These algorithms use an IRL algorithm, decided by the user, as *choice function*. We have derived from the Uniform Sampling IRL a sample complexity bound which, to the best of our knowledge, is the first sample complexity result for the IRL setting. TRAVEL, instead, adapts the sampling strategy to the specific environment at hand. Leveraging this characteristic of the algorithm, we have obtained a problem-dependent bound on its sample complexity. Despite the limitations of the considered setting, we believe that this paper makes a first step towards a better understanding of the theoretical aspects of IRL. Many appealing future research directions arise. One central theoretical question is:

(Q3) *Is Inverse Reinforcement Learning intrinsically more difficult than Reinforcement Learning? Is the sample complexity $\widetilde{\mathcal{O}}\left(\frac{SA}{(1-\gamma')^2(1-\gamma)^2\epsilon^2}\right)$ tight?*

We are currently unable to answer. From an algorithmic perspective, our setting limits to tabular MDPs and assumes access to a generative model. Future investigations should include the extension to episode-based interaction and the introduction of function approximation techniques to cope with continuous problems.

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In Brodley, C. E. (ed.), *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, volume 69 of *ACM International Conference Proceeding Series*. ACM, 2004. doi: 10.1145/1015330.1015430.

Boularias, A., Kober, J., and Peters, J. Relative entropy inverse reinforcement learning. In Gordon, G., Dunson, D., and Dudík, M. (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 182–189, Fort Lauderdale, FL, USA, 11–13 Apr 2011. JMLR Workshop and Conference Proceedings.

Chatzigeorgiou, I. Bounds on the lambert function and their application to the outage analysis of user cooperation. *IEEE Communications Letters*, 17(8):1505–1508, 2013.

Fu, J., Luo, K., and Levine, S. Learning robust rewards with adversarial inverse reinforcement learning. *CoRR*, abs/1710.11248, 2017.

Hayat, A., Singh, U., and Namboodiri, V. P. Inforl: Interpretable reinforcement learning using information maximization. *CoRR*, abs/1905.10404, 2019.

Ho, J. and Ermon, S. Generative adversarial imitation learning. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4565–4573, 2016.

Judah, K., Fern, A., and Dietterich, T. G. Active imitation learning via reduction to I.I.D. active learning. In de Freitas, N. and Murphy, K. P. (eds.), *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, August 14-18, 2012*, pp. 428–437. AUAI Press, 2012.

Juozapaitis, Z., Koul, A., Fern, A., Erwig, M., and Doshi-Velez, F. Explainable reinforcement learning via reward decomposition. In *IJCAI/ECAI Workshop on Explainable Artificial Intelligence*, 2019.

Kakade, S. M. and Langford, J. Approximately optimal approximate reinforcement learning. In Sammut, C. and Hoffmann, A. G. (eds.), *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002), University of New South Wales, Sydney, Australia, July 8-12, 2002*, pp. 267–274. Morgan Kaufmann, 2002.

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

Levine, S., Popovic, Z., and Koltun, V. Nonlinear inverse reinforcement learning with gaussian processes. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. C. N., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pp. 19–27, 2011.

Lopes, M., Melo, F., and Montesano, L. Active learning for reward estimation in inverse reinforcement learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 31–46. Springer, 2009.

Metelli, A. M., Pirotta, M., and Restelli, M. Compatible reward inverse reinforcement learning. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 2050–2059, 2017.

Metelli, A. M., Pirotta, M., and Restelli, M. On the use of the policy gradient and hessian in inverse reinforcement learning. *Intelligenza Artificiale*, 14(1):117–150, 2020. doi: 10.3233/IA-180011.

Ng, A. Y. and Russell, S. J. Algorithms for inverse reinforcement learning. In Langley, P. (ed.), *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pp. 663–670. Morgan Kaufmann, 2000.

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., and Peters, J. An algorithmic perspective on imitation learning. *Found. Trends Robotics*, 7(1-2):1–179, 2018. doi: 10.1561/2300000053.

Pirotta, M. and Restelli, M. Inverse reinforcement learning through policy gradient minimization. In Schuurmans, D. and Wellman, M. P. (eds.), *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pp. 1993–1999. AAAI Press, 2016.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Ramachandran, D. and Amir, E. Bayesian inverse reinforcement learning. In Veloso, M. M. (ed.), *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pp. 2586–2591, 2007.

Ramponi, G., Drappo, G., and Restelli, M. Inverse reinforcement learning from a gradient-based learner. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020a.

Ramponi, G., Likmeta, A., Metelli, A. M., Tirinzoni, A., and Restelli, M. Truly batch model-free inverse reinforcement learning about multiple intentions. In Chiappa, S. and Calandra, R. (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2359–2369. PMLR, 2020b.

Ratliff, N. D., Bagnell, J. A., and Zinkevich, M. Maximum margin planning. In Cohen, W. W. and Moore, A. W. (eds.), *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, pp. 729–736. ACM, 2006a. doi: 10.1145/1143844.1143936.

Ratliff, N. D., Bradley, D. M., Bagnell, J. A., and Chestnutt, J. E. Boosting structured prediction for imitation learning. In Schölkopf, B., Platt, J. C., and Hofmann, T. (eds.), *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pp. 1153–1160. MIT Press, 2006b.

Rockafellar, R. T. and Wets, R. J.-B. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Ross, S. and Bagnell, D. Efficient reductions for imitation learning. In Teh, Y. W. and Titterington, D. M. (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pp. 661–668. JMLR.org, 2010.

Ross, S., Gordon, G. J., and Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In Gordon, G. J., Dunson, D. B., and Dudík, M. (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, volume 15 of *JMLR Proceedings*, pp. 627–635. JMLR.org, 2011.

Russell, J. and Santos, E. Explaining reward functions in markov decision processes. In Barták, R. and Brawner, K. W. (eds.), *Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference, Sarasota, Florida, USA, May 19-22 2019*, pp. 56–61. AAAI Press, 2019.

Sutton, R. S., Barto, A. G., et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

Zanette, A., Kochenderfer, M. J., and Brunskill, E. Almost horizon-free structure-aware best policy identification with a generative model. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 5626–5635, 2019.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K. Maximum entropy inverse reinforcement learning. In Fox, D. and Gomes, C. P. (eds.), *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*, pp. 1433–1438. AAAI Press, 2008.

Ziebart, B. D., Bagnell, J. A., and Dey, A. K. Modeling interaction via the principle of maximum causal entropy. In Fürnkranz, J. and Joachims, T. (eds.), *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pp. 1255–1262. Omnipress, 2010.

A. Notation and Symbols

In this appendix, we report the explicit definitions of the operators and functions that we have employed in the main paper and in the appendix (Table 1).

Symbol	Signature	Definition
E	$\mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$	$(Ef)(s,a)=f(s)$
P	$\mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$	$(Pf)(s,a)=\sum_{s'\in\mathcal{S}} p(s'\|s,a)f(s')$
π	$\mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}}$	$(\pi f)(s)=\sum_{a\in\mathcal{A}}\pi(a\|s)f(s,a)$
$E^\intercal$	$\mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}}$	$(E^\intercal f)(s)=\sum_{a\in\mathcal{A}} f(s,a)$
$P^\intercal$	$\mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}}$	$(P^\intercal f)(s)=\sum_{(s',a)\in\mathcal{S}\times\mathcal{A}} p(s\|s',a)f(s',a)$
$\pi^\intercal$	$\mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$	$(\pi^\intercal f)(s,a)=\pi(a\|s)f(s)$
M	$\mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}}$	$(Mf)(s)=\max_{a\in\mathcal{A}} f(s,a)$
B^π	$\mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$	$(B^\pi f)(s,a)=f(s,a)\mathbb{1}\{\pi(a\|s)>0\}$
$\overline{B}^\pi$	$\mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$	$(\overline{B}^\pi f)(s,a)=f(s,a)\mathbb{1}\{\pi(a\|s)=0\}$
$I_{\mathcal{S}\times\mathcal{A}}$	$\mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$	$(I_{\mathcal{S}\times\mathcal{A}} f)(s,a)=f(s,a)$
$I_{\mathcal{S}}$	$\mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$	$(I_{\mathcal{S}} f)(s)=f(s)$
$(I_{\mathcal{S}\times\mathcal{A}}-\gamma P\pi)^{-1}$	$\mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$	$((I_{\mathcal{S}\times\mathcal{A}}-\gamma P\pi)^{-1} f)(s,a)=\sum_{i\in\mathbb{N}}\left((\gamma P\pi)^i f\right)(s,a)$
$(I_{\mathcal{S}}-\gamma \pi P)^{-1}$	$\mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$	$((I_{\mathcal{S}}-\gamma\pi P)^{-1} f)(s)=\sum_{i\in\mathbb{N}}\left((\gamma\pi P)^i f\right)(s)$
$\mathbf{1}_{\mathcal{S}}$	$\mathbb{R}^{\mathcal{S}}$	$\mathbf{1}_{\mathcal{S}}(s)=1$
$\mathbf{1}_{\mathcal{S}\times\mathcal{A}}$	$\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$	$\mathbf{1}_{\mathcal{S}\times\mathcal{A}}(s,a)=1$
$\mathbf{0}_{\mathcal{S}}$	$\mathbb{R}^{\mathcal{S}}$	$\mathbf{0}_{\mathcal{S}}(s)=0$
$\mathbf{0}_{\mathcal{S}\times\mathcal{A}}$	$\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$	$\mathbf{0}_{\mathcal{S}\times\mathcal{A}}(s,a)=0$

Table 1. Operators and functions.

B. Proofs and Derivations

In this section, we provide the proofs of the results that are reported in the main paper.

B.1. Proofs of Section 3

Lemma 3.1 (Feasible Reward Set Implicit (Ng & Russell, 2000))**.** *Let $\mathfrak{P}=(\mathcal{M},\pi^E)$ be an IRL problem. Let $r\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, then r is a feasible reward, i.e., $r\in\mathcal{R}_{\mathfrak{P}}$ if and only if for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ it holds that:*

$$\text{(i)} \qquad Q^{\pi^E}_{\mathcal{M}\cup r}(s,a)-V^{\pi^E}_{\mathcal{M}\cup r}(s)=0 \qquad \text{if } \pi^E(a|s)>0,$$

$$\text{(ii)} \qquad Q^{\pi^E}_{\mathcal{M}\cup r}(s,a)-V^{\pi^E}_{\mathcal{M}\cup r}(s)\leqslant 0 \qquad \text{if } \pi^E(a|s)=0.$$

Furthermore, if condition (ii) holds with the strict inequality, π^E is the unique optimal policy under r, i.e., $\Pi^*_{\mathcal{M}\cup r}=\{\pi^E\}$.

Proof. The proof is analogous of that of Theorem 3 of (Ng & Russell, 2000). For every state $s\in\mathcal{S}$, it must be that all actions $a^E\in\mathcal{A}$ such that $\pi^E(a^E|s)>0$ fulfill for all other actions $a\in\mathcal{A}$ that $Q^{\pi^E}_{\mathcal{M}\cup r}(s,a^E)\geqslant Q^{\pi^E}_{\mathcal{M}\cup r}(s,a)$. Furthermore, all actions $a^E\in\mathcal{A}$ such that $\pi^E(a^E|s)>0$ must yield the same performance equal to $Q^{\pi^E}_{\mathcal{M}\cup r}(s,a^E)=V^{\pi^E}_{\mathcal{M}\cup r}(s)$. □

Lemma B.1. *Let $\mathfrak{P}=(\mathcal{M},\pi^E)$ be an IRL problem. A Q-function satisfies condition of Lemma 3.1 if and only if there exist $\zeta\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\geqslant 0}$ and $V\in\mathbb{R}^{\mathcal{S}}$ such that:*

$$Q_{\mathcal{M}\cup r}=-\overline{B}^{\pi^E}\zeta+EV,$$

Furthermore, $\|V\|_\infty\leqslant\|Q_{\mathcal{M}\cup r}\|_\infty$.

Proof. We can easily see that all the Q-functions of the form $Q_{\mathcal{M}\cup r}=-\overline{B}^{\pi^E}\zeta+EV$ satisfy the conditions of Lemma 3.1. First notice that the corresponding value function is given by $V_{\mathcal{M}\cup r}=\pi^E Q_{\mathcal{M}\cup r}=V$, having observed that $\pi^E\overline{B}^{\pi^E}=\mathbf{0}_{\mathcal{S}}$ and $\pi^E E=I_{\mathcal{S}}$. Let $s\in\mathcal{S}$

and let $a\in\mathcal{A}$ be such that $\pi^E(a|s)>0$, then we have $Q_{\mathcal{M}\cup r}(s,a)=V(s)=V_{\mathcal{M}\cup r}(s)$, that is condition (i) of Lemma 3.1. Instead, if $a\in\mathcal{A}$ such that $\pi^E(a|s)=0$, then we have $Q_{\mathcal{M}\cup r}(s,a)=-\zeta(s,a)+V(s)=-\zeta(s,a)+V_{\mathcal{M}\cup r}(s)\leqslant V_{\mathcal{M}\cup r}(s)$, that is condition (ii) of Lemma 3.1. For the converse, suppose that $Q_{\mathcal{M}\cup r}$ satisfies conditions of Lemma 3.1. We take $V=V_{\mathcal{M}\cup r}$ and $\zeta=EV_{\mathcal{M}\cup r}-Q_{\mathcal{M}\cup r}\geqslant 0$.

For the second statement, consider a state $s\in\mathcal{S}$ and an action $a\in\mathcal{A}$ such that $\pi^E(a|s)>0$. Then, we have $Q_{\mathcal{M}\cup r}(s,a)=V(s)$. Consequently, $\|V\|_\infty\leqslant\|Q_{\mathcal{M}\cup r}\|_\infty$. □

Lemma 3.2 (Feasible Reward Set Explicit)**.** *Let $\mathfrak{P}=(\mathcal{M},\pi^E)$ be an IRL problem. Let $r\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, then r is a feasible reward, i.e., $r\in\mathcal{R}_{\mathfrak{P}}$ if and only if there exist $\zeta\in\mathbb{R}_{\geqslant 0}^{\mathcal{S}\times\mathcal{A}}$ and $V\in\mathbb{R}^{\mathcal{S}}$ such that:*

$$r=-\overline{B}^{\pi^E}\zeta+(E-\gamma P)V.$$

Proof. Simply recall that $Q_{\mathcal{M}\cup r}=\left(I_{\mathcal{S}\times\mathcal{A}}-\gamma P\pi^E\right)^{-1}r$ and that for $\gamma<1$ the matrix is invertible. In other words, having fixed π^E, P, and $\gamma<1$, there is a one-to-one correspondence between Q-functions and rewards. From Lemma B.1, we have:

$$r=\left(I_{SA}-\gamma P\pi^E\right)\left(-\overline{B}^{\pi^E}\zeta+EV\right)=-\overline{B}^{\pi^E}\zeta+\gamma P\pi^E\overline{B}^{\pi^E}\zeta+\left(E-\gamma P\pi^E E\right)V,$$

and we observe that $\pi^E\overline{B}^{\pi^E}=\mathbf{0}_{\mathcal{S}}$ and $\pi^E E=I_{\mathcal{S}}$. □

Theorem 3.1 (Error Propagation)**.** *Let $\mathfrak{P}=(\mathcal{M},\pi^E)$ and $\widehat{\mathfrak{P}}=(\widehat{\mathcal{M}},\widehat{\pi}^E)$ be two IRL problems. Then, for any $r\in\mathcal{R}_{\mathfrak{P}}$ such that $r=-\overline{B}^{\pi^E}\zeta+(E-\gamma P)V$ and $\|r\|_\infty\leqslant R_{\max}$ there exists $\widehat{r}\in\mathcal{R}_{\widehat{\mathfrak{P}}}$ such that element-wise it holds that:*

$$|r-\widehat{r}|\leqslant\overline{B}^{\pi^E}B^{\widehat{\pi}^E}\zeta+\gamma\left|\left(P-\widehat{P}\right)V\right|.$$

Furthermore, $\|\zeta\|_\infty\leqslant\frac{R_{\max}}{1-\gamma}$ and $\|V\|_\infty\leqslant\frac{R_{\max}}{1-\gamma}$.

Proof. From Lemma 3.2, we can express the reward functions belonging to $\mathcal{R}_{\mathfrak{P}}$ and $\mathcal{R}_{\widehat{\mathfrak{P}}}$ as:

$$r=-\overline{B}^{\pi^E}\zeta+(E-\gamma P)V,$$
$$\widehat{r}=-\overline{B}^{\widehat{\pi}^E}\widehat{\zeta}+\left(E-\gamma\widehat{P}\right)\widehat{V},$$

where $V,\widehat{V}\in\mathbb{R}^{\mathcal{S}}$ and $\zeta,\widehat{\zeta}\in\mathbb{R}_{\geqslant 0}^{\mathcal{S}\times\mathcal{A}}$. Since we look for the existence of $\widehat{r}\in\mathcal{R}_{\widehat{\mathfrak{P}}}$, we provide a specific choice of $\widehat{V}$ and $\widehat{\zeta}$: $\widehat{V}=V$ and $\widehat{\zeta}=\overline{B}^{\pi^E}\zeta$. Thus, we have:

$$\begin{aligned}r-\widehat{r}&=-\left(\overline{B}^{\pi^E}\zeta-\overline{B}^{\widehat{\pi}^E}\overline{B}^{\pi^E}\zeta\right)-\gamma\left(P-\widehat{P}\right)V\\&=-\left(I_{\mathcal{S}\times\mathcal{A}}-\overline{B}^{\widehat{\pi}^E}\right)\overline{B}^{\pi^E}\zeta-\gamma\left(P-\widehat{P}\right)V\\&=-B^{\widehat{\pi}^E}\overline{B}^{\pi^E}\zeta-\gamma\left(P-\widehat{P}\right)V,\end{aligned}$$

having observed that $\overline{B}^{\pi}$ is linear and commutative and by observing that $B^\pi+\overline{B}^\pi=I_{\mathcal{S}\times\mathcal{A}}$. We now take the absolute value and by applying the triangular inequality, we obtain:

$$|r-\widehat{r}|\leqslant B^{\widehat{\pi}^E}\overline{B}^{\pi^E}\zeta+\gamma\left|\left(P-\widehat{P}\right)V\right|.$$

We now bound the L_∞-norms by using the condition $\|r\|_\infty\leqslant R_{\max}$. First of all, we observe that it must be that $\|\zeta\|_\infty\leqslant\frac{R_{\max}}{1-\gamma}$ being the advantage function. Then, from Lemma B.1 we have $\|V\|_\infty\leqslant\frac{R_{\max}}{1-\gamma}$. □

B.2. Proofs of Section 4

Lemma B.2 (Simulation Lemma 0)**.** *Let $\mathcal{M}'=(\mathcal{S},\mathcal{A},p',\gamma')$ be an MDP\R, let $r,\widehat{r}\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ be two reward functions, and let $\pi\in\Delta_{\mathcal{S}}^{\mathcal{A}}$ be a policy. Then, the following equality holds element-wise:*

$$V^\pi_{\mathcal{M}'\cup r}-V^\pi_{\mathcal{M}'\cup\widehat{r}}=\left(I_{\mathcal{S}}-\gamma'\pi P'\right)^{-1}\pi(r-\widehat{r}).$$

Proof. The result is a simple application of the Bellman's equation:

$$V^\pi_{\mathcal{M}'\cup r}-V^\pi_{\mathcal{M}'\cup\widehat{r}}=\left(I_{\mathcal{S}}-\gamma'\pi P'\right)^{-1}\pi r-\left(I_{\mathcal{S}}-\gamma'\pi P'\right)^{-1}\pi\widehat{r}=\left(I_{\mathcal{S}}-\gamma'\pi P'\right)^{-1}\pi(r-\widehat{r}).$$

□

Lemma 4.1 (Simulation Lemma 1)**.** *Let $\mathcal{M}'=(\mathcal{S},\mathcal{A},p',\gamma')$ be an MDP\R, let $r,\widehat{r}\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ be two reward functions. Then, for every $\pi^*\in\Pi^*_{\mathcal{M}'\cup r}$ and $\widehat{\pi}^*\in\Pi^*_{\mathcal{M}'\cup\widehat{r}}$ optimal policies for the MDPs $\mathcal{M}'\cup r$ and $\mathcal{M}'\cup\widehat{r}$ respectively, the following inequalities hold element-wise:*

$$Q^*_{\mathcal{M}'\cup r}-Q^*_{\mathcal{M}'\cup\widehat{r}}\leqslant\left(I_{\mathcal{S}\times\mathcal{A}}-\gamma'P'\pi^*\right)^{-1}(r-\widehat{r}),$$
$$Q^*_{\mathcal{M}'\cup r}-Q^*_{\mathcal{M}'\cup\widehat{r}}\geqslant\left(I_{\mathcal{S}\times\mathcal{A}}-\gamma'P'\widehat{\pi}^*\right)^{-1}(r-\widehat{r}).$$

In particular, it holds that:

$$\left\|Q^*_{\mathcal{M}'\cup r}-Q^*_{\mathcal{M}'\cup\widehat{r}}\right\|_\infty\leqslant\max_{\pi\in\{\pi^*,\widehat{\pi}^*\}}\left\|\left(I_{\mathcal{S}\times\mathcal{A}}-\gamma'P'\pi\right)^{-1}(r-\widehat{r})\right\|_\infty.$$

Proof. We exploit the facts: $Q^*_{\mathcal{M}'\cup\widehat{r}}\geqslant Q^{\pi^*}_{\mathcal{M}'\cup\widehat{r}}$ and $Q^*_{\mathcal{M}'\cup r}=Q^{\pi^*}_{\mathcal{M}'\cup r}$:

$$\begin{aligned}Q^*_{\mathcal{M}'\cup r}-Q^*_{\mathcal{M}'\cup\widehat{r}}&\leqslant Q^{\pi^*}_{\mathcal{M}'\cup r}-Q^{\pi^*}_{\mathcal{M}'\cup\widehat{r}}\\&=\left(I_{\mathcal{S}\times\mathcal{A}}-\gamma'P'\pi^*\right)^{-1}r-\left(I_{\mathcal{S}\times\mathcal{A}}-\gamma'P'\pi^*\right)^{-1}\widehat{r}\\&=\left(I_{\mathcal{S}\times\mathcal{A}}-\gamma'P'\pi^*\right)^{-1}(r-\widehat{r}).\end{aligned}$$

For the second inequality, we exploit the facts: $Q^*_{\mathcal{M}'\cup r}\geqslant Q^{\widehat{\pi}^*}_{\mathcal{M}'\cup r}$ and $Q^*_{\mathcal{M}'\cup\widehat{r}}=Q^{\widehat{\pi}^*}_{\mathcal{M}'\cup\widehat{r}}$. The result is obtained by the very same derivation reversing the roles of π^* and $\widehat{\pi}^*$.

For the L_∞-norm inequality, we simply observe:

$$\begin{aligned}\|Q^*_{\mathcal{M}'\cup r}-Q^*_{\mathcal{M}'\cup\widehat{r}}\|_\infty&\leqslant\max\left\{\left\|\left(I_{\mathcal{S}\times\mathcal{A}}-\gamma'P'\pi^*\right)^{-1}(r-\widehat{r})\right\|_\infty,\left\|\left(I_{\mathcal{S}\times\mathcal{A}}-\gamma'P'\widehat{\pi}^*\right)^{-1}(r-\widehat{r})\right\|_\infty\right\}\\&=\max_{\pi\in\{\pi^*,\widehat{\pi}^*\}}\left\|\left(I_{\mathcal{S}\times\mathcal{A}}-\gamma'P'\pi\right)^{-1}(r-\widehat{r})\right\|_\infty.\end{aligned}$$

□

Lemma B.3 (Sum of losses (Zanette et al., 2019))**.** *Let $\mathcal{M}'=(\mathcal{S},\mathcal{A},p',\gamma')$ be an MDP\R, let $r\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ be a reward function, and let $\pi\in\Delta^{\mathcal{A}}_{\mathcal{S}}$ be a policy. Then, the following equality holds element-wise:*

$$V^*_{\mathcal{M}'\cup r}-V^\pi_{\mathcal{M}'\cup r}=-\left(I_{\mathcal{S}}-\gamma'P'\pi\right)^{-1}\pi A^*_{\mathcal{M}'\cup r},$$

where $A^*_{\mathcal{M}'\cup r}=Q^*_{\mathcal{M}'\cup r}-EV^*_{\mathcal{M}'\cup r}$ is the advantage function.

Proof. The result is proved in Lemma 6.1 of (Kakade & Langford, 2002). □

B.3. Proofs of Section 5

We introduce the notion of *good event*, i.e., the event under which the confidence intervals hold uniformly over $\mathcal{S}\times\mathcal{A}$ and for every $k\geqslant1$. At each iteration $k\in[K]$, we need to guarantee that the good event holds simultaneously for one target reward $\overline{r}\in\mathcal{R}_{\mathfrak{P}}$ (for instance the expert's reward function r^E) and one target reward $\check{r}_k\in\mathcal{R}_{\widehat{\mathfrak{P}}_k}$ (for instance the one outputted by an IRL algorithm $\mathscr{A}(\mathcal{R}_{\widehat{\mathfrak{P}}_k})$. We decompose the reward functions based on Lemma 3.2 as $\overline{r}=-\overline{B}^{\pi^E}\zeta+(E-\gamma P)V$ and $\check{r}_k=-\overline{B}^{\widehat{\pi}^E_k}\widehat{\zeta}_k+(E-\gamma\widehat{P}_k)\widehat{V}_k$.

Lemma B.4 (Good Event)**.** *Let $\delta\in(0,1)$, define the* good event *$\mathcal{E}$ as the event such that the following inequalities hold simultaneously for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $k\geqslant1$:*

$$\left(\overline{B}^{\pi^E}B^{\widehat{\pi}^E_k}\zeta\right)(s,a)\leqslant\frac{R_{\max}}{1-\gamma}\mathbb{1}\{N_k(s)=0\},$$
$$\left(\overline{B}^{\widehat{\pi}^E_k}B^{\pi^E}\widehat{\zeta}_k\right)(s,a)\leqslant\frac{R_{\max}}{1-\gamma}\mathbb{1}\{N_k(s)=0\},$$
$$\left|\left(P-\widehat{P}_k\right)V\right|(s,a)\leqslant\frac{R_{\max}}{1-\gamma}\sqrt{\frac{2\ell_k(s,a)}{N^+_k(s,a)}},$$
$$\left|\left(P-\widehat{P}_k\right)\widehat{V}_k\right|(s,a)\leqslant\frac{R_{\max}}{1-\gamma}\sqrt{\frac{2\ell_k(s,a)}{N^+_k(s,a)}},$$

where ζ, $\widehat{\zeta}_k$, V, and $\widehat{V}_k$ are defined in Theorem 3.1 and $\ell_k(s,a)=\log\left(\frac{12SA(N^+_k(s,a))^2}{\delta}\right)$. Then, $\Pr(\mathcal{E})\geqslant1-\delta$.

Proof. First of all, we observe that the bound for the policy term is independent on the probability and that $\|\zeta\|_\infty, \|\widehat{\zeta}_k\|_\infty \leqslant \frac{R_{\max}}{1-\gamma}$. Thus, we focus on the transition model. As explained in (Lattimore & Szepesvári, 2020, Section 4.4), suppose we visit m times a state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, thus we will change at most m times the estimate of $p(\cdot|s,a)$. For this reason, we will denote with $p_{[m]}$ the estimate of the transition model made with m samples. For the sake of brevity, we denote $\beta_{N_k(s,a)}(s,a) = \frac{R_{\max}}{1-\gamma}\sqrt{\frac{2\ell_k(s,a)}{N_k^+(s,a)}}$. We just prove the statement for V enforcing to hold with probability $\frac{\delta}{2}$:

$$\begin{aligned}
\Pr\left(\exists k \geqslant 1, \exists (s,a) \in \mathcal{S} \times \mathcal{A} : \left|\left(P - \widehat{P}_k\right)V\right|(s,a) > \beta_{N_k(s,a)}(s,a)\right) &\leqslant \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Pr\left(\exists k \geqslant 1 : \left|\left(P - \widehat{P}_k\right)V\right|(s,a) > \beta_{N_k(s,a)}(s,a)\right) && \text{(P.1)} \\
&= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Pr\left(\exists m \geqslant 0 : \left|\left(P - \widehat{P}_{[m]}\right)V\right|(s,a) > \beta_m(s,a)\right) && \text{(P.2)} \\
&\leqslant \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{m \geqslant 0} \frac{\delta}{6SA(m^+)^2} && \text{(P.3)} \\
&= \frac{\delta}{6}\left(1 + \frac{\pi^2}{6}\right) \leqslant \frac{\delta}{2},
\end{aligned}$$

where line (P.1) derives from a union bound over $\mathcal{S} \times \mathcal{A}$, line (P.2) follows from observing that we just need to enforce the condition when the transition model estimate changes, line (P.3) from a union bound over the possible values of m and applying Höeffding's inequality, having recalled that $\|V\|_\infty, \|\widehat{V}_k\|_\infty \leqslant \frac{R_{\max}}{1-\gamma}$. □

Corollary B.1. *Let $\mathscr{S}$ be an iterative sampling strategy. Let $\mathcal{R}_{\mathfrak{P}}$ be the exact feasible set and $\mathcal{R}_{\widehat{\mathfrak{P}}_K}$ the estimated feasible set after K iterations. Under the good event $\mathcal{E}$, the conditions of Definition 5.1 are satisfied when either of the following conditions are satisfied:*

(i) $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{1}{1-\gamma'} \mathcal{C}_K(s,a) \leqslant \epsilon$;

(ii) $\sup_{\pi\in\Pi^\dagger} \sup_{\mu_0\in\Delta^{\mathcal{S}\times\mathcal{A}}} \mu_0^\intercal \left(I_{\mathcal{S}\times\mathcal{A}} - \gamma' P'\pi\right)^{-1} \mathcal{C}_K \leqslant \epsilon$, *where* $\Pi^\dagger = \left(\bigcap_{r\in\mathcal{R}_{\mathfrak{P}}} \Pi^*_{\mathcal{M}'\cup r}\right) \cup \left(\bigcap_{\widehat{r}\in\mathcal{R}_{\widehat{\mathfrak{P}}_K}} \Pi^*_{\mathcal{M}'\cup\widehat{r}}\right)$.

Proof. We apply Lemma 4.1 followed by the reward error propagation of Theorem 3.1:

$$\inf_{\widehat{r}_K\in\mathcal{R}_{\widehat{\mathfrak{P}}_K}} \left\|Q^*_{\mathcal{M}'\cup\overline{r}} - Q^*_{\mathcal{M}'\cup\widehat{r}_K}\right\|_\infty \leqslant \inf_{\widehat{r}_K\in\mathcal{R}_{\widehat{\mathfrak{P}}_K}} \max_{\pi\in\{\overline{\pi}^*,\widehat{\pi}^*_K\}} \left\|\left(I_{\mathcal{S}\times\mathcal{A}} - \gamma' P'\pi\right)^{-1}(\overline{r} - \widehat{r}_K)\right\|_\infty \leqslant \max_{\pi\in\{\overline{\pi}^*,\widehat{\pi}^*_K\}} \left\|\left(I_{\mathcal{S}\times\mathcal{A}} - \gamma' P'\pi\right)^{-1}\mathcal{C}_K\right\|_\infty,$$

$$\inf_{r\in\mathcal{R}_{\mathfrak{P}}} \left\|Q^*_{\mathcal{M}'\cup\breve{r}_K} - Q^*_{\mathcal{M}'\cup r}\right\|_\infty \leqslant \inf_{r\in\mathcal{R}_{\mathfrak{P}}} \max_{\pi\in\{\breve{\pi}^*_K,\pi^*\}} \left\|\left(I_{\mathcal{S}\times\mathcal{A}} - \gamma' P'\pi\right)^{-1}(\overline{r}_K - r)\right\|_\infty \leqslant \max_{\pi\in\{\breve{\pi}^*_K,\pi^*\}} \left\|\left(I_{\mathcal{S}\times\mathcal{A}} - \gamma' P'\pi\right)^{-1}\mathcal{C}_K\right\|_\infty,$$

holding for policies arbitrarily selected in the corresponding sets: $\overline{\pi}^* \in \Pi^*_{\mathcal{M}'\cup\overline{r}}$, $\widehat{\pi}^*_K \in \Pi^*_{\mathcal{M}'\cup\widehat{r}_K}$, $\breve{\pi}^*_K \in \Pi^*_{\mathcal{M}'\cup\breve{r}_K}$, and $\pi^* \in \Pi^*_{\mathcal{M}'\cup r}$, with also $\widehat{r}_K \in \operatorname{arginf}_{\widehat{r}_K\in\mathcal{R}_{\widehat{\mathfrak{P}}_K}} |\overline{r} - \widehat{r}_K|$ and $r \in \operatorname{arginf}_{r\in\mathcal{R}_{\mathfrak{P}}} |\breve{r}_K - r|$. Since $\overline{r}, r \in \mathcal{R}_{\mathfrak{P}}$ and $\widehat{r}_K, \breve{r}_K \in \mathcal{R}_{\widehat{\mathfrak{P}}_K}$, we can select:

$$\overline{\pi}^*, \pi^* \in \Pi^\dagger_{\mathcal{M}'\cup\mathcal{R}_{\mathfrak{P}}} := \bigcap_{r\in\mathcal{R}_{\mathfrak{P}}} \Pi^*_{\mathcal{M}'\cup r} \qquad \text{and} \qquad \widehat{\pi}^*_K, \breve{\pi}^*_K \in \Pi^\dagger_{\mathcal{M}'\cup\mathcal{R}_{\widehat{\mathfrak{P}}_K}} := \bigcap_{\widehat{r}\in\mathcal{R}_{\widehat{\mathfrak{P}}_K}} \Pi^*_{\mathcal{M}'\cup\widehat{r}}.$$

Therefore, having defined $\Pi^\dagger = \Pi^\dagger_{\mathcal{M}'\cup\mathcal{R}_{\mathfrak{P}}} \cup \Pi^\dagger_{\mathcal{M}'\cup\mathcal{R}_{\widehat{\mathfrak{P}}_K}}$, we have:

$$\begin{aligned}
\max\left\{\inf_{\widehat{r}_K\in\mathcal{R}_{\widehat{\mathfrak{P}}_K}} \left\|Q^*_{\mathcal{M}'\cup\overline{r}} - Q^*_{\mathcal{M}'\cup\widehat{r}_K}\right\|_\infty, \inf_{r\in\mathcal{R}_{\mathfrak{P}}} \left\|Q^*_{\mathcal{M}'\cup\breve{r}_K} - Q^*_{\mathcal{M}'\cup r}\right\|_\infty\right\} &\leqslant \sup_{\pi\in\Pi^\dagger} \left\|\left(I_{\mathcal{S}\times\mathcal{A}} - \gamma' P'\pi\right)^{-1}\mathcal{C}_K\right\|_\infty \\
&= \sup_{\pi\in\Pi^\dagger} \sup_{\mu_0\in\Delta^{\mathcal{S}\times\mathcal{A}}} \mu_0^\intercal \left(I_{\mathcal{S}\times\mathcal{A}} - \gamma' P'\pi\right)^{-1}\mathcal{C}_K.
\end{aligned}$$

This proves the statement (ii). For the statement (i), it suffices to observe that:

$$\sup_{\pi\in\Pi^\dagger} \sup_{\mu_0\in\Delta^{\mathcal{S}\times\mathcal{A}}} \mu_0^\intercal \left(I_{\mathcal{S}\times\mathcal{A}} - \gamma' P'\pi\right)^{-1}\mathcal{C}_K \leqslant \frac{1}{1-\gamma'} \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{C}_K(s,a),$$

having observed that $\left\|\mu_0^\intercal \left(I_{\mathcal{S}\times\mathcal{A}} - \gamma' P'\pi\right)^{-1}\right\|_\infty \leqslant \frac{1}{1-\gamma'}$. □

We now move to study the sample complexity of the uniform sampling strategy.

Theorem 5.1 (Sample Complexity of Uniform Sampling IRL)**.** *If Algorithm 1 stops at iteration K with accuracy ϵ_K, then with probability at least $1-\delta$ it fulfills Definition 5.1, for arbitrary target reward functions $\overline{r}$ and $\breve{r}$, with a number of samples* *upper bounded by:*

$$n\leqslant\widetilde{\mathcal{O}}\left(\frac{\gamma^2 R_{\max}^2 SA}{(1-\gamma')^2(1-\gamma)^2\epsilon_K^2}\right).$$

Proof. We start from Corollary B.1 and we further bound:

$$\frac{1}{1-\gamma'}\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathcal{C}_k(s,a)=\frac{R_{\max}}{(1-\gamma')(1-\gamma)}\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left(\mathbb{1}\{N_k(s)=0\}+\gamma\sqrt{\frac{2\ell_k(s,a)}{N_k^+(s,a)}}\right).$$

After K iterations having collected a total of $N_K=Kn_{\max}$ samples, we know that $N_K^+(s,a)\geqslant\frac{N_K}{SA}\geqslant 1$ and, therefore, $\mathbb{1}\{N_K(s)=0\}=0$. Thus, it suffices to enforce for every $(s,a)\in\mathcal{S}\times\mathcal{A}$:

$$\frac{R_{\max}\gamma}{(1-\gamma')(1-\gamma)}\sqrt{\frac{2\ell_K(s,a)}{N_K^+(s,a)}}=\epsilon_K\quad\Longrightarrow\quad N_K(s,a)=\frac{2\gamma^2 R_{\max}^2\ell_K(s,a)}{(1-\gamma')^2(1-\gamma)^2\epsilon_K^2}.$$

From Lemma B.9, we conclude that the following number of samples is sufficient to ensure the accuracy ϵ:

$$\begin{aligned}N_K(s,a)&\leqslant\frac{-4\gamma^2R_{\max}^2}{(1-\gamma')^2(1-\gamma)^2\epsilon_K^2}W_{-1}\left(-\frac{(1-\gamma')^2(1-\gamma)^2\epsilon_K^2}{4\gamma^2R_{\max}^2}\sqrt{\frac{\delta}{12SA}}\right)\\&\leqslant\frac{8\gamma^2R_{\max}^2}{(1-\gamma')^2(1-\gamma)^2\epsilon_K^2}\log\left(\frac{4\gamma^2R_{\max}^2}{(1-\gamma')^2(1-\gamma)^2\epsilon_K^2}\sqrt{\frac{12SA}{\delta}}\right)\\&=\widetilde{\mathcal{O}}\left(\frac{\gamma^2R_{\max}^2}{(1-\gamma')^2(1-\gamma)^2\epsilon_K^2}\right).\end{aligned}$$

By summing $n=\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}N_K(s,a)$, we obtain the result. □

B.4. Proofs of Section 6

In this section, we perform the sample complexity analysis of the TRAVEL algorithm.

B.4.1. Correctness of the Algorithm

The goal of this section is to prove that the algorithm will retain, in every iteration k, some optimal policy π^* of MDP $\mathcal{M}'\cup\bar{r}$, some optimal policy $\widehat{\pi}^*_{k+1}$ of MDP $\mathcal{M}'\cup\check{r}_k$, and $\widehat{r}_k=\mathscr{A}(\mathcal{R}_{\widehat{\mathfrak{P}}_k})$ is obtained by using IRL algorithm $\mathscr{A}$. For every iteration $k\in[K]$, we define the following symbols:

$$\begin{aligned}&\epsilon_0=\frac{1}{4(1-\gamma')},\\&\epsilon_k^\pi=\sup_{\mu_0\in\Delta^{\mathcal{S}\times\mathcal{A}}}\mu_0^\intercal\left(I_{\mathcal{S}\times\mathcal{A}}-\gamma'P'\pi\right)\mathcal{C}_k,\\&\epsilon_k=\max_{\pi\in\Pi_{k-1}^{\mathscr{A}}}\epsilon_k^\pi,\\&\Pi_k^{\mathscr{A}}=\left\{\pi\in\Delta_{\mathcal{S}}^{\mathcal{A}}:\sup_{\mu_0\in\Delta^{\mathcal{S}}}\mu_0^\intercal\left(V^*_{\mathcal{M}'\cup\widehat{r}_k}-V^\pi_{\mathcal{M}'\cup\widehat{r}_k}\right)\leqslant 4\epsilon_k\right\},\end{aligned}$$

Moreover, we recall the two special sets of policies, for every $k\in[K]$:

$$\pi^*\in\Pi^\dagger_{\mathcal{M}'\cup\mathcal{R}_{\mathfrak{P}}}=\bigcap_{r\in\mathcal{R}_{\mathfrak{P}}}\Pi^*_{\mathcal{M}'\cup r},\qquad\widehat{\pi}^*_k\in\Pi^\dagger_{\mathcal{M}'\cup\mathcal{R}_{\widehat{\mathfrak{P}}_k}}=\bigcap_{\widehat{r}\in\mathcal{R}_{\widehat{\mathfrak{P}}_k}}\Pi^*_{\mathcal{M}'\cup\widehat{r}}$$

The following analysis will be conducted by using the symbols π^* and $\widehat{\pi}^*_k$ as *arbitrary* policies belonging to the corresponding sets. Notice, by the way, that, under Assumption 4.1, we have $(\pi')^E\in\Pi^\dagger_{\mathcal{M}'\cup\mathcal{R}_{\mathfrak{P}}}$.

Lemma B.5 (π^* propagation)**.** *Under the good event $\mathcal{E}$, if $\pi^*,\widehat{\pi}^*_k\in\Pi^{\mathscr{A}}_{k-1}$ then $\pi^*\in\Pi^{\mathscr{A}}_k$.*

Proof. Let us start with the following inequality, defined in terms of $r\in\mathcal{R}_{\mathfrak{P}}$, that will be specified later:

$$V^*_{\mathcal{M}'\cup\widehat{r}_k}-V^{\pi^*}_{\mathcal{M}'\cup\widehat{r}_k}=\underbrace{V^*_{\mathcal{M}'\cup\widehat{r}_k}-V^*_{\mathcal{M}'\cup r}}_{\text{(a)}}+\underbrace{V^*_{\mathcal{M}'\cup r}-V^{\pi^*}_{\mathcal{M}'\cup\widehat{r}_k}}_{\text{(b)}}.$$

For (b) we observe that $V^{*}_{\mathcal{M}'\cup r}=V^{\pi^*}_{\mathcal{M}'\cup r}$, that follows by definition of π^*. Now we have:

$$\left|V^{\pi^*}_{\mathcal{M}'\cup r}-V^{\pi^*}_{\mathcal{M}'\cup \widehat{r}_k}\right|\leqslant\left(I_{\mathcal{S}}-\gamma'\pi^*P'\right)^{-1}\pi^*\left|\widehat{r}_k-r\right| \tag{P.4}$$

$$\leqslant\left(I_{\mathcal{S}}-\gamma'\pi^*P'\right)^{-1}\pi^*\mathcal{C}_k, \tag{P.5}$$

where line (P.4) derives from Lemma B.2, line (P.5) from the reward error propagation (Theorem 3.1), having taken $r\in\operatorname{arginf}_{r\in\mathcal{R}_{\mathfrak{P}}}|\widehat{r}_k-r|$, and from the good event definition (Lemma B.4). By taking the L_∞-norm, and recalling that $\pi^*\in\Pi^{\mathscr{A}}_{k-1}$, we deduce:

$$\sup_{\mu_0\in\Delta^{\mathcal{S}}}\mu_0^{\mathsf{T}}\left(V^{*}_{\mathcal{M}'\cup r}-V^{\pi^*}_{\mathcal{M}'\cup \widehat{r}_k}\right)=\epsilon_k^{\pi^*}\leqslant\max_{\pi\in\Pi^{\mathscr{A}}_{k-1}}\epsilon_k^{\pi}=\epsilon_k. \tag{P.6}$$

For (a) instead we observe that:

$$V^{*}_{\mathcal{M}'\cup \widehat{r}_k}-V^{*}_{\mathcal{M}'\cup r}=V^{\widehat{\pi}^*_k}_{\mathcal{M}'\cup \widehat{r}_k}-V^{\pi^*}_{\mathcal{M}'\cup r} \tag{P.7}$$

$$=\left|\max_{\pi\in\Pi^{\mathscr{A}}_{k-1}}V^{\pi}_{\mathcal{M}'\cup \widehat{r}_k}-\max_{\pi\in\Pi^{\mathscr{A}}_{k-1}}V^{\pi}_{\mathcal{M}'\cup r}\right| \tag{P.8}$$

$$\leqslant\max_{\pi\in\Pi^{\mathscr{A}}_{k-1}}\left|V^{\pi}_{\mathcal{M}'\cup \widehat{r}_k}-V^{\pi}_{\mathcal{M}'\cup r}\right|,$$

where line (P.7) follows by definitions of $\widehat{\pi}^*_k$ and π^* and line (P.8) is obtained by recalling that $\pi^*,\widehat{\pi}^*_k\in\Pi^{\mathscr{A}}_{k-1}$. Thus, we have:

$$\sup_{\mu_0\in\Delta^{\mathcal{S}}}\max_{\pi\in\Pi^{\mathscr{A}}_{k-1}}\mu_0^{\mathsf{T}}\left(V^{\pi}_{\mathcal{M}'\cup \widehat{r}_k}-V^{\pi}_{\mathcal{M}'\cup r}\right)=\epsilon_k,$$

with an analogous derivation as the one employed starting from line (P.4), taking again $r\in\operatorname{arginf}_{r\in\mathcal{R}_{\mathfrak{P}}}|\widehat{r}_k-r|$. It follows that $\sup_{\mu_0\in\Delta^{\mathcal{S}}}\mu_0^{\mathsf{T}}\left(V^{*}_{\mathcal{M}'\cup \widehat{r}_k}-\widehat{V}^{\pi^*}_{\mathcal{M}'\cup \widehat{r}_k}\right)\leqslant 2\epsilon_k\leqslant 4\epsilon_k$, from which $\pi^*\in\Pi^{\mathscr{A}}_{k}$. □

Lemma B.6 (Accuracy Improvement)**.** *Under the good event $\mathcal{E}$, for every $\pi\in\Delta^{\mathcal{A}}_{\mathcal{S}}$ and $k'>k$ there exists $\widehat{r}_{k'}\in\mathcal{R}_{\widehat{\mathfrak{P}}_{k'}}$ such that:*

$$\sup_{\mu_0\in\Delta^{\mathcal{S}}}\mu_0^{\mathsf{T}}\left(V^{\pi}_{\mathcal{M}'\cup \widehat{r}_{k'}}-V^{\pi}_{\mathcal{M}'\cup \widehat{r}_k}\right)\leqslant 2\epsilon_k^{\pi}.$$

Proof. Under the good event, the confidence intervals are non-increasing in k, i.e., for $k'>k$ we have $\mathcal{C}_{k'}\leqslant\mathcal{C}_k$. Consequently, having fixed a policy $\pi\in\Delta^{\mathcal{A}}_{\mathcal{S}}$, it follows that $\epsilon^{\pi}_{k'}\leqslant\epsilon^{\pi}_{k}$. Thus, let $r\in\mathcal{R}_{\mathfrak{P}}$, that will be specified later, and $\widehat{r}_{k'}\in\mathcal{R}_{\widehat{\mathfrak{P}}_{k'}}$. We have:

$$V^{\pi}_{\mathcal{M}'\cup \widehat{r}_{k'}}-V^{\pi}_{\mathcal{M}'\cup \widehat{r}_k}=V^{\pi}_{\mathcal{M}'\cup \widehat{r}_{k'}}-V^{\pi}_{\mathcal{M}'\cup r}+V^{\pi}_{\mathcal{M}'\cup r}-V^{\pi}_{\mathcal{M}'\cup \widehat{r}_k}$$

$$\leqslant\left(I_{\mathcal{S}}-\gamma'\pi P'\right)^{-1}\pi\left(|\widehat{r}_{k'}-r|+|r-\widehat{r}_k|\right) \tag{P.9}$$

$$\leqslant\left(I_{\mathcal{S}}-\gamma'\pi P'\right)^{-1}\pi\left(\mathcal{C}_{k'}+\mathcal{C}_k\right), \tag{P.10}$$

where line (P.9) is obtained from Lemma B.2, line (P.10) derives from the reward error propagation (Theorem 3.1), having taken $\widehat{r}_{k'}\in\operatorname{arginf}_{\widehat{r}_{k'}\in\mathcal{R}_{\widehat{\mathfrak{P}}_k}}|\widehat{r}_{k'}-r|$ for the first addendum and $r\in\operatorname{arginf}_{r\in\mathcal{R}_{\mathfrak{P}}}|r-\widehat{r}_k|$ for the second addendum, and from the good event definition (Lemma B.4). Thus, we have:

$$\sup_{\mu_0\in\Delta^{\mathcal{S}}}\mu_0^{\mathsf{T}}\left(V^{\pi}_{\mathcal{M}'\cup \widehat{r}_{k'}}-V^{\pi}_{\mathcal{M}'\cup \widehat{r}_k}\right)\leqslant\sup_{\mu_0\in\Delta^{\mathcal{S}}}\mu_0^{\mathsf{T}}\left(I_{\mathcal{S}}-\gamma'\pi P'\right)^{-1}\pi\mathcal{C}_{k'}+\sup_{\mu_0\in\Delta^{\mathcal{S}}}\mu_0^{\mathsf{T}}\left(I_{\mathcal{S}}-\gamma'\pi P'\right)^{-1}\pi\mathcal{C}_{k}=\epsilon^{\pi}_{k'}+\epsilon^{\pi}_{k}\leqslant 2\epsilon^{\pi}_{k},$$

having just recalled that $\epsilon^{\pi}_{k'}\leqslant\epsilon^{\pi}_{k}$. □

Lemma B.7. *Under the good event $\mathcal{E}$, if $\widehat{\pi}^*_k,\xi\in\Pi^{\mathscr{A}}_{k-1}$ and $\xi\notin\Pi^{\mathscr{A}}_{k}$ then ξ is suboptimal for some reward $\widehat{r}_{k'}\in\mathcal{R}_{\widehat{\mathfrak{P}}_{k'}}$ for all $k'\geqslant k$.*

Proof. Let us consider the following decomposition:

$$\begin{aligned}V^{\xi}_{\mathcal{M}'\cup \widehat{r}_{k'}}-V^{*}_{\mathcal{M}'\cup \widehat{r}_{k'}}&\leqslant V^{\xi}_{\mathcal{M}'\cup \widehat{r}_{k'}}-V^{\widehat{\pi}^*_k}_{\mathcal{M}'\cup \widehat{r}_{k'}}\\&=\underbrace{V^{\xi}_{\mathcal{M}'\cup \widehat{r}_{k'}}-V^{\xi}_{\mathcal{M}'\cup \widehat{r}_{k}}}_{\text{(a)}}+\underbrace{V^{\xi}_{\mathcal{M}'\cup \widehat{r}_{k}}-V^{\widehat{\pi}^*_k}_{\mathcal{M}'\cup \widehat{r}_{k}}}_{\text{(b)}}+\underbrace{V^{\widehat{\pi}^*_k}_{\mathcal{M}'\cup \widehat{r}_{k}}-V^{\widehat{\pi}^*_k}_{\mathcal{M}'\cup \widehat{r}_{k'}}}_{\text{(c)}},\end{aligned}$$

where the inequality follows from $V^{*}_{\mathcal{M}'\cup \widehat{r}_{k'}}\geqslant V^{\widehat{\pi}^*_k}_{\mathcal{M}'\cup \widehat{r}_{k'}}$. For (a) and (c) we apply Lemma B.6, having taken $\widehat{r}_{k'}\in\operatorname{arginf}_{\widehat{r}_{k'}\in\mathcal{R}_{\widehat{\mathfrak{P}}_k}}|\widehat{r}_{k'}-r|$

and $r \in \arg\inf_{r \in \mathcal{R}_{\mathfrak{P}}} |r - \widehat{r}_k|$ and we recall that $\xi, \widehat{\pi}^*_k \in \Pi^{\mathscr{A}}_{k-1}$ obtaining:

$$\sup_{\mu_0 \in \Delta^{\mathcal{S}}} \mu_0^{\intercal} \left(V^{\xi}_{\mathcal{M}' \cup \widehat{r}_{k'}} - V^{\xi}_{\mathcal{M}' \cup \widehat{r}_k} \right) \leqslant 2\epsilon^{\xi}_k \leqslant 2\epsilon_k,$$

$$\sup_{\mu_0 \in \Delta^{\mathcal{S}}} \mu_0^{\intercal} \left(V^{\widehat{\pi}^*_k}_{\mathcal{M}' \cup \widehat{r}_k} - V^{\widehat{\pi}^*_k}_{\mathcal{M}' \cup \widehat{r}_{k'}} \right) \leqslant 2\epsilon^{\widehat{\pi}^*_k}_k \leqslant 2\epsilon_k.$$

For (b),we first note that $V^{\widehat{\pi}^*_k}_{\mathcal{M}' \cup \widehat{r}_k} = V^*_{\mathcal{M}' \cup \widehat{r}_k}$ and then we observe that $\xi \notin \Pi^{\mathscr{A}}_k$ from which:

$$\sup_{\mu_0 \in \Delta^{\mathcal{S}}} \mu_0^{\intercal} \left(V^{\widehat{\pi}^*_k}_{\mathcal{M}' \cup \widehat{r}_k} - V^{\xi}_{\mathcal{M}' \cup \widehat{r}_k} \right) = \sup_{\mu_0 \in \Delta^{\mathcal{S}}} \mu_0^{\intercal} \left(V^*_{\mathcal{M}' \cup \widehat{r}_k} - V^{\xi}_{\mathcal{M}' \cup \widehat{r}_k} \right) > 4\epsilon_k.$$

Putting all together we have that $V^{\xi}_{\mathcal{M}' \cup \widehat{r}_{k'}} - V^*_{\mathcal{M}' \cup \widehat{r}_{k'}} < 0$ from which we conclude that ξ cannot be optimal for $k' \geqslant k$. □

Corollary B.2. *If $\epsilon_0 = \frac{1}{4(1-\gamma)}$, then for every $k \geqslant 0$, it holds that $\pi^*, \widehat{\pi}^*_{k+1} \in \Pi^{\mathscr{A}}_k$.*

Proof. We prove the result by induction on k. For $k=0$ we have that for every policy $\pi \in \Delta^{\mathcal{A}}_{\mathcal{S}}$ we have $\sup_{\mu_0 \in \Delta^{\mathcal{S}}} \mu_0^{\intercal} \left(V^*_{\mathcal{M}' \cup \widehat{r}_0} - V^{\pi}_{\mathcal{M}' \cup \widehat{r}_0} \right) \leqslant \frac{1}{1-\gamma} \leqslant 4\epsilon_0$. Thus, $\Pi^{\mathscr{A}}_0 = \Delta^{\mathcal{A}}_{\mathcal{S}}$, and in particular $\pi^*, \widehat{\pi}^*_1 \in \Pi^{\mathscr{A}}_0$. Suppose that for every $k' < k$ the statement hold, we prove it that it holds for k. Take $k' = k-1$, from the inductive hypothesis we have that $\pi^*, \widehat{\pi}^*_k \in \Pi^{\mathscr{A}}_{k-1}$. Then, from Lemma B.5 it holds that $\pi^* \in \Pi^{\mathscr{A}}_k$. By contradiction, suppose $\widehat{\pi}^*_{k+1} \notin \Pi^{\mathscr{A}}_k$. Then, let $j \leqslant k$ be the iteration such that $\widehat{\pi}^*_{k+1} \in \Pi^{\mathscr{A}}_{j-1}$ and $\widehat{\pi}^*_{k+1} \notin \Pi^{\mathscr{A}}_j$. From the inductive hypotesis, we have that $\widehat{\pi}^*_j \in \Pi^{\mathscr{A}}_{j-1}$. Thus, from Lemma B.7, it must be that $\widehat{\pi}^*_{k+1}$ is suboptimal for all $j' \geqslant j$, in particular for $j' = k+1$, that is a contradiction. □

Lemma B.8. *Under the good event, let $\widetilde{r} \in \arg\inf_{r \in \mathcal{R}_{\mathfrak{P}}} \|r - \widehat{r}_k\|_\infty$ where $\widehat{r}_k = \mathscr{A}(\mathcal{R}_{\widehat{\mathfrak{P}}_k})$, if $\pi \in \Pi^{\mathscr{A}}_k$ and $\pi^*, \widehat{\pi}^*_k \in \Pi^{\mathscr{A}}_{k-1}$ then $\sup_{\mu_0 \in \Delta^{\mathcal{S}}} \mu_0^{\intercal} \left(V^*_{\mathcal{M}' \cup \widetilde{r}} - V^{\pi}_{\mathcal{M}' \cup \widetilde{r}} \right) \leqslant 6\epsilon_k$.*

Proof. Let us consider the following derivation for $\widehat{r}_k \in \mathscr{A}(\mathcal{R}_{\widehat{\mathfrak{P}}_k})$ and $\widetilde{r} \in \arg\inf_{r \in \mathcal{R}_{\mathfrak{P}}} \|r - \widehat{r}_k\|_\infty$:

$$V^*_{\mathcal{M}' \cup \widetilde{r}} - V^{\pi}_{\mathcal{M}' \cup \widetilde{r}} = \underbrace{V^*_{\mathcal{M}' \cup \widetilde{r}} - V^*_{\mathcal{M}' \cup \widehat{r}_k}}_{\text{(a)}} + \underbrace{V^*_{\mathcal{M}' \cup \widehat{r}_k} - V^{\pi}_{\mathcal{M}' \cup \widehat{r}_k}}_{\text{(b)}} + \underbrace{V^{\pi}_{\mathcal{M}' \cup \widehat{r}_k} - V^{\pi}_{\mathcal{M}' \cup \widetilde{r}}}_{\text{(c)}}.$$

For (b), we have that $\sup_{\mu_0 \in \Delta^{\mathcal{S}}} \mu_0^{\intercal} \left(V^*_{\mathcal{M}' \cup \widehat{r}_k} - V^{\pi}_{\mathcal{M}' \cup \widehat{r}_k} \right) \leqslant 4\epsilon_k$ since $\pi \in \Pi^{\mathscr{A}}_k$. For (a), we have already proved in Equation (P.6) that $\sup_{\mu_0 \in \Delta^{\mathcal{S}}} \mu_0^{\intercal} \left(V^*_{\mathcal{M}' \cup \widetilde{r}} - V^*_{\mathcal{M}' \cup \widehat{r}_k} \right) \leqslant \epsilon_k$ recalling the definition of $\widetilde{r}$. For (c), by definition of ϵ^{π}_k and recalling the definition of $\widetilde{r}$ we have that $\sup_{\mu_0 \in \Delta^{\mathcal{S}}} \mu_0^{\intercal} \left(V^{\pi}_{\mathcal{M}' \cup \widehat{r}_k} - V^{\pi}_{\mathcal{M}' \cup \widetilde{r}} \right) \leqslant \epsilon^{\pi}_k \leqslant \epsilon_k$. □

B.4.2. Sample complexity

Theorem 6.1 (Sample Complexity of TRAVEL). *If Algorithm 2 stops at iteration K with accuracy ϵ_K and accuracy ϵ_{K-1} at the previous iteration, then with probability at least $1-\delta$ it fulfills Definition 5.1, for arbitrary target reward functions $\overline{r}$ and $\check{r}$, with a number of samples upper bounded by $n = \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} N_K(s,a)$ where:*

$$N_K(s,a) \leqslant \widetilde{\mathcal{O}} \left(\min \left\{ \frac{\gamma^2 R^2_{\max}}{(1-\gamma')^2 (1-\gamma)^2 \epsilon^2_K}, \frac{\gamma^2 R^2_{\max} \epsilon^2_{K-1}}{(1-\gamma)^2 (-A(s,a))^2 \epsilon^2_K} \right\} \right).$$

Proof. First of all, we observe that TRAVEL is optimizing a tighter bound (Corollary B.1 (ii)) compared to that of Uniform Sampling IRL (Corollary B.1 (i)). Thus, it follows that the sample complexity of Uniform Sampling IRL applies to TRAVEL too. We now turn to the problem-dependent analysis. Let us recall the definition of advantage function $A^*_{\mathcal{M}' \cup \widetilde{r}}(s,a) = Q^*_{\mathcal{M}' \cup \widetilde{r}}(s,a) - V^*_{\mathcal{M}' \cup \widetilde{r}}(s)$. We first derive a value of $N_K(s,a)$ so that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ it holds that:

$$\mathcal{C}_K(s,a) = \frac{R_{\max}}{1-\gamma} \left(\mathbb{1}\{N_k(s) = 0\} + \gamma \sqrt{\frac{2\ell_k(s,a)}{N^+_k(s,a)}} \right) \leqslant \frac{-A^*_{\mathcal{M}' \cup \widetilde{r}}(s,a) \epsilon_K}{6\epsilon_{K-1}}.$$

From which we obtain, having applied Lemma B.9:

$$\begin{aligned} N_K(s,a) &= \frac{72\gamma^2 R^2_{\max} \ell_K(s,a) \epsilon^2_{K-1}}{(1-\gamma)^2 (-A^*_{\mathcal{M}' \cup \widetilde{r}}(s,a))^2 \epsilon^2_K} \leqslant \frac{288\gamma^2 R^2_{\max} \epsilon^2_{K-1}}{(1-\gamma)^2 (-A^*_{\mathcal{M}' \cup \widetilde{r}}(s,a))^2 \epsilon^2_K} \log \left(\frac{144\gamma^2 R^2_{\max} \epsilon^2_{K-1}}{(1-\gamma)^2 (-A^*_{\mathcal{M}' \cup \widetilde{r}}(s,a))^2 \epsilon^2_K} \sqrt{\frac{12SA}{\delta}} \right) \\ &= \widetilde{\mathcal{O}} \left(\frac{\gamma^2 R^2_{\max} \epsilon^2_{K-1}}{(1-\gamma)^2 (-A^*_{\mathcal{M}' \cup \widetilde{r}}(s,a))^2 \epsilon^2_K} \right). \end{aligned}$$

From Lemma B.8, we know that $\sup_{\mu_0 \in \Delta^{\mathcal{S}}} \mu_0^\intercal \left(V^*_{\mathcal{M}' \cup \tilde{r}} - V^\pi_{\mathcal{M}' \cup \tilde{r}}\right) \leqslant 6\epsilon_K$. Thus, we have for every $\pi \in \Pi^{\mathscr{A}}_K$:

$$\begin{aligned}\left(I_{\mathcal{S}} - \gamma' \pi P'\right)^{-1} \pi \mathcal{C}_{K-1} &\leqslant -\frac{\epsilon_K}{6\epsilon_{K-1}} \left(I_{\mathcal{S}} - \gamma' \pi P'\right)^{-1} \pi A^*_{\mathcal{M}' \cup \tilde{r}} \\ &= \frac{\epsilon_K}{6\epsilon_{K-1}} \left(V^*_{\mathcal{M}' \cup \tilde{r}} - V^\pi_{\mathcal{M}' \cup \tilde{r}}\right) \leqslant \epsilon_K,\end{aligned}$$

where the equality follows from Lemma B.3. □

B.5. Technical Lemmas

Lemma B.9. *Let $a, b \geqslant 0$ such that $2a\sqrt{b} > e$. Then, the inequality $x \geqslant a \log(bx^2)$ is satisfied for all $x \geqslant -2aW_{-1}\left(-\frac{1}{2a\sqrt{b}}\right)$, where W_{-1} is the secondary component of the Lambert W function. Moreover, $-2aW_{-1}\left(-\frac{1}{2a\sqrt{b}}\right) \leqslant 4a \log(2a\sqrt{b})$.*

Proof. Let us consider the following derivation:

$$x \geqslant a \log(bx^2) \quad \Longrightarrow \quad e^x \geqslant b^a x^{2a} \quad \Longrightarrow \quad -\frac{x}{2a} e^{-\frac{x}{2a}} \geqslant -\frac{1}{2a\sqrt{b}}.$$

Now we apply the Lambert W function, under the assumption $-\frac{1}{2a\sqrt{b}} \leqslant -\frac{1}{e}$:

$$x \leqslant -2aW_0\left(-\frac{1}{2a\sqrt{b}}\right) \quad \text{or} \quad x \geqslant -2aW_{-1}\left(-\frac{1}{2a\sqrt{b}}\right),$$

where W_0 is the principal component of the Lambert W function. We consider only the second inequality. Now, we bound the Lambert W function starting from the inequality (Chatzigeorgiou, 2013): $W_{-1}(-e^{-u-1}) \geqslant -1 - \sqrt{2u} - u$. From which, we obtain:

$$-2aW_{-1}\left(-\frac{1}{2a\sqrt{b}}\right) \leqslant 2\sqrt{2}a\sqrt{\log(2a\sqrt{b}) - 1} + 2a\log(2a\sqrt{b}) \leqslant 4a\log(2a\sqrt{b}),$$

having bounded $\sqrt{x-1} \leqslant \frac{1}{2}x$. □

C. Related Works

The IRL problem was introduced by Ng & Russell (2000). Most early IRL algorithms assume that the dynamics of the system are known. Many criteria were proposed for selecting a *good* reward function in the feasible reward set, based on features matching (Abbeel & Ng, 2004), maximum margin (Ratliff et al., 2006a), maximum entropy (Ziebart et al., 2008; 2010), Bayesian framework (Ramachandran & Amir, 2007), boosting methods (Ratliff et al., 2006b) and Gaussian processes (Levine et al., 2011). A limited number of IRL algorithms can be considered model-free: Relative Entropy Inverse Reinforcement Learning (Boularias et al., 2011), Generative Adversarial Imitation Learning (Ho & Ermon, 2016), Gradient-based Inverse Reinforcement Learning (Pirotta & Restelli, 2016) and its extensions (Metelli et al., 2017; 2020; Ramponi et al., 2020b;a). Other works on Imitation Learning use an active approach, such as the one used in this paper. Judah et al. (2012) draw a reduction from active imitation learning to i.i.d. active learning. In (Ross & Bagnell, 2010) and (Ross et al., 2011), the authors propose two approaches based on executing the estimated policy and asking an oracle for a dataset containing the action performed by the expert. In these papers, however, no guarantees on the sample complexity are provided. The closest work to ours is by Lopes et al. (2009), which propose a method to actively ask for samples from a generator to perform IRL, adopting a Bayesian approach. However, they assume knowledge of the real transition model and the main effort lies in estimating the expert's policy. Since we do not assume the knowledge of the transition model, this work is not fully comparable to our setting.

D. Efficient Implementation

In this appendix, we provide a formulation of the optimization problem in Equation (2) that is convex. We follow an implementation similar to (Zanette et al., 2019). As first step, we move the optimization from the policies π on the visitation distribution μ. This allow, as in (Zanette et al., 2019), to put the inner maximization into a matrix form:

$$\max_{\boldsymbol{x}} \begin{bmatrix} \mathcal{C}_{k+1} \\ \mathbf{0}_{\mathcal{S}\times\mathcal{A}} \\ \mathbf{0}_{\mathcal{S}} \end{bmatrix}^\intercal \boldsymbol{x}$$

$$\begin{aligned}\text{s.t.} & \begin{bmatrix} E^\intercal - \gamma'(P')^\intercal & -I_{\mathcal{S}} & \mathbf{0}_{\mathcal{S}} \\ \mathbf{0}_{\mathcal{S}\times\mathcal{A}}{}^\intercal & \mathbf{1}_{\mathcal{S}}{}^\intercal & 0 \\ \widehat{r}_k^\intercal & -(EV^*_{\mathcal{M}'\cup\widehat{r}_k})^\intercal & -1 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \mathbf{0}_{\mathcal{S}} \\ 1 \\ -4\epsilon_k \end{bmatrix} \\ & \boldsymbol{x} \geqslant 0\end{aligned}$$

where $\boldsymbol{x} = \begin{bmatrix} \mu \\ \mu_0 \\ t \end{bmatrix} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}+\mathcal{S}+1}$ and $\mathcal{C}_{k+1}$ is equal to:

$$\mathcal{C}_{k+1}(s,a) = \frac{R_{\max}}{1-\gamma}\left(\mathbb{1}\{N_{k+1}(s)=0\} + \sqrt{\frac{2\ell_k(s,a)}{N^+_{k+1}(s,a)}}\right)$$

Now we formulate the dual of this linear program:

$$\begin{aligned}\min_{\boldsymbol{y}} & \begin{bmatrix} \mathbf{0}_{\mathcal{S}} \\ 1 \\ -4\epsilon_k \end{bmatrix}^T \boldsymbol{y} \\ \text{s.t.} & \begin{bmatrix} E^\intercal - \gamma'(P')^\intercal & -I_{\mathcal{S}} & \mathbf{0}_{\mathcal{S}} \\ \mathbf{0}_{\mathcal{S}\times\mathcal{A}}{}^\intercal & \mathbf{1}_{\mathcal{S}}{}^\intercal & 0 \\ \widehat{r}_k^\intercal & -(EV^*_{\mathcal{M}'\cup\widehat{r}_k})^\intercal & -1 \end{bmatrix}^\intercal \boldsymbol{y} \geqslant \begin{bmatrix} \mathcal{C}_{k+1} \\ \mathbf{0}_{\mathcal{S}\times\mathcal{A}} \\ \mathbf{0}_{\mathcal{S}} \end{bmatrix}\end{aligned}$$

From that we can formulate the minimax optimization problem in Equation (2) as a convex minimization problem, recalling that $N_{k+1}(s,a) = N_k(s,a) + n_{k+1}(s,a)$:

$$\begin{aligned}\min_{n_{k+1},\boldsymbol{y}} & \begin{bmatrix} \mathbf{0}_{\mathcal{S}} \\ 1 \\ -4\epsilon_k \end{bmatrix}^T \boldsymbol{y} \\ \text{s.t.} & \begin{bmatrix} E^\intercal - \gamma'(P')^\intercal & -I_{\mathcal{S}} & \mathbf{0}_{\mathcal{S}} \\ \mathbf{0}_{\mathcal{S}\times\mathcal{A}}{}^\intercal & \mathbf{1}_{\mathcal{S}}{}^\intercal & 0 \\ \widehat{r}_k^\intercal & -(EV^*_{\mathcal{M}'\cup\widehat{r}_k})^\intercal & -1 \end{bmatrix}^\intercal \boldsymbol{y} \geqslant \begin{bmatrix} \mathcal{C}_{k+1} \\ \mathbf{0}_{\mathcal{S}\times\mathcal{A}} \\ \mathbf{0}_{\mathcal{S}} \end{bmatrix} \\ & n_{k+1}(s,a) \geqslant 0 \\ & \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} n_{k+1}(s,a) \leqslant n_{\max}\end{aligned}$$

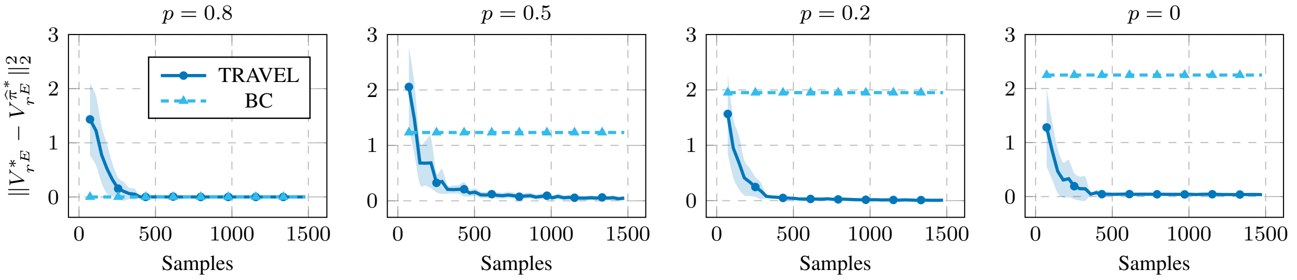


Figure 2. Comparison between TRAVEL and Behavioral Cloning (BC) on Gridworld environment, with different values of obstacle's probability for the target MDP $\mathcal{M}'$. 200 runs, 98% c.i.

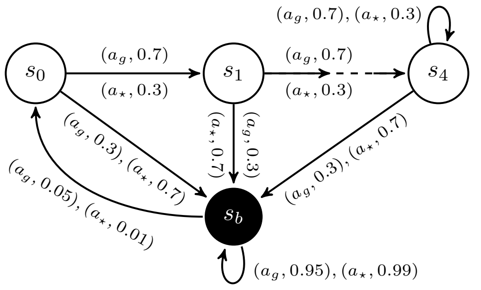

s_0
s_1
s_4
s_b
$(a_g, 0.7)$
$(a_\star, 0.3)$
$(a_g, 0.7)$
$(a_\star, 0.3)$
$(a_g, 0.7), (a_\star, 0.3)$
$(a_g, 0.3), (a_\star, 0.7)$
$(a_\star, 0.7)$
$(a_g, 0.3)$
$(a_g, 0.3), (a_\star, 0.7)$
$(a_g, 0.05), (a_\star, 0.01)$
$(a_g, 0.95), (a_\star, 0.99)$

Figure 3. MDP employed in Section E.2. States s_2 and s_3 behave exactly as s_1. $a_\star$ denotes any action in $\{a_1,...,a_9\}$.

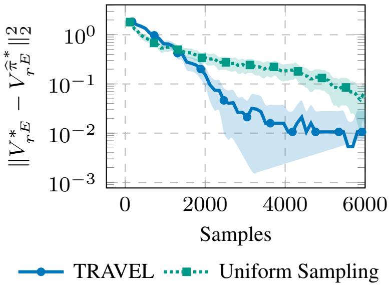


Figure 4. Comparison between Uniform Sampling IRL and TRAVEL. 300 runs, 98% c.i.

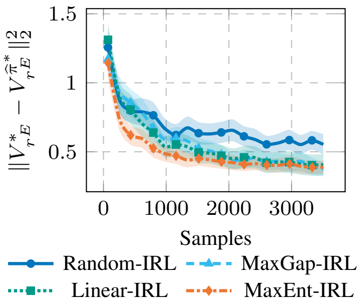


Figure 5. TRAVEL using different IRL algorithms on random MDPs. 200 runs, 98% c.i.

E. Experiment

In this section, we provide the experimental evaluation of TRAVEL with a threefold goal. In the first experiment, we motivate the need for employing IRL over BC when our goal is to transfer knowledge to a target environment (Section E.1). Then, we highlight the benefits of the sampling strategy of TRAVEL over Uniform Sampling (Section E.2). Finally, we show how TRAVEL can be combined with different IRL algorithms (Section E.3). In all the experiments, we employ reward functions that depend on the state only and the algorithms are evaluated according to the following performance index $\|V^*_{\mathcal{M}'\cup r^E} - V^{\hat{\pi}^*}_{\mathcal{M}'\cup r^E}\|_2^2$, where the symbols are defined in the previous sections (we omit the subscript $\mathcal{M}'\cup$ in the plots).

E.1. IRL vs Behavioral Cloning

In this section, we show the benefits of IRL over BC when we want to learn a transferable reward function. In BC, the collected samples are used to estimate a policy describing the expert's behavior directly. The recovered policy is typically highly dependent on the environment in which the expert is acting, therefore, in many cases, it cannot be transferred to different environments. We use a 3×3 Gridworld environment with an obstacle in the central cell that makes the agent bouncing back with probability p and surpassing it with probability $1-p$. If $p\simeq 0$ the optimal policy is to collide with the obstacle until the agent reaches the goal state. While, if $p\simeq 1$, an optimal agent gets around the obstacle. The source MDP has obstacle's probability $p=0.8$ and target MDPs are four Gridworlds with obstacle's probabilities $p\in\{0,0.2,0.5,0.8\}$. For this experiment, we use a simple IRL algorithm that enforces the conditions of Lemma 3.1 and chooses the reward function to maximize the minimum action gap; we call it *MaxGap-IRL*. The results in Figure 2 show that the performance of BC deteriorates as the source and target MDPs become more dissimilar, as expected. Differently, TRAVEL combined with MaxGap-IRL allows recovering a reward function that leads to an optimal policy. Thus, as long as the target and source environments are the same (Figure 2 first plot) BC is a valid alternative, but IRL becomes unavoidable when the need for transferring knowledge arises.

E.2. TRAVEL vs Uniform Sampling

As discussed in Section 5.2, Uniform Sampling IRL and TRAVEL differ from the strategy used to allocate samples to the state-action pairs. While Uniform Sampling queries the generative model uniformly, TRAVEL actively allocates samples in the state-action pairs that will carry “more information”. We consider a chain MDP composed by 6 states $\mathcal{S}=\{s_0,\ldots,s_4,s_b\}$ and 10 actions $\mathcal{A}=\{a_g,a_1,\ldots,a_9\}$ (Figure 3). We have tested both algorithms and the results are shown in Figure 4. Although Uniform Sampling IRL seems to perform better with a small number of samples, we observe that TRAVEL recovers a reward function that allows achieving a near-optimal performance in less than half of the samples needed by Uniform Sampling.

E.3. TRAVEL with Different IRL Algorithms

In this section, we show the performance of TRAVEL paired with different IRL algorithms: MaxGap-IRL, MaxEnt-IRL (Ziebart et al., 2008), Linear-IRL (Abbeel & Ng, 2004), and *Random-IRL*. The Random IRL selects a random reward function from the estimated feasible set of rewards. We compare the algorithms on 200 random generated MDPs. The results in Figure 5 show that MaxGap-IRL, Linear-IRL, and MaxEnt-IRL display a faster convergence rate than Random. This is the expected behavior as these IRL algorithms choose the reward function in the feasible set with a meaningful criterion. However, the curve of Random-IRL shows an improvement, proving that the feasible set shrinks, but it struggles harder to reach a near-zero error as it likely selects less discriminating rewards. This underlines how a reasonable choice of the reward function within the feasible set can have a positive impact on performance.