Inverse Reinforcement Learning with Sub-optimal Experts

Anonymous Author
Anonymous Institution

Abstract

Inverse Reinforcement Learning (IRL) techniques deal with the problem of deducing a reward function that explains the behavior of an expert agent who is assumed to act optimally in an underlying unknown task. In several problems of interest, however, it is possible to observe the behavior of multiple experts with different degree of optimality (e.g., racing drivers whose skills ranges from amateurs to professionals). For this reason, in this work, we extend the IRL formulation to problems where, in addition to demonstrations from the optimal agent, we can observe the behavior of multiple sub-optimal experts. Given this problem, we first study the theoretical properties of the class of reward functions that are compatible with a given set of experts, i.e., the *feasible reward set*. Our results show that the presence of multiple sub-optimal experts can significantly shrink the set of compatible rewards. Furthermore, we study the statistical complexity of estimating the feasible reward set with a generative model. To this end, we analyze a uniform sampling algorithm that results in being minimax optimal whenever the sub-optimal experts' performance level is sufficiently close to the one of the optimal agent.

1 INTRODUCTION

Inverse Reinforcement Learning (IRL, Ng et al., 2000) deals with the problem of recovering a reward function that explains the behavior of an expert agent who is assumed to act optimally in an underlying unknown task. Over the years, the IRL problem has consistently captured the attention of the research community (see, for instance, Arora and Doshi (2021) and Adams et al. (2022) for in-depth surveys). Indeed, this general scenario, where the reward function needs to be learned, emerges in numerous real-world applications. A prime example of this arises from human-in-the-loop settings (Mosqueira-Rey et al., 2023), where the expert is a human solving a task, and an explicit specification of the human's goal in the form of a reward function is often unavailable. Notably, humans encounter difficulty in expressing their intentions in the form of an underlying reward signal, preferring instead to demonstrate what they perceive as the correct behavior. Once we retrieve a reward function, (i) we obtain explicit information for understanding the expert's choices, and, furthermore, (ii) we can utilize it to train reinforcement learning agents, even under shifts in the features of the underlying system.

Since the seminal work of Ng et al. (2000), IRL has emerged as a significantly complex task. Indeed, one of its primary challenges lies in the intrinsic *ill-posed* nature of the problem, as multiple reward functions that are compatible with the expert's behavior exist. Recently, a promising avenue of research (Metelli et al., 2021; Lindner et al., 2022; Metelli et al., 2023) has tackled this ambiguity issue from an intriguing perspective. Specifically, this strand of works focuses on estimating *all* the reward functions that are compatible with the observed demonstration, thereby postponing the selection of the reward function and directing their focus solely on the expert's intentions.

Nevertheless, these approaches fall short in modeling more complex situations that arise in the real world. Indeed, in several problems of interest, it is possible to observe the behavior of multiple agents with different degrees of expertise. As an illustrative example, we can consider the human-in-the-loop settings mentioned above. Imagine, indeed, that we are interested in recovering reward functions that explain the intent behind racing drivers. In this scenario, racing car companies typically have access to a variety of drivers with diverse skills, including professionals, test drivers, and emerging talents from developmental programs. In

Preliminary work. Under review by AISTATS 2024. Do not distribute.

this context, while the focus is typically on the reward function of professional drivers, we expect a proficient IRL method to effectively leverage demonstrations and information provided by drivers with lower expertise. Indeed, from an intuitive perspective, if we have information on the degree of expertise of other drivers, we can expect that, by exploiting their demonstrations, we can reduce the inherent ambiguity of IRL problems. For this reason, in this work, we extend the IRL formulation to settings where, in addition to demonstrations from an optimal agent, we can observe the behavior of multiple sub-optimal experts, of which we know an index of their sub-optimality.

More specifically, we will be primarily focused in answering the following theoretical questions:

(Q1) How does the presence of sub-optimal experts affects the class of reward functions that are compatible with the observed behavior? Can they limit the intrinsic ambiguity that affects IRL problems?

(Q2) What is the statistical complexity of estimating the set of reward functions that are compatible with a given set of experts? How does it compare against the one of single-experts IRL problems?

Contributions and Outline After providing the necessary notation and background, we introduce the novel problem of Inverse Reinforcement Learning with multiple and sub-optimal experts (Section 2). We then proceed by studying the *theoretical properties* of the class of reward functions that are compatible with a given set of experts under the assumption that an upper bound on the performance between a sub-optimal agent and the optimal expert is available to the designer of the IRL system (Section 3). More precisely, our findings indicate that having multiple sub-optimal experts can significantly shrink the set of compatible rewards, thereby *limiting* the ambiguity issue that affects the IRL problem. Leveraging our previous results, we continue by studying the *statistical complexity* of estimating the feasible reward set with a generative model (Section 4). To this end, after formally introducing a Probabilistic Approximately Correct (PAC, Even-Dar et al., 2002) framework, we derive a novel lower bound on the number of samples that are required to obtain an accurate estimate of the feasible reward set. Then, we present a uniform sampling algorithm and analyze its theoretical guarantees. Our results show that (i) the IRL problem with sub-optimal experts is statistically harder than the single agent IRL setting, and (ii) that the uniform sampling algorithm is minimax optimal whenever the sub-optimal experts' performance level is sufficiently close to the one of the optimal agent. Finally, we conclude with a discussion on existing works (Section 5) and by highlighting potential avenues for future research (Section 6).

2 PRELIMINARIES

In this section, we provide the notation and essential concepts employed throughout this document.

Notation Consider a finite set $\mathcal{X}$, we denote with $\Delta^{\mathcal{X}}$ the set of probability measures over $\mathcal{X}$. Let $\mathcal{Y}$ be a set, we denote with $\Delta_{\mathcal{Y}}^{\mathcal{X}}$ the set of functions $f : \mathcal{Y} \to \Delta^{\mathcal{X}}$. Given $f \in \mathbb{R}^n$, we denote with $\|f\|_\infty$ the infinite norm of f. Let $\mathcal{X}$ and $\mathcal{X}'$ be two non-empty subsets of a metric space $(\mathcal{Y}, d)$, we define the Hausdorff distance (Rockafellar and Wets, 2009) between $\mathcal{X}$ and $\mathcal{X}'$ as:

$$H_d(\mathcal{X}, \mathcal{X}') = \max\left\{\sup_{x\in\mathcal{X}}\inf_{x'\in\mathcal{X}'} d(x,x'), \sup_{x'\in\mathcal{X}'}\inf_{x\in\mathcal{X}} d(x,x')\right\}.$$

Notice that the Hausdorff distance is directly dependent on the metric d. Finally, given an integer $x \in \mathbb{N}_{>0}$, we denote with $\mathbf{1}_x$ the x-dimensional vector given by $(1, \dots, 1)^\top$.

Markov Decision Processes A Markov Decision Process *without a reward function* (MDP\R) is defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, \gamma)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $p \in \Delta_{\mathcal{S}\times\mathcal{A}}^{\mathcal{S}}$ denotes the transition probability kernel, and $\gamma \in [0, 1)$ is the discount factor. In this paper, we consider finite state and action spaces, namely $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$. A Markov Decision Process (MDP, Puterman, 2014) is obtained by combining an MDP\R $\mathcal{M}$ with a reward function $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. Without loss of generality, we assume reward functions bounded in $[0, 1]$. We denote with $\mathcal{M} \cup r$ the resulting MDP. The behavior of an agent is described by a policy $\pi \in \Delta_{\mathcal{S}}^{\mathcal{A}}$, that, for each state, prescribes a probability distribution over actions.

Operators Consider $f \in \mathbb{R}^{\mathcal{S}}$ and $g \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. We denote with P and π the operators that are induced by the transition model p and the policy π respectively. More specifically, $Pf(s,a) = \sum_{s'\in\mathcal{S}} p(s'|s,a)f(s')$, and $\pi g(s) = \sum_{a\in\mathcal{A}} \pi(a|s)g(s,a)$. Moreover, we introduce the operators E and $\bar{B}^\pi$ defined in the following way: $Ef(s,a) = f(s)$ and $\left(\bar{B}^\pi g\right)(s,a) = \mathbb{1}\{\pi(a|s) = 0\} g(s,a)$. Finally, we define $d^\pi f$ as the expectation of f under the discounted occupancy measure. More formally $d^\pi f = (I_{\mathcal{S}} - \gamma\pi P)^{-1} f = \sum_{t=0}^{+\infty} (\gamma\pi P)^t f$.

Value Functions and Optimality Given an MDP $\mathcal{M} \cup r$ and a policy π, the *Q-function* $Q^\pi_{\mathcal{M}\cup r}(\cdot)$ represents the expected discounted sum of rewards collected

in $\mathcal{M} \cup r$ starting from (s, a) and following policy π. More formally:

$$Q^{\pi}_{\mathcal{M}\cup r}(s,a) = \mathbb{E}\left[\sum_{t=0}^{+\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a\right],$$

where the expectation is taken w.r.t. the stochasticity of the policy and the environment, that is $s_{t+1} \sim p(\cdot|s_t, a_t)$ and $a_t \sim \pi(\cdot|s_t)$. Similarly, the *V-function* $V^{\pi}_{\mathcal{M}\cup r}$ represents the expectation of the Q-function over the action space, namely $V^{\pi}_{\mathcal{M}\cup r} = \pi Q^{\pi}_{\mathcal{M}\cup r}$. The *advantage function* $A^{\pi}_{\mathcal{M}\cup r} = Q^{\pi}_{\mathcal{M}\cup r} - EV^{\pi}_{\mathcal{M}\cup r}$ represents the immediate gain of taking a given action, rather than following policy π. A policy π^* is optimal if it has non-positive advantage in each-state action pair; namely $A^{\pi^*}_{\mathcal{M}\cup r} \leq 0$ holds element-wise.

Inverse Reinforcement Learning An Inverse Reinforcement Learning (IRL, Ng et al., 2000) problem is defined as a tuple $\mathfrak{B} = (\mathcal{M}, \pi_E)$, where $\mathcal{M}$ is an MDP\R and $\pi_E \in \Delta^{\mathcal{A}}_{\mathcal{S}}$ is an expert policy. Given a reward function $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, we say that r is *feasible* for $\mathfrak{B}$ if it is compatible with the behavior of the expert, namely π_E is an optimal policy for the MDP $\mathcal{M} \cup r$. We denote with $\mathcal{R}_{\mathfrak{B}}$ the set of feasible reward functions, namely:

$$\mathcal{R}_{\mathfrak{B}} = \left\{r \in [0,1]^{\mathcal{S}\times\mathcal{A}} : A^{\pi_E}_{\mathcal{M}\cup r} \leq 0\right\}. \tag{1}$$

The set $\mathcal{R}_{\mathfrak{B}}$ takes the name of *feasible reward set* (Metelli et al., 2021; Lindner et al., 2022; Metelli et al., 2023). To characterize the set $\mathcal{R}_{\mathfrak{B}}$, Metelli et al. (2021) have shown that a reward function r belongs to $\mathcal{R}_{\mathcal{B}}$ if and only if there exists $\zeta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\geq 0}$ and $V \in \mathbb{R}^{\mathcal{S}}$ such that:

$$r = -\bar{B}^{\pi_E}\zeta + (E - \gamma P)V. \tag{2}$$

In other words, each reward function in $\mathcal{R}_{\mathfrak{B}}$, is expressed as a sum of two components. The first one, $-\bar{B}^{\pi_E}\zeta$, which is non-zero only when $\pi_E(a|s) = 0$, can be interpreted as the advantage function $A^{\pi_E}_{\mathcal{M}\cup r}$. The second one, $(E - \gamma P)V$, instead, can be interpreted as a reward-shaping via function V, which is widely recognized to maintain the optimality of the expert's policy (Ng et al., 2000). Given this interpretation, it follows that $\|V\|_\infty \leq (1-\gamma)^{-1}$ and $\|\zeta\|_\infty \leq (1-\gamma)^{-1}$.

IRL with Sub-optimal Experts In this work, we extend the IRL formulation to problems where, in addition to demonstrations from an optimal expert, we can observe the behaviors of multiple and sub-optimal agents. More precisely, we define an Inverse Reinforcement Learning problem with multiple and Sub-optimal Experts (IRL-SE) as a tuple $\bar{\mathfrak{B}} = \left(\mathcal{M}, \pi_{E_1}, (\pi_{E_i})_{i=2}^{n+1}, (\xi_i)_{i=2}^{n+1}\right)$, where $\mathcal{M}$ is an MDP\R, π_{E_1} is the policy of an optimal agent, and $(\pi_{E_i})_{i=2}^{n}$ are a collection of n sub-optimal policies with known degree of sub-optimality $\xi_i \in \mathbb{R}_{>0}$.[1] A reward function $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, is feasible for $\bar{\mathfrak{B}}$ if π_{E_1} is an optimal policy for the MDP $\mathcal{M} \cup r$ and, furthermore, if:

$$\left\|V^{\pi_{E_1}}_{\mathcal{M}\cup r} - V^{\pi_{E_i}}_{\mathcal{M}\cup r}\right\|_\infty \leq \xi_i, \tag{3}$$

holds for all $i \in \{2, \dots, n+1\}$. In this sense, ξ_i (i.e., the degree of sub-optimality of policy π_{E_i}) represents a known upper bound on the performance between the optimal expert and the i-th sub-optimal agent. We denote by $\mathcal{R}_{\bar{\mathfrak{B}}}$ the set of feasible rewards for $\bar{\mathfrak{B}}$. More formally, $r \in [0,1]^{\mathcal{S}\times\mathcal{A}}$ belongs to $\mathcal{R}_{\bar{\mathfrak{B}}}$ if (i) $A^{\pi_{E_1}}_{\mathcal{M}\cup r} \leq 0$ and (ii) Equation (3) holds for all $i \in \{2, \dots, n+1\}$. Notice that, whenever no sub-optimal expert is present, we directly recover the definition of the feasible set for single-agent IRL problems, i.e., $\mathcal{R}_{\mathfrak{B}}$ in Equation (1).

Empirical Estimates Let $\mathcal{D}_t$ be a dataset of transitions of t tuples $\mathcal{D}_t = \left\{\left(s_j, a_j, s'_j, (a^{(i)}_j)_{i=1}^{n+1}\right)\right\}_{j=1}^{t}$, where $s'_j \sim p(\cdot|s_j, a_j)$, and $a^{(i)}_j \sim \pi_{E_i}(\cdot|s_j)$. Given $\mathcal{D}_t$, it is possible to define the empirical transition model $\hat{p}$ and the empirical experts' policy $\hat{\pi}_{E_i}$ as follows:

$$\begin{aligned}
\hat{p}(s'|s,a) &= \begin{cases} \frac{N_t(s,a,s')}{N_t(s,a)} & \text{if } N_t(s,a) > 0 \\ \frac{1}{S} & \text{otherwise} \end{cases}, \\
\hat{\pi}_{E_i}(a|s) &= \begin{cases} \frac{N^{(i)}_t(s,a)}{N_t(s)} & \text{if } N_t(s) > 0 \\ \frac{1}{A} & \text{otherwise} \end{cases},
\end{aligned} \tag{4}$$

where $N_t(s,a,s')$ denotes the number of times in which (s_j, a_j, s'_j) is equal to (s,a,s'), $N_t(s,a) = \sum_{s'} N_t(s,a,s')$, $N_t(s) = \sum_{a,s'} N_t(s,a,s')$, and, finally, $N^{(i)}_t(s,a)$ counts the number of times in which $(s_j, a^{(i)}_j)$ is equal to (s,a). Given these definitions, we denote with $\hat{\bar{\mathfrak{B}}}_t$ the empirical IRL problem that is induced by $\hat{p}$ and $\{\hat{\pi}_{E_i}\}_{i=1}^{n+1}$. We denote with $\mathcal{R}_{\hat{\bar{\mathfrak{B}}}_t}$ its corresponding feasible reward region.

3 SUB-OPTIMAL EXPERTS AND THE FEASIBLE REWARD SET

In this section, we lay down the foundations for the problem of Inverse Reinforcement Learning in the presence of multiple and sub-optimal experts. Specifically, given the formulation introduced in Section 2, we now delve into an in-depth examination of the theoretical properties of the feasible reward set $\mathcal{R}_{\bar{\mathfrak{B}}}$. We will tackle the problem from two different perspectives.

[1] For the sake of exposition, we consider a single optimal agent. The extension to cases where multiple optimal policies are available is direct. Futher details on this point are provided in Appendix A.

First, we present an implicit formulation of $\mathcal{R}_{\bar{\mathfrak{B}}}$ that will allow us to characterize the properties of the feasible set by means of Q and V function (Section 3.1). Then, we will present an explicit formulation that will provide us with a precise mathematical description of $\mathcal{R}_{\bar{\mathfrak{B}}}$ (Section 3.2). As we shall see, these results indicate that the presence of sub-optimal experts can significantly shrink the feasible set of compatible rewards.

3.1 Implicit Formulation of $\mathcal{R}_{\bar{\mathfrak{B}}}$

As mentioned above, we begin by providing an implicit description of the feasible reward set $\mathcal{R}_{\bar{\mathfrak{B}}}$. To this end, we derive the following result (proof in Appendix B).

Lemma 1. *Let $\bar{\mathfrak{B}}$ be an IRL problem with sub-optimal experts. Let $r \in [0,1]^{\mathcal{S}\times\mathcal{A}}$. Then, $r \in \mathcal{R}_{\bar{\mathfrak{B}}}$ if and only if the following conditions are satisfied:*

$$\textit{(i)}\ Q^{\pi_{E_1}}_{\mathcal{M}\cup r}(s,a) = V^{\pi_{E_1}}_{\mathcal{M}\cup r}(s) \qquad \forall (s,a) : \pi_{E_1}(a|s) > 0$$

$$\textit{(ii)}\ Q^{\pi_{E_1}}_{\mathcal{M}\cup r}(s,a) \leq V^{\pi_{E_1}}_{\mathcal{M}\cup r}(s) \qquad \forall (s,a) : \pi_{E_1}(a|s) = 0$$

$$\textit{(iii)}\ V^{\pi_{E_1}}_{\mathcal{M}\cup r} \leq V^{\pi_{E_i}}_{\mathcal{M}\cup r} + \mathbf{1}_S \xi_i \qquad \forall i \in \{2,\dots,n+1\}.$$

Lemma 1 provides necessary and sufficient conditions for determining whether a reward function r belongs to the feasible set $\mathcal{R}_{\bar{\mathfrak{B}}}$. More precisely, condition (i) and (ii) directly encodes the optimality of policy π_{E_1} for $\mathcal{M}\cup r$, i.e., the advantage function $A^{\pi_{E_1}}_{\mathcal{M}\cup r}$ is non-positive in each state-action pair. Condition (iii), on the other hand, arises from the presence of sub-optimal experts, and it is directly related to Equation (3).

At this point, by closely examining Lemma 1, it is possible to gain insight into the limitations and advantages associated with the additional presence of multiple and sub-optimal experts. Consider, indeed, the following illustrative examples.

Example 1. Suppose that $\pi_{E_i} = \pi_{E_1}$ holds for all $i \in \{2,\dots,n+1\}$. In this case, condition (iii) is clearly satisfied for any reward function r. It follows that the feasible reward set $\mathcal{R}_{\bar{\mathfrak{B}}}$ is purely determined by the requirement that the advantage function of π_{E_1} is non-negative, and, as a consequence, the set $\mathcal{R}_{\bar{\mathfrak{B}}}$ coincides with the one of the single-expert IRL problem, namely $\mathcal{R}_{\bar{\mathfrak{B}}} = \mathcal{R}_{\mathfrak{B}}$. Analogously, if $\xi_i \geq (1-\gamma)^{-1}$ holds for all sub-optimal experts, condition (iii) is vacuous, and, similarly to the previous case, $\mathcal{R}_{\bar{\mathfrak{B}}}$ reduces to $\mathcal{R}_{\mathfrak{B}}$.

Example 2. Consider the MDP with 2 states depicted in Figure 1, and suppose, for the sake of exposition, that only one additional sub-optimal expert is present. In this case, the optimal agent and the sub-optimal agent follows completely different policies in S_0. By developing the conditions in Lemma 1, it is easy to see that, in addition to the constraint that $r(S_0,A_1) \geq r(S_0,A_2)$ (i.e., π_{E_1} is an optimal policy), condition (iii) introduces a further relationship between $r(S_0,A_1)$ and $r(S_0,A_2)$, that is $r(S_0,A_1) - r(S_0,A_2) \leq \xi_i$. In this sense, if ξ_i is sufficiently small (i.e., $\xi_i < 1$ in this case), the presence of the sub-optimal agents can significantly reduce $\mathcal{R}_{\bar{\mathfrak{B}}}$ compared to $\mathcal{R}_{\mathfrak{B}}$.

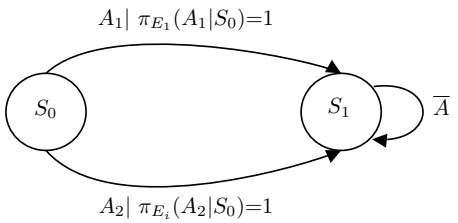

Figure 1: MDP example, with 2 states and 2 experts, that highlights the benefits of sub-optimal agents (Example 2). In S_1 both π_{E_1} and π_{E_i} are identical, i.e., $\pi_{E_1}(\bar{A}|S_1) = \pi_{E_i}(\bar{A}|S_1) = 1$.

Abstracting away from the previous examples, we can notice that whenever (a) the sub-optimal agents exhibit significant differences in behavior from the optimal expert and (b) their performance level is sufficiently close to being optimal, $\mathcal{R}_{\bar{\mathfrak{B}}}$ can notably shrink compared to $\mathcal{R}_{\mathfrak{B}}$. In the next section, through the explicit formulation of the feasible reward set, we will analyze this phenomenon quantitatively and in more detail.

3.2 Explicit Formulation of $\mathcal{R}_{\bar{\mathfrak{B}}}$

We now continue by providing an explicit formulation of the feasible set $\mathcal{R}_{\bar{\mathfrak{B}}}$. The following result (proof in Appendix B) summarizes our findings.

Theorem 3. *Let $\bar{\mathfrak{B}}$ be an IRL problem with sub-optimal experts. Let $r \in [0,1]^{\mathcal{S}\times\mathcal{A}}$. Then, $r \in \mathcal{R}_{\bar{\mathfrak{B}}}$ if and only if there exists $\zeta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\geq 0}$ and $V \in \mathbb{R}^{\mathcal{S}}$ such that the following conditions are satisfied:*

$$r = -\bar{B}^{\pi_{E_1}}\zeta + (E - \gamma P)V, \tag{5}$$

and, for all $i \in \{2,\dots,n+1\}$:

$$d^{\pi_{E_i}}\pi_{E_i}\bar{B}^{\pi_{E_1}}\zeta \leq \mathbf{1}_S \xi_i. \tag{6}$$

Theorem 3 deserves some comments. First of all, from Equation (5), we can see that a necessary condition for having $r \in \mathcal{R}_{\bar{\mathfrak{B}}}$ is that it can be expressed as the sum of two different components, namely $-\bar{B}^{\pi_{E_1}}\zeta$ and $(E-\gamma P)V$. This sort of result is a direct consequence of the fact that π_{E_1} is an optimal policy for $\mathcal{M}\cup r$, and, in this sense, it recovers the existing results of single expert IRL settings Metelli et al. (2021). Indeed, we

notice that it exactly matches Equation (2), and, consequently, it does not depend at all on the presence of the sub-optimal experts.[2] The role of the sub-optimal agents, on the other hand, is completely expressed by Equation (6).[3] More precisely, each additional expert introduces a set of *linear* constraints on the values that ζ can assume.[4] We recall that $-\bar{B}^{\pi_{E_1}}\zeta$ can be interpreted as the advantage function for the optimal policy π_{E_1}. In this sense, Equation (6) limits how sub-optimal the values of actions not played by π_{E_1} can be. Specifically, we notice that the resulting Q function of the optimal expert π_{E_1}, for a given choice of r, can be expressed as $Q^{\pi_{E_1}}_{\mathcal{M}\cup r} = -\bar{B}^{\pi_{E_1}}\zeta + EV$ (Metelli et al., 2021). In this sense, we can appreciate that by limiting the values of ζ, we are restricting the sub-optimality gaps, expressed in terms of Q functions, of actions that the optimal expert does not play. At this point, we notice that the linear constraints in Equation (6) are expressed in terms of $\pi_{E_i}\bar{B}^{\pi_{E_1}}\zeta$. As a consequence, they will only affect state-action pairs (s, a) that are played by the sub-optimal experts (i.e., $\pi_{E_i}(a|s) > 0$) and that are not played by the optimal agent (i.e., $\pi_{E_1}(a|s) = 0$). Therefore, as previously highlighted with the implicit formulation of $\mathcal{R}_{\bar{\mathfrak{B}}}$, a sub-optimal expert π_{E_i} should behave differently w.r.t. the optimal agent π_{E_1} in order to provide meaningful information and reduce the feasible reward set. Furthermore, the limitations introduced over ζ are directly dependent on the expected discounted occupancy of π_{E_i}. Given these considerations, we can appreciate that Equation (6) has provided a precise mathematical description of the phenomenon we identified at the end of the previous section.

As a final remark, we comment on the maximum values that ζ can assume. We recall that, for classical IRL problems, $\|\zeta\|_\infty \leq (1-\gamma)^{-1}$. For the sub-optimal experts case, instead, let us analyze Equation (6) in greater detail. Fix a state $s' \in \mathcal{S}$ and a sub-optimal agent $i \in \{2, \dots, n+1\}$; in this case, the s'-th constraint in Equation (6) can be written as:

$$\sum_{s\in\mathcal{S}} d^{\pi_{E_i}}_{s'}(s) \sum_{a:\pi_{E_1}(a|s)=0} \pi_{E_i}(a|s)\zeta(s,a) \leq \xi_i, \qquad (7)$$

where $d^{\pi_{E_i}}_{s'}(s)$ denotes the discounted expected number of times that policy π_{E_i} visits state s starting from state s'. From Equation (7), we can obtain necessary conditions on the values of ζ that can generate compatible reward functions. More specifically, let $\mathcal{X}(s,a) \subset \{2, \dots, n+1\}$ be the subset of optimal experts such that $\pi_{E_i}(a|s) > 0$. Then, for each state-action pair (s, a) such that $\pi_{E_1}(a|s) = 0$ and $\pi_{E_i}(a|s) > 0$, we have that:

$$\zeta(s,a) \leq \min\left\{k(s,a), \frac{1}{1-\gamma}\right\} \coloneqq g(s,a), \qquad (8)$$

where $k(s,a)$ is given by:

$$k(s,a) \coloneqq \min_{i\in\mathcal{X}(s,a), s'\in\mathcal{S}} \frac{\xi_i}{d^{\pi_{E_i}}_{s'}(s)\pi_{E_i}(a|s)}. \qquad (9)$$

More specifically, the term $k(s,a)$ directly follows from Equation (7), while $(1-\gamma)^{-1}$ is the maximum value that any $\zeta(s,a)$ can assume, and arises, as in the classical IRL setting, from the fact that advantage functions are bounded by $(1-\gamma)^{-1}$ for any possible reward function. In this sense, as shown in the following example, Equation (8) implies a significant potential reduction in the maximum values that the advantage function can take, i.e., how much sub-optimal, in terms of Q-function, an action not played by π_{E_1} can be.

Example 4. Consider a IRL problems with only one additional expert. Suppose that π_{E_1} and π_{E_i} are deterministic. For all state-action pairs in which $\pi_{E_1}(a|s) = 0$ and $\pi_{E_i}(a|s) = 1$, Equation (8) implies that $\zeta(s,a) \leq \min\{\xi_i, (1-\gamma)^{-1}\}$. If ξ_i is significantly smaller than $(1-\gamma)^{-1}$, we obtain a notable restriction on the set of feasible reward functions.

4 LEARNING THE FEASIBLE SET

So far, we have investigated the theoretical properties of the class of reward functions that belong to the feasible set. In this section, we leverage these results to tackle the statistical complexity of estimating $\mathcal{R}_{\bar{\mathfrak{B}}}$ with a *generative model*. Specifically, we first introduce a Probabilistic Approximately Correct (PAC) framework (Section 4.1). Then, we study the statistical complexity of the problem by presenting lower bounds on the number of samples that any algorithm requires in order to correctly identify the feasible set (Section 4.1). Finally, we propose a uniform sampling algorithm and analyze its theoretical guarantees (Section 4.3). As a summary, our results show that (i) the IRL problem with sub-optimal experts is statistically more demanding than the single agent IRL setting, and (ii) that the uniform sampling is minimax optimal whenever the sub-optimal experts' performance level is sufficiently close to the one of the optimal agent. For the sake of presentation, all results are presented under the assumption that π_{E_1} is deterministic. The extension to the case in which π_{E_1} is stochastic is presented in Appendix D.

[2] We notice that, however, contrary to single-agent IRL problems, now Equation (5) is only a necessary condition for having $r \in \mathcal{R}_{\bar{\mathfrak{B}}}$.

[3] We remark that whenever $n = 1$ (i.e., we have only access to the optimal expert π_{E_1}), Theorem 3 simply reduces to Equation (5), and, consequently, it smoothly generalizes existing results for the classical IRL problem.

[4] As a consequence of the linearity, testing whether a given ζ satisfies Equation (6) is computationally efficient.

4.1 PAC Framework

We define a learning algorithm for an IRL problem $\bar{\mathfrak{B}}$ as a tuple $\mathfrak{A} = (\tau, \nu)$, τ is a stopping time that controls the end of the data acquisition phase, and $\nu = (\nu_t)_{t\in\mathbb{N}}$ is a history-dependent sampling strategy over $\mathcal{S} \times \mathcal{A}$. More precisely, $\nu_t \in \Delta_{\mathcal{D}_t}^{\mathcal{S}\times\mathcal{A}}$, where $\mathcal{D}_t = \left(\mathcal{S} \times \mathcal{A} \times \mathcal{S} \times (\mathcal{A})^{n+1}\right)^t$. At each time step $t \in \mathbb{N}$, the algorithm selects a state-action pair $(S_t, A_t) \sim \nu_t$, and observes a sample $S'_t \sim p(\cdot|S_t, A_t)$ from the environment, together with actions sampled from the experts' policy, namely $\left(A_t^{(i)}\right)_{i=1}^{n+1}$, where $A_t^{(i)} \sim \pi_{E_i}(\cdot|S_t)$. The observed realizations are then used to update the sampling strategy ν_t, and the process goes on until the stopping rule is satisfied. At the end of the data acquisition phase, the algorithm leverages the collected data to output the estimate of the feasible reward set $\mathcal{R}_{\widehat{\bar{\mathfrak{B}}}_\tau}$ that is induced by the resulting empirical IRL problem $\widehat{\bar{\mathfrak{B}}}_\tau$. Given this formalism, we are interested in designing learning algorithms that, for any desired accuracy $\epsilon \in (0, 1)$ and any risk parameter $\delta \in (0, 1)$, guarantee that:

$$\mathbb{P}_{\mathfrak{A},\bar{\mathfrak{B}}}\left(H_\infty(\mathcal{R}_{\bar{\mathfrak{B}}}, \mathcal{R}_{\widehat{\bar{\mathfrak{B}}}_\tau}) > \epsilon\right) \leq \delta. \tag{10}$$

We refer to these algorithms as (ϵ, δ)-correct identification strategies. For (ϵ, δ)-correct strategies, we define their sample complexity as the total number of interaction rounds with the generative model before stopping. In other words, the sample complexity is given by τ.

4.2 Statistical Lower Bound

In this section, we present lower bounds on the number of queries to the generative model that any (ϵ, δ)-correct algorithm needs to perform in order to correctly identify the feasible reward set $\mathcal{R}_{\bar{\mathfrak{B}}}$. The following theorem (proof in Appendix C) reports our result.

Theorem 5. *Let $\mathfrak{A}$ be a (ϵ, δ)-correct algorithm for the IRL problem with sub-optimal experts. There exists a problem instance $\bar{\mathfrak{B}}$ such that the expected sample complexity is lower bounded by:*

$$\mathbb{E}_{\mathfrak{A},\bar{\mathfrak{B}}}[\tau] \geq \Omega\left(\frac{SA}{\epsilon^2(1-\gamma)^2}\left(\log\left(\frac{1}{\delta}\right) + S\right)\right), \tag{11}$$

where $\Omega(\cdot)$ hides constant dependencies. Furthermore, let $\pi_{\min}$ be:

$$\pi_{\min} \coloneqq \min_{i\in\{2,\dots n+1\}} \max_{(s,a):\pi_{E_i}(a|s)>0} \pi_{E_i}(a|s), \tag{12}$$

and define $q_0 \coloneqq \pi_{\min}^{-1} \max_{i\in\{2,\dots,n+1\}} \xi_i$. Then there exists an instance $\bar{\mathfrak{B}}'$ in which $q_0 < 1$ such that:

$$\mathbb{E}_{\mathfrak{A},\bar{\mathfrak{B}}'}[\tau] \geq \Omega\left(\frac{q_0^2 S \log\left(\frac{1}{\delta}\right)}{\epsilon^2 \pi_{\min}}\right). \tag{13}$$

Theorem 5 provides two distinct lower bounds (i.e., Equations (11) and (13)) for IRL problems with sub-optimal experts. As a consequence, we notice that whenever $q_0 < 1$ holds, the lower bound for the IRL-SE setting can be expressed as the maximum between Equation (11) and (13). At this point, we will comment in-depth on these two equations.

Concerning Equation (11), as our analysis reveals, it directly arises from the problem of estimating rewards functions that are compatible with π_{E_1} (i.e., with Equation (5) in Theorem 3). In this sense, it represents the complexity of single-agent IRL problems.[5] As a precise consequence of the structure of the feasible region we derived in Theorem 3, this results in a lower bound also for the multiple sub-optimal experts setting. Therefore, Equation (11) formally shows that the sub-optimal expert setting is always at least as difficult as the single agent IRL problem.

Equation (13), on the other hand, is strongly related to the presence of sub-optimal experts. More precisely, under the assumption that $q_0 < 1$ (e.g., for sufficiently small values of ξ_i), it shows a dependency in the lower bound of a factor $\pi_{\min}^{-1}$, where $\pi_{\min}$ represents the minimum probability with which sub-optimal experts plays their actions. From an intuitive perspective, its presence is related to the difficulty in estimating reward functions that are compatible with Equation (6) in Theorem 3. Indeed, as we have shown in Section 3, the presence of sub-optimal agents can limit the value of ζ with a relationship that involves $\pi_{\min}^{-1}$ (i.e., Equation (8)). As our analysis will reveal, the proof of Equation (13) is directly related to these worst-case upper-bounds on ζ (and, in order to exploit them successfully, we needed to restrict ourselves to the case in which $q_0 < 1$). At this point, it has to be remarked that, according to the value of $\pi_{\min}$, Equation (13) can be significantly larger than Equation (6), thus showing an increased difficulty in the statistical complexity that is related to the stochasticity of sub-optimal experts.

At this point, it has to be noticed that the generative model we defined in Section 4.1 is significantly more powerful than the one adopted in a classical IRL setting (see, e.g., Metelli et al., 2021, 2023). For single-agent problems, indeed, a query to the generative model provides only samples from the environment and from the expert agent π_{E_1}. In our context, on the other hand, for each query, the generative model provides demonstrations from *each* sub-optimal expert. It can be shown that, by slightly modifying the

[5] We notice that similar results were presented in Metelli et al. (2023) for the finite-horizon single expert IRL problem. In this work, we extend their construction and analysis to the infinite-horizon IRL model.

Algorithm 1 Uniform Sampling for Inverse RL with Suboptimal Experts (US-IRL-SE)

Require: samples collected in each (s, a) pair m
1: **for** $t = 1, 2, \dots, m$ **do**
2: Collect one tuple $(s', (a^{(i)})_{i=1}^{n+1})$ where $s' \sim p(\cdot|s, a)$ and $a^{(i)} \sim \pi_{E_i}(\cdot|s)$ from each $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $i \in \{1 \dots, n+1\}$
3: Update $\hat{p}$ and $(\hat{\pi}_{E_i})_{i=1}^{n+1}$ according to Equation (4)
4: **end for**

learning formalism, Equation (13) actually represents a lower bound to the number of samples that should be gathered from *each* sub-optimal agent.[6] In this sense, the statistical complexity increases significantly in the sub-optimal expert setting compared to the single agent one. Therefore, as a concluding remark, we notice that, in order to gain the reduction in the feasible reward set that we discussed in Section 5, we need to gather additional data in terms of demonstrations from the sub-optimal experts. This unavoidable trade-off is a direct consequence of the structure of the feasible set $\mathcal{R}_{\widehat{\mathfrak{B}}}$ that we derived in Theorem 3, and, indeed, it arises from the statistical complexity of estimating reward functions that are compatible with the linear constraints of Equation (6).

4.3 Uniform Sampling Algorithm

In this section, we present the Uniform Sampling algorithm for Inverse RL with Suboptimal Experts (US-IRL-SE). The pseudo-code can be found in Algorithm 1. As we can see, US-IRL-SE receives the number of samples m that will be queried to the generative model in each state-action pair. Then, it uniformly gathers data across the entire state-action space, and it updates the empirical estimates $\hat{p}$ and $(\hat{\pi}_{E_i})_{i=1}^{n+1}$.

The following theorem (proof in Appendix C), describes the theoretical guarantees of US-IRL-SE.

Theorem 6. *Let* $q_1 = \min\left\{\pi_{\min}^{-1} \max_i \xi_i, (1-\gamma)^{-1}\right\}$, *and* $q_2 = \max\{1, q_1\}$. *Then, with a total budget of:*

$$\widetilde{\mathcal{O}}\left(\max\left\{\frac{q_1^2 S \log\left(\frac{1}{\delta}\right)}{\pi_{\min}\epsilon^2}, \frac{q_2^2 SA(S + \log\left(\frac{1}{\delta}\right))}{\epsilon^2(1-\gamma)^2}\right\}\right), \quad (14)$$

US-IRL-SE is (ϵ, δ)*-correct and* $\widetilde{\mathcal{O}}(\cdot)$ *hides constant and logarithmic dependencies.*

Theorem 6 deserves some comments. First of all, it formally shows that when the total number of queries to the generative is sufficiently large, US-IRL-SE is (ϵ, δ)-correct, and its sample complexity is provided in Equation (14). In this sense, we notice that, since m represents the number of calls to the generative model in each state-action pair, its expression can simply be calculated by dividing Equation (14) by SA.[7] As a consequence, we remark that, in order to compute the value of m, the algorithm requires knowledge of the minimum probability with which sub-optimal experts play their actions.

We now proceed by analyzing in detail the sample complexity guarantee. Equation (14) is the maximum between two terms whose expressions closely resemble the lower bound that we presented in Theorem 5. Specifically, the only difference arises in the definition of q_0, q_1 and q_2. Currently, we are unsure whether this gap arises from the lower bound or the algorithm analysis, and we leave this gap to be filled in for future work. Nevertheless, it has to be remarked that, whenever the sub-optimal expert's performance level is sufficiently close to the one of the optimal agent (i.e., $\pi_{\min}^{-1}\xi_i \leq 1$ for all $i \in \{2, \dots, n+1\}$), Equation (14) exactly recovers the lower bound that we presented in Theorem 5.[8] We remark that according to Theorem 3, as the values of ξ_i's decrease, the feasible reward set is substantially reduced. In this sense, US-IRL-SE enjoys minimax optimality in the most interesting scenarios where the presence of sub-optimal experts is particularly valuable for mitigating the intrinsic ambiguity that affects inverse reinforcement learning problems.

Technical Remark To conclude, we highlight that, although the algorithm is relatively simple, the proof of Theorem 6 requires significant technical effort. The main challenge arises from studying how the Hausdorff distance between $\mathcal{R}_{\widehat{\mathfrak{B}}}$ and $\mathcal{R}_{\widehat{\mathfrak{B}}_t}$ decreases as we collect more data from the generative model. Indeed, we recall that these feasible reward sets are subject to the peculiar structure that we identified in Theorem 3. More specifically, the set of constraints of Equation (6) that arises from the presence of sub-optimal experts complicates significantly the study of $H_\infty\left(\mathcal{R}_{\widehat{\mathfrak{B}}}, \mathcal{R}_{\widehat{\mathfrak{B}}_t}\right)$. For further details on this point, we invite the reader to consult our proofs Appendix C.

5 RELATED WORKS

Inverse Reinforcement Learning Historically, solving an IRL problem (Adams et al., 2022) involves determining a reward function that is compatible with the behavior of an optimal expert. Since the seminal work of Ng et al. (2000), the problem has been recognized as ill-posed, as multiple reward functions that

[6] For further details on this point, we defer the reader to Appendix E.

[7] The exact expression of m (i.e., constants and hidden logarithmic factors) is provided in Appendix F.

[8] More precisely, under the condition that $\pi_{\min}^{-1}\xi_i \leq 1$, it holds that $q_0 = q_1$, and $q_2 = 1$.

satisfies this requirement exists (Skalse et al., 2023). For this reason, over the years, several algorithmic criteria have been introduced to address this ambiguity issue. These criteria includes maximum margin (Ratliff et al., 2006), Bayesian approaches (Ramachandran and Amir, 2007), maximum entropy (Ziebart et al., 2008), and many others (e.g., Majumdar et al., 2017; Metelli et al., 2017; Zeng et al., 2022). More recently, a new line of works have circumvented the ambiguity issue by redefining the IRL task as the problem of estimating the entire feasible reward set (Metelli et al., 2021; Lindner et al., 2022; Metelli et al., 2023). In our work, we take this novel perspective, and, in this sense, this recent research strand is the most related to our document. Specifically, of particular interests is the work of Metelli et al. (2023). In their work, the authors study, for the first time, lower bounds for the single-agent IRL problem in finite horizon settings; furthermore, they show that uniform sampling algorithm is minimax optimal for this task. Nevertheless, it has to remarked that this recent strand of research focuses entirely on single expert problems. As we have shown, however, the extension to the multiple and sub-optimal agents setting requires non-trivial effort. Indeed, the feasible reward set significantly differ (see, e.g., Theorem 3), and the problem is harder from a statistical perspective (see, e.g., Theorem 5).

Multiple and/or Sub-optimal Experts The presence of multiple/sub-optimal experts has garnered attention in the Imitation Learning (IL, Hussein et al., 2017) community. In IL problems, contrary to IRL, the goal lies in directly leveraging demonstrations of optimal behavior to accelerate the training process of reinforcement learning algorithms. In this context, works that are close in spirit to ours are Kurenkov et al. (2020); Jing et al. (2020); Cheng et al. (2020); Liu et al. (2023); here, the authors extends the IL formulation to account for the fact that demonstrations are provided from multiple and/or sub-optimal experts. However, unlike our specific focus, their emphasis is on understanding how to effectively exploit imperfect demonstrations to improve training of RL agents. In our work, instead, we exploit the presence of sub-optimal agents to reduce the intrinsic ambiguity that affects the IRL formulation. In this sense, our work is complementary to several studies that analyzed how to improve the identifiability of the reward function in IRL problems by making additional structural assumptions. These include the possibility of observing an optimal agent interacting with several MDPs (e.g., Ratliff et al., 2006; Amin and Singh, 2016; Amin et al., 2017) and focusing on peculiar types of MDPs that allows for strong theoretical guarantees (e.g., Dvijotham and Todorov, 2010; Kim et al., 2021; Cao et al., 2021). Along this line of work, the most related to ours is Rolland et al. (2022). Here, the authors study how the presence of multiple experts impact the identifiability of the reward function. Contrary to our work, however, the authors assume each agent to follow an entropy regularized objective and, furthermore, they focus on the case in which all experts act optimally in the underlying environment. In this sense, our work encompasses a wider spectrum of applications, as we do not require optimality for each of the agent, nor an entropy regularized objective. Finally, it has to be remarked that the multiple expert setting and IRL have been studied in Likmeta et al. (2021) with the goal of providing practical algorithms that can be used in real-world applications. Also in this scenario, each agent is assumed to act optimally in the underlying domain.

6 CONCLUSIONS

In this work, we studied the novel problem of Inverse RL where, in addition to demonstrations from an optimal expert, we can observe the behavior of multiple and sub-optimal agents. More precisely, we first investigated the theoretical properties of the class of reward functions that are compatible with a given set of experts, i.e., the feasible reward set. Our results formally show that, by exploiting this additional structure, it is possible to significantly reduce the intrinsic ambiguity that affects the IRL formulation. Secondly, we have tackled the statistical complexity of estimating the feasible reward set from a generative model. More precisely, we have shown that a uniform sampling algorithm is minimax optimal whenever the performance level of the sub-optimal expert is sufficiently close to the one of the optimal agent.

Our research opens up intriguing avenues for future studies. For instance, since we have shown that sub-optimal experts can improve the identifiability of the reward function, future research should focus on building practical algorithms that can exploit this additional structure. To this end, as an intermediate step, it might be interesting to extend our results to the case in which the reward function is expressed as a linear combination of features. This approach would enable addressing infinite state-spaces (e.g., Ng et al., 2000).

References

Adams, S., Cody, T., and Beling, P. A. (2022). A survey of inverse reinforcement learning. *Artificial Intelligence Review*, 55(6):4307–4346.

Amin, K., Jiang, N., and Singh, S. (2017). Repeated inverse reinforcement learning. *Advances in neural information processing systems*, 30.

Amin, K. and Singh, S. (2016). Towards resolving

unidentifiability in inverse reinforcement learning. *arXiv preprint arXiv:1601.06569*.

Arora, S. and Doshi, P. (2021). A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500.

Cao, H., Cohen, S. N., and Szpruch, L. (2021). Identifiability in inverse reinforcement learning. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Cheng, C.-A., Kolobov, A., and Agarwal, A. (2020). Policy improvement via imitation of multiple oracles. *Advances in Neural Information Processing Systems*, 33:5587–5598.

Dvijotham, K. and Todorov, E. (2010). Inverse optimal control with linearly-solvable mdps. In *Proceedings of the 27th International conference on machine learning (ICML-10)*, pages 335–342.

Even-Dar, E., Mannor, S., and Mansour, Y. (2002). Pac bounds for multi-armed bandit and markov decision processes. In *Computational Learning Theory: 15th Annual Conference on Computational Learning Theory, COLT 2002 Sydney, Australia, July 8–10, 2002 Proceedings 15*, pages 255–270. Springer.

Hussein, A., Gaber, M. M., Elyan, E., and Jayne, C. (2017). Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35.

Jing, M., Ma, X., Huang, W., Sun, F., Yang, C., Fang, B., and Liu, H. (2020). Reinforcement learning from imperfect demonstrations under soft expert guidance. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5109–5116.

Jonsson, A., Kaufmann, E., Ménard, P., Darwiche Domingues, O., Leurent, E., and Valko, M. (2020). Planning in markov decision processes with gap-dependent sample complexity. *Advances in Neural Information Processing Systems*, 33:1253–1263.

Kim, K., Garg, S., Shiragur, K., and Ermon, S. (2021). Reward identification in inverse reinforcement learning. In *International Conference on Machine Learning*, pages 5496–5505. PMLR.

Kurenkov, A., Mandlekar, A., Martin-Martin, R., Savarese, S., and Garg, A. (2020). Ac-teach: A bayesian actor-critic method for policy learning with an ensemble of suboptimal teachers. In *Proceedings of the Conference on Robot Learning*.

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

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

Liu, X., Yoneda, T., Wang, C., Walter, M., and Chen, Y. (2023). Active policy improvement from multiple black-box oracles. In *International Conference on Machine Learning*, pages 22320–22337. PMLR.

Majumdar, A., Singh, S., Mandlekar, A., and Pavone, M. (2017). Risk-sensitive inverse reinforcement learning via coherent risk models. In *Robotics: science and systems*, volume 16, page 117.

Metelli, A. M., Lazzati, F., and Restelli, M. (2023). Towards theoretical understanding of inverse reinforcement learning. *arXiv preprint arXiv:2304.12966*.

Metelli, A. M., Pirotta, M., and Restelli, M. (2017). Compatible reward inverse reinforcement learning. *Advances in neural information processing systems*, 30.

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

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., and Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56(4):3005–3054.

Ng, A. Y., Russell, S., et al. (2000). Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, volume 1, page 2.

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

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

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

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

Rolland, P., Viano, L., Schürhoff, N., Nikolov, B., and Cevher, V. (2022). Identifiability and generalizability from multiple experts in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 35:550–564.

Skalse, J. M. V., Farrugia-Roberts, M., Russell, S., Abate, A., and Gleave, A. (2023). Invariance in policy optimisation and partial identifiability in reward

learning. In *Proceedings of the 40th International Conference on Machine Learning*.

Zeng, S., Li, C., Garcia, A., and Hong, M. (2022). Maximum-likelihood inverse reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 35:10122–10135.

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

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, mathematical setting are provided in Section 2 and Section 4.1. For the algorithm, we included a pseudo-code that explains its behavior (see Algorithm 1).
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes. The statistical complexity of the algorithm is described in Theorem 6. For computational complexity, we refer the reader to Appendix F.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Not Applicable.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes. Each theoretical statements is precise.
   (b) Complete proofs of all theoretical results. Yes, complete proof of all theoretical results are presented in Appendix B and C.
   (c) Clear explanations of any assumptions. Yes, below each theoretical results we include an in-depth discussion that explains the results, together with the theoretical requirements. These discussions includes details on the assumptions needed.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Not Applicable.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Not Applicable.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Not Applicable.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Not Applicable.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable.
   (b) The license information of the assets, if applicable. Not Applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.
   (d) Information about consent from data providers/curators. Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.