Journal of Machine Learning Research 27 (2026) 1-50 Submitted 1/25; Revised 9/26; Published 11/26

Reward Compatibility: A Framework for Inverse RL

Filippo Lazzati FILIPPO.LAZZATI@POLIMI.IT
Politecnico di Milano
Milan, Italy

Mirco Mutti MIRCO.M@TECHNION.AC.IL
Technion
Haifa, Israel

Alberto Maria Metelli ALBERTOMARIA.METELLI@POLIMI.IT
Politecnico di Milano
Milan, Italy

Editor: Zhaoran Wang

Abstract

We provide an original theoretical study of Inverse Reinforcement Learning (IRL) through the lens of *reward compatibility*, a novel framework to quantify the compatibility of a reward with the given expert's demonstrations. Intuitively, a reward is more *compatible* with the demonstrations the closer the performance of the expert's policy computed with that reward is to the optimal performance for that reward. This generalizes the notion of *feasible reward set*, the most common framework in the theoretical IRL literature, for which a reward is either compatible or not compatible. The *grayscale* introduced by the reward compatibility is the key to extend the realm of provably efficient IRL far beyond what is attainable with the feasible reward set: from tabular to *large-scale* MDPs. We analyze the IRL problem across various settings, including optimal and *suboptimal* expert's demonstrations and both online and *offline* data collection. For all of these dimensions, we provide a tractable algorithm and corresponding sample complexity analysis, as well as various insights on reward compatibility and how the framework can pave the way to yet more general problem settings.

Keywords: Inverse reinforcement learning, Linear MDPs, sample complexity, reward-free exploration, identifiability

1. Introduction

Inverse Reinforcement Learning (IRL) is the problem of inferring the reward function of an agent, named the expert agent, from demonstrations of behavior (Russell, 1998; Ng and Russell, 2000). Since its formulation, much research effort has been put into the design of efficient algorithms for solving the IRL problem (Arora and Doshi, 2018; Adams et al., 2022), with the promise to open the door to a variety of interesting applications, including Apprenticeship Learning (AL, Abbeel and Ng, 2004; Abbeel et al., 2006), reward design (Hadfield-Menell et al., 2017), interpretability of the expert's behavior (Hadfield-Menell et al., 2016), and transferability of behavior to new environments (Fu et al., 2017).

©2026 Filippo Lazzati, Mirco Mutti, Alberto Maria Metelli.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/25-0105.html.

Despite the relevance of these applications and the abundance of practical solutions, IRL has long escaped a formal and coherent theoretical characterization to support the empirical research. Indeed, the *vanilla* formulation of IRL, coarsely "given a set of demonstrations from an expert's policy recover the reward function the expert is maximizing", is notoriously ill-posed (Ng and Russell, 2000; Metelli et al., 2021, 2023), as several rewards (infinitely many) are *feasible* to explain the evidence provided by expert's demonstrations, no matter the number of available demonstrations. To resolve this ambiguity, it has been proposed to consider *additional* demonstrations from multiple environments (Amin and Singh, 2016; Cao et al., 2021) or from multiple experts (Rolland et al., 2022; Poiani et al., 2024), or to consider *additional* human feedback (Jeon et al., 2020; Skalse et al., 2023). However, this extra information is not always available.

Only recently, the notion of *feasible reward set*, introduced by Metelli et al. (2021), has convincingly sorted out the issue, emerging as a common theoretical framework for the study of IRL (Lindner et al., 2022; Metelli et al., 2023; Lazzati et al., 2024a; Zhao et al., 2024; Poiani et al., 2024) and related problems (Lazzati and Metelli, 2025; Yue et al., 2024; Freihaut and Ramponi, 2024). In the absence of conclusive information coming from expert's demonstrations, Metelli et al. (2021) propose to extract not just one feasible reward, but the set of all of the rewards that make the expert's policy optimal, called the *feasible reward set*. This formulation lead to the development of provably efficient IRL algorithms and clarified the statistical barriers of the problem (Metelli et al., 2023). However, the application of the feasible reward set framework is essentially limited to settings where the space of states and actions is finite and small, falling into the *tabular* Markov Decision Process (MDP) formalism. As we shall see, this constraint is inherent to the feasible set and cannot be overcome.

Unfortunately, important potential applications of IRL do not comply with this property. Think about the problem of extracting a reward from demonstrations collected by an expert driver (Ziebart et al., 2008) or an helicopter control policy (Abbeel et al., 2006). These settings typically involve large or continuous *state spaces*, thus invalidating the tabular representation of the space. Previous works have studied IRL beyond tabular MDPs (Michini et al., 2013; Finn et al., 2016; Fu et al., 2017; Barnes et al., 2024), but none of them from the theoretical viewpoint of the feasible set.

From these considerations, the need for a unifying framework that incorporates the important traits of the feasible reward set on the one hand, i.e., provides a formal and coherent theoretical characterization of the sample complexity of IRL, yet extends the notion to MDPs with large state spaces naturally arises.

In order to be useful and general, such framework should be easily applicable not only to the common IRL problem of Ng and Russell (2000), where (i) the expert's demonstrations are collected with an *optimal* policy, and (ii) the IRL algorithm can *actively explore* the environment to collect the data it needs, but also to other settings in which these properties do not hold. In fact, note that in all the settings in which demonstrations are provided by humans, property (i) may also be violated. Humans are characterized by a bounded rationality and may not be able to optimally solve the demonstrated task. This is why a more realistic setting, in which demonstrations of behavior come from suboptimal experts,

has been widely studied in the literature (Ziebart et al., 2008; Jing et al., 2019; Kurenkov et al., 2019; Cheng et al., 2020; Liu et al., 2023; Poiani et al., 2024). Moreover, while letting the IRL algorithm actively interact with the environment helps improve its performance in many cases (e.g., in simulation problems Neu and Szepesvári 2007), it may not be possible in settings where safety is critical. Think about the possibility of damaging the helicopter in Abbeel et al. (2006) or harm other people in an autonomous driving scenario (Arora and Doshi, 2018). This is why previous works considered access to a batch of demonstrations and no interaction with the environment (Jarboui and Perchet, 2021; Zhao et al., 2024; Lazzati et al., 2024a) invalidating property (*ii*).

In this paper, we build on these premises as follows.

Contributions. We propose the *reward compatibility* framework as a unifying learning framework for the development of efficient algorithms for IRL. The main contributions of the current work are summarized as follows:

- We demonstrate that, without additional assumptions, the notion of feasible set can *not* be learned efficiently in Markov Decision Processes (MDPs) when the state space is large/continuous, even under the structure enforced by Linear MDPs (Jin et al., 2020b; Yang and Wang, 2019), in which we assume that the reward function and the transition model can be expressed as linear combinations of known features. This fact entails that the learning framework of the feasible reward set is not powerful enough to manage all the meaningful IRL problem settings (Section 3).
- As a unifying learning framework for IRL, we propose *Reward Compatibility*, a novel scheme that formalizes the intuitive notion of compatibility of a reward function with expert demonstrations. It generalizes the feasible set and allows us to define an original problem, *IRL classification*. We devise a provably-efficient algorithm, `CATY-IRL` (`C`omp`AT`ibilit`Y` for `IRL`), for solving the IRL classification problem in both the online settings with optimal and suboptimal expert, and in both tabular and Linear MDPs (see Section 4 and Table 1).
- In addition, we focus on the offline IRL setting in tabular MDPs. We show that, given the partial coverage of the space induced by the batch data set, the offline setting requires suitable "robust" notions of reward compatibility, that we provide. Then, we develop a provably-efficient algorithm, `CATY-OFF-IRL` (`C`omp`AT`ibilit`Y` for `OFF`line `IRL`), for solving the problem in both the optimal and suboptimal expert settings (see Section 6 and Table 1).
- We conclude with a discussion on reward compatibility that ranges from Reward Learning (ReL) (Jeon et al., 2020), to other IRL problem settings (e.g., Ziebart et al., 2008), to the practical usage of learned reward functions, and we identify a variety of interesting research directions for future works.

This paper unifies and extends the previous conference papers (Lazzati et al., 2024a,b). The former demonstrates the need of suitable "robust" notions of feasible set in the offline setting, and studies them. Instead, the latter demonstrates the limitations of the feasible set in Linear MDPs, formalizes the reward compatibility framework, and presents a version of `CATY-IRL` for the setting with optimal expert. In this paper, we extend the contributions of Lazzati et al. (2024a,b) as follows:

	Tabular MDPs	Linear MDPs
Online (CATY-IRL)	$\widetilde{\mathcal{O}}\left(\frac{H^3SA}{\epsilon^2}\left(S+\log\frac{1}{\delta}\right)\right)$	$\widetilde{\mathcal{O}}\left(\frac{H^5d}{\epsilon^2}\left(d+\log\frac{1}{\delta}\right)\right)$
Offline (CATY-OFF-IRL)	$\widetilde{\mathcal{O}}\left(\frac{H^4\log\frac{1}{\delta}}{\epsilon^2 d_{\min}}\left(S+\log\frac{1}{\delta}\right)\right)$	/

Table 1: In this table, we summarize the theoretical guarantees of sample complexity concerning the estimation of the transition model for the proposed algorithms. Interestingly, our algorithms enjoy the *same rates in both the optimal and suboptimal expert settings*. In tabular MDPs, S denotes the size of the state space, A that of the action space, H is the horizon, ϵ the accuracy, and δ the failure probability. d represents the feature dimension in Linear MDPs, while $d_{\min}$ keeps into account the coverage of the environment provided by the batch data set. We leave for future works the development of efficient algorithms for Linear MDPs in the offline setting. Indeed, the notion of compatibility that we adopted for the offline tabular setting is based on a definition of coverage of the space that cannot be applied to Linear MDPs. For this reason, a new notion of reward compatibility that exploits the existing definitions of coverage for Linear MDPs (Wang et al., 2020a; Jin et al., 2021b) should be developed, falling outside the scope of this paper.

- We extend the results on the limitations of the feasible set in Lazzati et al. (2024b) to the setting with suboptimal expert (Theorem 2), and to other function approximation settings beyond Linear MDPs (Theorem 3 and a discussion in Section 3).
- We extend the reward compatibility framework and the CATY-IRL algorithm in Lazzati et al. (2024b) to the setting with suboptimal experts, and we provide theoretical guarantees of sample complexity for the algorithm (Theorem 4, the claim concerning suboptimal experts). Moreover, we provide sufficient conditions to extend CATY-IRL to any function approximation setting beyond tabular and linear MDPs (see Section 5.3).
- We demonstrate the need of suitable "robust" notions of reward compatibility if we want to extend the framework to the offline setting (Theorem 5), and we devise them (Definition 11), similarly to what derived in Lazzati et al. (2024a) for the feasible set.
- We extend CATY-IRL to both the offline settings with optimal and suboptimal expert in tabular MDPs, and we provide theoretical guarantees of sample complexity (Theorem 7), in a setting analogous to that analyzed in Lazzati et al. (2024a).

2. Preliminaries

Notation. Given an integer $N \in \mathbb{N}$, we define $[\![N]\!] \coloneqq \{1, \dots, N\}$. We denote by $\Delta^{\mathcal{X}}$ the probability simplex over $\mathcal{X}$, and by $\Delta_{\mathcal{Y}}^{\mathcal{X}}$ the set of functions from $\mathcal{Y}$ to $\Delta^{\mathcal{X}}$. Sometimes, we denote the dot product between vectors x, y as $\langle x, y \rangle \coloneqq x^\intercal y$. We employ $\mathcal{O}, \Omega, \Theta$ for the common asymptotic notation and $\widetilde{\mathcal{O}}, \widetilde{\Omega}, \widetilde{\Theta}$ to omit logarithmic terms. Given an equivalence relation $\equiv \subseteq \mathcal{X} \times \mathcal{X}$ and an item $x \in \mathcal{X}$, we denote by $[x]_{\equiv}$ the equivalence class of x.

Markov Decision Processes. A finite-horizon Markov Decision Process (MDP) without reward (Puterman, 1994) is defined as a tuple $\mathcal{M} \coloneqq (\mathcal{S}, \mathcal{A}, H, d_0, p)$, where $\mathcal{S}$ and $\mathcal{A}$ are the

measurable state and action spaces, $H \in \mathbb{N}$ is the horizon, $d_0 \in \Delta^{\mathcal{S}}$ is the initial-state distribution,[1] and $p \in \mathcal{P} := \Delta^{\mathcal{S}}_{\mathcal{S}\times\mathcal{A}\times[\![H]\!]}$ is the transition model. Given a (deterministic) reward function $r \in \mathfrak{R} := [-1,1]^{\mathcal{S}\times\mathcal{A}\times[\![H]\!]}$, we denote by $\overline{\mathcal{M}} := \mathcal{M}\cup\{r\}$ the MDP obtained by pairing $\mathcal{M}$ and r. Each policy $\pi \in \Pi := \Delta^{\mathcal{A}}_{\mathcal{S}\times[\![H]\!]}$ induces in $\overline{\mathcal{M}}$ a state-action probability distribution $d^{p,\pi} := \{d_h^{p,\pi}\}_{h\in[\![H]\!]}$ (we omit d_0 for simplicity) that assigns, to each subset $\mathcal{Z} \subseteq \mathcal{S}\times\mathcal{A}$, the probability of being in $\mathcal{Z}$ at stage $h \in [\![H]\!]$ when playing π in $\overline{\mathcal{M}}$. We denote with $\mathcal{S}_h^{p,\pi}$ (resp. $\mathcal{Z}_h^{p,\pi}$) the set of states (resp. state-action pairs) supported by $d_h^{p,\pi}$ at stage h, and with $\mathcal{S}^{p,\pi}$ (resp. $\mathcal{Z}^{p,\pi}$) the disjoint union of sets $\{\mathcal{S}_h^{p,\pi}\}_{h\in[\![H]\!]}$ (resp. $\{\mathcal{Z}_h^{p,\pi}\}_{h\in[\![H]\!]}$). For distribution $d^{p,\pi}$, we define $d^{p,\pi}_{\min} := \min_{(s,a,h)\in\mathcal{Z}^{p,\pi}} d_h^{p,\pi}(s,a)$. The Q-function of policy π in MDP $\overline{\mathcal{M}}$ is defined at every $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[\![H]\!]$ as $Q_h^\pi(s,a;p,r) := \mathbb{E}_{p,\pi}[\sum_{t=h}^H r_t(s_t,a_t)|s_h = s, a_h = a]$, and the optimal Q-function as $Q_h^*(s,a;p,r) := \sup_{\pi\in\Pi} Q_h^\pi(s,a;p,r)$, where the expectation $\mathbb{E}_{p,\pi}$ is computed over the stochastic process generated by playing policy π in the MDP $\overline{\mathcal{M}}$. Similarly, we define the V-function of policy π at (s,h) as $V_h^\pi(s;p,r) := \mathbb{E}_{p,\pi}[\sum_{t=h}^H r_t(s_t,a_t)|s_h = s]$, and the optimal V-function as $V_h^*(s;p,r) := \sup_{\pi\in\Pi} V_h^\pi(s;p,r)$. We define the utility of π as $J^\pi(r;p) := \mathbb{E}_{s\sim d_0}[V_1^\pi(s;p,r)] = \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} d_h^{p,\pi}(s,a) r_h(s,a)$, and the optimal utility as $J^*(r;p) := \mathbb{E}_{s\sim d_0}[V_1^*(s;p,r)]$. A forward (sampling) model of the environment permits to collect samples starting from $s \sim d_0$ and following some policy.[2] A generative (sampling) model consists in an oracle that, given an arbitrary state-stage pair s,h (resp. state-action-stage triple s,a,h) in input, returns a sampled action $a' \sim \pi_h(\cdot|s)$ (resp. a sampled next state $s' \sim p_h(\cdot|s,a)$). For simplicity, when $\mathcal{S},\mathcal{A}$ are finite, given a set $\overline{\mathcal{Z}} \subseteq \mathcal{S}\times\mathcal{A}\times[\![H]\!]$, we define the equivalence relation $\equiv_{\overline{\mathcal{Z}}} \subseteq \mathcal{P}\times\mathcal{P}$ such that, for any pair $p,p' \in \mathcal{P}$ of transition models, we have $p \equiv_{\overline{\mathcal{Z}}} p'$ if and only if $\forall (s,a,h) \in \overline{\mathcal{Z}} :\ p_h(\cdot|s,a) = p'_h(\cdot|s,a)$.

BPI and RFE. In both Best-Policy Identification (BPI) (Menard et al., 2021) and Reward-Free Exploration (RFE) (Jin et al., 2020a), the learner has to explore the *unknown* MDP to optimize a certain reward function. In BPI, the learner observes the reward function r during exploration, and its goal is to output a policy $\widehat{\pi}$ such that, in the true MDP with transition model p, we have $\mathbb{P}\big(J^*(r;p) - J^{\widehat{\pi}}(r;p) \leqslant \epsilon\big) \geqslant 1-\delta$ for every $\epsilon,\delta \in (0,1)$. RFE considers the setting in which the reward is revealed *a posteriori* of the exploration phase. Thus, the goal of the agent in RFE is to compute an estimate $\widehat{p}$ of the true dynamics p so that $\mathbb{P}\big(\sup_{r\in\mathfrak{R}}(J^*(r;p) - J^{\widehat{\pi}_r}(r;p)) \leqslant \epsilon\big) \geqslant 1-\delta$ for every $\epsilon,\delta \in (0,1)$, where $\widehat{\pi}_r$ is the optimal policy in the MDP with $\widehat{p}$ as transition model and r as reward function.

IRL. In the most common IRL setting, we are given an MDP without reward $\mathcal{M} := (\mathcal{S},\mathcal{A},H,d_0,p)$ (p may be known or unknown) and a data set of n state-action trajectories $\mathcal{D} = \{(s_1^i,a_1^i,\dots,s_H^i,a_H^i,s_{H+1}^i)\}_{i\in[\![n]\!]}$ obtained by executing n times the expert's policy π^E in $\mathcal{M}$. The underlying assumption is that there exists a true expert's reward r^E that "someway" relates to π^E, and the IRL objective is to use the knowledge of $\mathcal{M}$ and the data set $\mathcal{D}$ to find a reward $\widehat{r} \approx r^E$ that represents a good approximation for the unknown r^E according to some performance index. In literature, different works have analysed different kinds of relationships between r^E and π^E. To make some examples, Ng and Russell (2000) assume that π^E is optimal under r^E, Ziebart et al. (2010) assume that π^E is the maximum causal entropy policy for r^E, Ramachandran and Amir (2007) assume that π^E is Boltzmann

1. In most results, w.l.o.g. we will consider a deterministic initial state distribution $d_0(s_0) = 1$.
2. See Kakade (2003) or Azar et al. (2013).

rational for r^E, and Poiani et al. (2024) assume that the suboptimality of π^E under r^E is no more than some given threshold ξ. We remark that different assumptions give birth to different *solution concepts* for IRL, i.e., given the same expert's policy π^E, these works aim to estimate different reward functions. The IRL solution concept of the *feasible set* has been formulated more recently (Metelli et al., 2021, 2023).

3. Limitations of the Feasible Set

In this section, we show that the notion of *feasible set*, an important solution concept for IRL, unfortunately, cannot be estimated efficiently in MDPs with a large state space, even under the structure provided by certain function approximation settings, particularly Linear MDPs. This limitation, along with some implementability issues, urges the introduction of a more powerful solution concept for IRL, that overcomes these limitations while preserving the desirable properties of the feasible set. Let us begin by formalizing the notion of feasible set in two important IRL settings.

3.1 The Feasible Set

We consider the IRL settings where we are given demonstrations collected by the expert's policy π^E, and we assume that π^E is optimal (Ng and Russell, 2000) or suboptimal (Poiani et al., 2024) under the true, unknown, expert's reward r^E. In both cases, even the exact knowledge of π^E is not sufficient for uniquely identifying r^E, because many other reward functions satisfy the same constraint of making π^E (sub)optimal. The *feasible set* is defined as the set of rewards that comply with such constraint (Metelli et al., 2021, 2023). Intuitively, we do not know the "true reason" (r^E) why the expert plays π^E because we do not have enough information, but we know that if it plays π^E then its "goal" is one of those in the feasible set. Thus, simply put, the feasible set represents the entire set of objectives that the expert may be optimizing. Formally, with optimal expert:[3]

Definition 1 (Feasible Set for Optimal Expert)**.** *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$ be an MDP without reward and let π^E be the expert's policy, that is optimal for the true unknown reward r^E. The* feasible set $\mathcal{R}_{\mathcal{M},\pi^E}$ *of rewards compatible with π^E in $\mathcal{M}$ is defined as*

$$\mathcal{R}_{\mathcal{M},\pi^E} \coloneqq \left\{ r \in \mathfrak{R} \,\middle|\, J^{\pi^E}(r;p) = J^*(r;p) \right\}. \tag{1}$$

Simply put, we are told that the true reward r^E satisfies the constraint $J^{\pi^E}(r^E;p) = J^*(r^E;p)$, thus, all the rewards in $\mathcal{R}_{\mathcal{M},\pi^E}$ represent equally-plausible candidates for r^E.

Analogously, we can define the feasible set for the setting in which the expert is suboptimal with a suboptimality contained in the (known) range $[L, U]$, for some $U \geqslant L \geqslant 0$. Note that this setting generalizes the setting in Poiani et al. (2024).

Definition 2 (Feasible Set for Suboptimal Expert)**.** *Let $U \geqslant L \geqslant 0$. Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$ be an MDP without reward and let π^E be the expert's policy, with subopti-*

3. To be precise, previous works (Metelli et al., 2021, 2023; Poiani et al., 2024) adopt more restrictive definitions of feasible set, that assume the existence of the expert's policy at all (s, a, h) of the state-action space. Our definitions relax such requirement (Lazzati et al., 2024a).

mality in $[L, U]$ *under the true unknown reward* r^E*. The* feasible set $\mathcal{R}^{L,U}_{\mathcal{M},\pi^E}$ *of rewards compatible with* π^E *in* $\mathcal{M}$ *is defined as*

$$\mathcal{R}^{L,U}_{\mathcal{M},\pi^E} := \Big\{ r \in \mathfrak{R} \,\big|\, J^*(r;p) - J^{\pi^E}(r;p) \in [L,U] \Big\}. \tag{2}$$

Again, the notion of feasible set represents the set of rewards containing the true reward r^E. Note that $\mathcal{R}^{0,0}_{\mathcal{M},\pi^E} = \mathcal{R}_{\mathcal{M},\pi^E}$, and that $\mathcal{R}^{0,U}_{\mathcal{M},\pi^E} \supseteq \mathcal{R}_{\mathcal{M},\pi^E}$ for any $U \geqslant 0$. Moreover, observe that the setting with $L = 0, U > 0$ coincides with the setting of Poiani et al. (2024).

3.2 Learning the Feasible Set

In practical applications, the dynamics (d_0, p) and the expert's policy π^E are not known, but have to be estimated from samples. In tabular MDPs, in the simple setting with a generative model for both p and π^E, we have algorithms for learning the feasible set that require a number of calls to the generative model (i.e., sample complexity) that grows at most *quadratically* in the size of the state space. For instance, algorithm US-IRL of Metelli et al. (2023) can compute accurate estimates of (a variant of) the feasible set in Definition 1 with high probability (w.h.p.), while, for the setting with a suboptimal expert, algorithm US-IRL-SE of Poiani et al. (2024) can compute accurate estimates of (a variant of) the feasible set in Definition 2 w.h.p..

These algorithms are efficient as long as the cardinality of the state space is reasonably small, but when the state space is large, these algorithms can be very sample *inefficient* because of the quadratic dependence. Unfortunately, without any structural assumption on the MDP, no efficient algorithm can be developed when the state space is large, even using the simple generative model, due to the lower bounds to the sample complexity presented in Metelli et al. (2023) (optimal expert) and Poiani et al. (2024) (suboptimal expert), that grow directly with the size of the state space.[4]

For this reason, similarly to what is done in (forward) RL in environments with large or even continuous state spaces, we make the simple but strong structural assumption that the environment is a Linear MDP, and we try to design sample efficient algorithms. Specifically, we assume the existence of a feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, with $\|\phi(s,a)\|_2 \leqslant 1$ for every $(s,a) \in \mathcal{S} \times \mathcal{A}$, such that the considered MDP without reward $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$ has a dynamics that is linear in ϕ, i.e., $p_h(\cdot|s,a) = \langle \phi(s,a), \mu_h(\cdot) \rangle$ for all $h \in [\![H]\!]$, where $\mu_h = [\mu_h^1, \dots, \mu_h^d]^\intercal$ is a vector of d signed measures such that $\| |\mu_h|(\mathcal{S}) \|_2 \leqslant \sqrt{d}$. Linear MDPs also assume linear rewards. Thus, we consider the true expert's reward r^E to be linear in the feature mapping $r_h(s,a) = \langle \phi(s,a), \theta_h \rangle$ for all $h \in [\![H]\!]$, for some parameters θ_h s.t. $\|\theta_h\|_2 \leqslant \sqrt{d}$. For simplicity, let us define the set of rewards that are linear in ϕ as $\mathfrak{R}_\phi$:

$$\mathfrak{R}_\phi := \Big\{ r \in \mathfrak{R} \,\big|\, \forall h \in [\![H]\!]\, \exists \theta_h \in \mathbb{R}^d :\, \|\theta_h\|_2 \leqslant \sqrt{d} \;\wedge\; \forall (s,a) \in \mathcal{S} \times \mathcal{A} :\, r_h(s,a) = \langle \phi(s,a), \theta_h \rangle \Big\}.$$

We denote a Linear MDP without reward as $\mathcal{M}_\phi = (\mathcal{S}, \mathcal{A}, H, d_0, p, \phi)$.

4. Again, to be precise, the lower bounds address a definition of feasible set that is slightly different than Definition 1 and 2.

Since the Linear MDP assumption gives additional information on r^E, then our definitions of feasible set change. In particular, they contain only rewards that are linear in ϕ:

Definition 3 (Feasible Set for Optimal Expert in Linear MDPs)**.** *Let $\mathcal{M}_\phi$ be a Linear MDP without reward, with $\mathcal{M}_\phi = (\mathcal{S}, \mathcal{A}, H, d_0, p, \phi)$, and let π^E be the expert's policy, that is optimal for the true unknown reward r^E. The* feasible set $\mathcal{R}_{\mathcal{M}_\phi, \pi^E}$ *of rewards compatible with π^E in $\mathcal{M}_\phi$ is defined as*

$$\mathcal{R}_{\mathcal{M}_\phi, \pi^E} \coloneqq \Big\{ r \in \mathfrak{R}_\phi \,\Big|\, J^{\pi^E}(r; p) = J^*(r; p) \Big\}. \tag{3}$$

Definition 4 (Feasible Set for Suboptimal Expert in Linear MDPs)**.** *Let $U \geqslant L \geqslant 0$. Let $\mathcal{M}_\phi = (\mathcal{S}, \mathcal{A}, H, d_0, p, \phi)$ be a Linear MDP without reward and let π^E be the expert's policy, with suboptimality in $[L, U]$ under the true unknown reward r^E. The* feasible set $\mathcal{R}^{L,U}_{\mathcal{M}_\phi, \pi^E}$ *of rewards compatible with π^E in $\mathcal{M}_\phi$ is defined as*

$$\mathcal{R}^{L,U}_{\mathcal{M}_\phi, \pi^E} \coloneqq \Big\{ r \in \mathfrak{R}_\phi \,\Big|\, J^*(r; p) - J^{\pi^E}(r; p) \in [L, U] \Big\}. \tag{4}$$

Before discussing whether efficient IRL algorithms can be developed for learning the feasible set in Linear MDPs, we formalize what we mean by "efficient algorithm" in the simple setting with a generative model for p and π^E. The following definition customizes the general notion of PAC algorithm for IRL (e.g., see Metelli et al., 2023) to the Linear MDPs setting:

Definition 5 (PAC Algorithm)**.** *An algorithm $\mathfrak{A}$ is a PAC algorithm for the IRL setting with* optimal expert *w.r.t. a metric ρ between sets of rewards if, for any $\epsilon, \delta \in (0, 1)$, for any Linear MDP without reward $\mathcal{M}_\phi$ and expert's policy π^E, the algorithm $\mathfrak{A}$ outputs set $\widehat{\mathcal{R}}$ such that*

$$\mathop{\mathbb{P}}_{(\mathcal{M}_\phi, \pi^E), \mathfrak{A}} \Big(\rho\big(\widehat{\mathcal{R}}, \mathcal{R}_{\mathcal{M}_\phi, \pi^E}\big) \leqslant \epsilon \Big) \geqslant 1 - \delta,$$

where $\mathbb{P}_{(\mathcal{M}_\phi, \pi^E), \mathfrak{A}}$ denotes the probability measure induced by executing $\mathfrak{A}$ in the Linear IRL problem $(\mathcal{M}_\phi, \pi^E)$, and $\widehat{\mathcal{R}}$ is an estimate of the feasible set computed by $\mathfrak{A}$ using τ calls to the generative model of p and τ^E calls to the generative model of π^E.
We define a PAC algorithm for the IRL setting with $[L, U]-$suboptimal expert *(with arbitrary $U \geqslant L \geqslant 0$) analogously.*

We say that an IRL algorithm is *efficient* if it is PAC with values of τ and τ^E that do not depend on the cardinality of the state space.

Intuitively, Linear MDPs provide structure to the transition model p (and the reward function), but *not to the expert's policy* π^E. If we assume that π^E is known at all (s, h), then, it is possible to design a PAC algorithm for the optimal expert setting (Eq. (3)) whose sample complexity is independent of the size of the state space (see Algorithm 1 of Lazzati et al. (2024b)). However, if π^E is unknown, we prove that no efficient algorithm can be developed, i.e., in the worst case, any PAC algorithm has to collect a number of samples that depends on the size of the state space.

Theorem 1 (Statistical Inefficiency for Optimal Expert)**.** *Let $\mathfrak{A}$ be a PAC algorithm for the IRL setting with optimal expert. Then, there exists a Linear MDP without reward $\mathcal{M}_\phi$ with a state space with finite but arbitrarily large cardinality S, and a deterministic expert's policy π^E, in which $\mathfrak{A}$ requires at least $\tau^E \geqslant S$ calls to the generative model of π^E.*

Proof Let $\mathfrak{A}$ be a PAC algorithm for the setting with optimal expert, and let ρ be any metric between sets of rewards with respect to which $\mathfrak{A}$ is PAC. Let $\mathcal{M}_\phi = (\mathcal{S}, \mathcal{A}, H, d_0, p, \phi)$ be a Linear MDP without reward, where $\mathcal{S} = \{s_1, s_2, \dots, s_S\}$ is the finite state space, $\mathcal{A} = \{a_1, a_2\}$ is the action space, $H = 1$ is the horizon, $d_0(s) > 0 \ \forall s \in \mathcal{S}$ is the initial state distribution, and $\phi(s,a) = \mathbb{1}\{a = a_1\} \ \forall (s,a) \in \mathcal{S} \times \mathcal{A}$ is the feature map. Since $H = 1$, we do not have to specify p.

We construct a family $\mathbb{M} = \big((\mathcal{M}_\phi, \pi^{E,0}), (\mathcal{M}_\phi, \pi^{E,1}), \dots, (\mathcal{M}_\phi, \pi^{E,S})\big)$ of $S+1$ IRL problem instances that share the same environment $\mathcal{M}_\phi$, but differ in the expert's policy. Specifically, we define the deterministic policy $\pi^{E,0}$ that always plays action a_1: $\pi_1^{E,0}(a_1|s) = 1 \, \forall s \in \mathcal{S}$. For all $i \in [\![S]\!]$, we define the deterministic policy $\pi^{E,i}$ that plays action a_1 at all states except state s_i: $\pi_1^{E,i}(a_1|s) = 1 \, \forall s \in \mathcal{S} \backslash \{s_i\}$, $\pi_1^{E,i}(a_2|s_i) = 1$.

By direct calculation, the feasible sets of the various problem instances are:

$$\mathcal{R}_{\mathcal{M}_\phi, \pi^{E,0}} = \Big\{ r \in \mathfrak{R} \,\Big|\, \forall s \in \mathcal{S} : \, r_1(s, a_1) \in [0,1] \ \wedge \ r_1(s, a_2) = 0 \Big\},$$
$$\mathcal{R}_{\mathcal{M}_\phi, \pi^{E,1}} = \ldots = \mathcal{R}_{\mathcal{M}_\phi, \pi^{E,S}} = \Big\{ r \in \mathfrak{R} \,\Big|\, \forall (s,a) \in \mathcal{S} \times \mathcal{A} : \, r_1(s,a) = 0 \Big\}.$$

Since ρ is a metric and $\mathcal{R}_{\mathcal{M}_\phi, \pi^{E,0}} \neq \mathcal{R}_{\mathcal{M}_\phi, \pi^{E,1}}$, then there exists a constant $c > 0$ such that

$$\rho\Big(\mathcal{R}_{\mathcal{M}_\phi, \pi^{E,0}}, \mathcal{R}_{\mathcal{M}_\phi, \pi^{E,1}}\Big) \geqslant c.$$

We proceed by contradiction. Assume that $\mathfrak{A}$ is a PAC algorithm that requires less than $\tau^E < S$ calls to the generative model of the expert's policy in all possible IRL problem instances. When $\mathfrak{A}$ tackles problem instance $(\mathcal{M}_\phi, \pi^{E,0})$, irrespective of the randomization method it uses to choose the states from which collecting samples, since it collects less than S samples, then there is at least one state s_j with $j \in [\![S]\!]$ in which $\mathfrak{A}$ does not know whether the expert's policy plays a_1 or a_2. Therefore, $\mathfrak{A}$ cannot distinguish between $\pi^{E,0}$ and $\pi^{E,j}$, and thus it cannot know whether the true feasible set is $\mathcal{R}_{\mathcal{M}_\phi, \pi^{E,0}}$ or $\mathcal{R}_{\mathcal{M}_\phi, \pi^{E,j}} = \mathcal{R}_{\mathcal{M}_\phi, \pi^{E,1}}$. Nevertheless, by hypothesis, being $\mathfrak{A}$ a PAC algorithm, it must output a set $\widehat{\mathcal{R}}$ such that, for any $\epsilon, \delta \in (0,1)$:

$$\mathop{\mathbb{P}}_{(\mathcal{M}_\phi, \pi^{E,0}), \mathfrak{A}}\left(\Big\{\rho\big(\mathcal{R}_{\mathcal{M}_\phi, \pi^{E,0}}, \widehat{\mathcal{R}}\big) \leqslant \epsilon\Big\}\right) \geqslant 1 - \delta \ \wedge \ \mathop{\mathbb{P}}_{(\mathcal{M}_\phi, \pi^{E,0}), \mathfrak{A}}\left(\Big\{\rho\big(\mathcal{R}_{\mathcal{M}_\phi, \pi^{E,1}}, \widehat{\mathcal{R}}\big) \leqslant \epsilon\Big\}\right) \geqslant 1 - \delta.$$

In particular, for the choice of accuracy $\epsilon < c/2$ and failure probability $\delta < 1/2$, we have

$$\mathop{\mathbb{P}}_{(\mathcal{M}_\phi, \pi^{E,0}), \mathfrak{A}}\left(\Big\{\rho\big(\mathcal{R}_{\mathcal{M}_\phi, \pi^{E,0}}, \widehat{\mathcal{R}}\big) < \frac{c}{2}\Big\}\right) > \frac{1}{2} \ \wedge \ \mathop{\mathbb{P}}_{(\mathcal{M}_\phi, \pi^{E,0}), \mathfrak{A}}\left(\Big\{\rho\big(\mathcal{R}_{\mathcal{M}_\phi, \pi^{E,1}}, \widehat{\mathcal{R}}\big) < \frac{c}{2}\Big\}\right) > \frac{1}{2}.$$

However, this results in a contradiction:

$$1 = \frac{1}{2} + \frac{1}{2}$$

$$< \mathop{\mathbb{P}}_{(\mathcal{M}_\phi,\pi^{E,0}),\mathfrak{A}}\left(\left\{\rho\big(\mathcal{R}_{\mathcal{M}_\phi,\pi^{E,0}},\widehat{\mathcal{R}}\big) < \frac{c}{2}\right\}\right) + \mathop{\mathbb{P}}_{(\mathcal{M}_\phi,\pi^{E,0}),\mathfrak{A}}\left(\left\{\rho\big(\mathcal{R}_{\mathcal{M}_\phi,\pi^{E,1}},\widehat{\mathcal{R}}\big) < \frac{c}{2}\right\}\right)$$
$$\overset{(1)}{=} \mathop{\mathbb{P}}_{(\mathcal{M}_\phi,\pi^{E,0}),\mathfrak{A}}\left(\left\{\rho\big(\mathcal{R}_{\mathcal{M}_\phi,\pi^{E,0}},\widehat{\mathcal{R}}\big) < \frac{c}{2}\right\} \cup \left\{\rho\big(\mathcal{R}_{\mathcal{M}_\phi,\pi^{E,1}},\widehat{\mathcal{R}}\big) < \frac{c}{2}\right\}\right),$$

where at (1) we use that the two events are disjoint because ρ satisfies the triangle inequality. Since the probability of no event can be larger than 1, we get a contradiction that demonstrates the claim of the theorem. ■

Theorem 2 (Statistical Inefficiency for Suboptimal Expert)**.** *Let $0 \leqslant L \leqslant U < 1$ be arbitrary. Let $\mathfrak{A}$ be a PAC algorithm for the IRL setting with $[L,U]$-suboptimal expert. Then, there exists a Linear MDP without reward $\mathcal{M}_\phi$ with a state space with finite but arbitrarily large cardinality S, and a deterministic expert's policy π^E, in which $\mathfrak{A}$ requires at least $\tau^E \geqslant S$ calls to the generative model of π^E.*

Proof Let $0 \leqslant L \leqslant U < 1$ be arbitrary. Let $\mathfrak{A}$ be a PAC algorithm for the setting with $[L,U]$-suboptimal expert, and let ρ be any metric between sets of rewards with respect to which $\mathfrak{A}$ is PAC. We consider the same family $\mathbb{M}$ of IRL problem instances used in the proof of Theorem 1. Let us compute the feasible sets for these problems.

By definition of Linear MDPs, the reward parameter θ belongs to $[-1,+1]$. For all $\theta \in [-1,+1]$, we denote by $\Delta_0(\theta) \coloneqq J^*(r_\theta;p) - J^{\pi^{E,0}}(r_\theta;p)$ the suboptimality of $\pi^{E,0}$ in $\mathcal{M}_\phi$ using reward r_θ defined as $r_\theta(s,a) = \langle \phi(s,a),\theta\rangle \ \forall s,a$. Similarly, for all $i \in [\![S]\!]$, we denote by $\Delta_i(\theta) \coloneqq J^*(r_\theta;p) - J^{\pi^{E,i}}(r_\theta;p)$ the suboptimality of $\pi^{E,i}$ in $\mathcal{M}_\phi$ using reward r_θ defined previously. We distinguish three cases:

- If $\theta \in [-1,0) \rightarrow$ then $\Delta_0(\theta) = -\theta$ and $\Delta_i(\theta) = -\frac{S-1}{S}\theta \ \forall i \in [\![S]\!]$.
- If $\theta = 0 \rightarrow$ then $\Delta_0(\theta) = \Delta_i(\theta) = 0 \ \forall i \in [\![S]\!]$.
- If $\theta \in (0,+1] \rightarrow$ then $\Delta_0(\theta) = 0$ and $\Delta_i(\theta) = \frac{\theta}{S} \ \forall i \in [\![S]\!]$.

Based on these expressions, we can compute the feasible sets. When $L = 0$, the values of $\theta \in [-1,+1]$ that make, respectively, $\Delta_0(\theta)$ and $\Delta_i(\theta)$ in $[0,U]$ for all $i \in [\![S]\!]$ are:

$$\begin{aligned}
\mathcal{R}^{0,U}_{\mathcal{M}_\phi,\pi^{E,0}} &= \Big\{r \in \mathfrak{R} \,\Big|\, \exists \theta \in [-U,+1] : \ \forall (s,a) \in \mathcal{S}\times\mathcal{A} : \ r(s,a) = \langle \phi(s,a),\theta\rangle\Big\} \\
&= \Big\{r \in \mathfrak{R} \,\Big|\, \forall s \in \mathcal{S} : \ r_1(s,a_1) \in [-U,1] \ \wedge \ r_1(s,a_2) = 0\Big\}, \\
\mathcal{R}^{0,U}_{\mathcal{M}_\phi,\pi^{E,i}} &= \Big\{r \in \mathfrak{R} \,\Big|\, \exists \theta \in [\max\{-\frac{S}{S-1}U,-1\},\min\{SU,+1\}] : \\
&\qquad\qquad \forall (s,a) \in \mathcal{S}\times\mathcal{A} : \ r(s,a) = \langle \phi(s,a),\theta\rangle\Big\} \\
&= \Big\{r \in \mathfrak{R} \,\Big|\, \forall s \in \mathcal{S} : \ r_1(s,a_1) \in [\max\{-\tfrac{S}{S-1}U,-1\},\min\{SU,+1\}] \ \wedge \ r_1(s,a_2) = 0\Big\}.
\end{aligned}$$

Note that these sets are different for all $U \in [0,1), S \geqslant 2$, thus we can proceed analogously as in the proof of Theorem 1 to demonstrate the result. When $L \neq 0$, we have, for all

$i \in [\![S]\!]$:

$$\begin{aligned}
\mathcal{R}^{L,U}_{\mathcal{M}_\phi,\pi^{E,0}} &= \Big\{ r \in \mathfrak{R} \,\Big|\, \exists \theta \in [-U,-L] : \forall (s,a) \in \mathcal{S} \times \mathcal{A} : \, r(s,a) = \langle \phi(s,a), \theta \rangle \Big\} \\
&= \Big\{ r \in \mathfrak{R} \,\Big|\, \forall s \in \mathcal{S} : \, r_1(s,a_1) \in [-U,-L] \, \wedge \, r_1(s,a_2) = 0 \Big\}, \\
\mathcal{R}^{L,U}_{\mathcal{M}_\phi,\pi^{E,i}} &= \Big\{ r \in \mathfrak{R} \,\Big|\, \exists \theta \in [\max\{-\frac{S}{S-1}U,-1\}, \max\{-\frac{S}{S-1}L,-1\}] \\
&\qquad \cup [\min\{SL,1\}, \min\{SU,1\}] : \forall (s,a) \in \mathcal{S} \times \mathcal{A} : \, r(s,a) = \langle \phi(s,a), \theta \rangle \Big\} \\
&= \Big\{ r \in \mathfrak{R} \,\Big|\, \forall s \in \mathcal{S} : \, r_1(s,a_1) \in [\max\{-\frac{S}{S-1}U,-1\}, \max\{-\frac{S}{S-1}L,-1\}] \\
&\qquad \cup \, [\min\{SL,1\}, \min\{SU,1\}] \, \wedge \, r_1(s,a_2) = 0 \Big\}.
\end{aligned}$$

Again, note that these sets are different for all $L \in (0,1), U \in [L,1)$, thus we can proceed analogously as in the proof of Theorem 1 to demonstrate the result. This concludes the proof. ■

In words, the theorems above show that, even under the easiest learning conditions (i.e., generative model and deterministic expert), and even imposing the strong structure provided by Linear MDPs, the sample complexity of learning the feasible set scales directly with the cardinality of the state space S, making it is infeasible when S is large or even infinite.

Of course, note that the feasible set is not efficiently learnable if we impose structural assumptions milder than those imposed by Linear MDPs. For instance, consider the families of problems with low Bellman rank (Jiang et al., 2017), low Eluder dimension (Wang et al., 2020b), or low Bellman Eluder dimension (Jin et al., 2021a). As shown by Jin et al. (2021a), all of them contain the family of Linear MDPs, therefore, the structure imposed to the reward function, and, thus, to the feasible set, is *at most* as strong as that provided by Linear MDPs (see Eq. (3) and (4)) and at least as strong as no structure (see Eq. (1) and (2)). An analogous statement can be made for the transition model. For these reasons, if we adapt Definition 5 to these settings, we obtain that, in the IRL setting with optimal or suboptimal expert, any PAC algorithm for learning the feasible set requires at least $\tau^E \geqslant S$ calls to the generative model of π^E, because this is the minimum number of samples required with Linear MDPs (see Theorem 1 and Theorem 2).

We remark that the same negative results hold also for other function approximation settings, like Linear Mixture MDPs (Zhou et al., 2021):

Theorem (Informal) 3 (Statistical Inefficiency - Linear Mixture MDPs)**.** *In the IRL setting with optimal or suboptimal expert, any PAC algorithm for learning the feasible set in Linear Mixture MDPs requires at least $\tau^E \geqslant S$ calls to the generative model of π^E.*

Proof Sketch First, we observe that Linear Mixture MDPs do not impose structure on the reward function, thus the feasible sets coincide with those in Definition 1 and 2. Next, we construct problem instances analogous to those constructed in the proofs of Theorem 1 and Theorem 2, without the need of choosing specific values for the feature mapping. Finally, we recognize that if the learning algorithm does not "know" the expert's action in

a state, then it does not know which action in that state should have the larger reward, thus ending up in an estimate with finite error, that can be reduced only discovering the expert's action at all states, i.e., with $\tau^E = S$ samples (more than this if π^E is stochastic). The proof can be formally concluded by contradiction as we did in Theorem 1. ■

To sum up, intuitively, learning the feasible set efficiently requires the exact knowledge of (the support of) the expert's policy in the entire state space. While collecting (at least) one sample from each state is feasible in tabular MDPs, and there are many efficient algorithms doing so (Metelli et al., 2023; Zhao et al., 2024; Poiani et al., 2024), this becomes clearly infeasible when the state space is large. Due to the results presented above, i.e., that the common function approximation settings that permit the development of efficient algorithms for (forward) RL in problems with large state spaces fail to do so for IRL, then we can conclude that *sample-efficient IRL requires structural assumptions stronger than those required by sample-efficient RL*, as long as the solution concept considered for IRL is the notion of feasible set. In addition, we remark that, irrespective of the specific problem setting considered (e.g., optimal/suboptimal expert), the feasible set contains an *infinite* amount of reward functions, thus it is not possible to practically implementing an algorithm that outputs *all* the rewards in the feasible set. This fact limits the framework of the feasible set as a mere theoretical tool.

4. The Reward Compatibility Framework

In this section, we present the main contribution of this work: *Reward Compatibility*, a novel framework for IRL that allows us to conveniently rephrase the learning from demonstrations problem as a *classification* task. More specifically, to overcome the two issues highlighted in Section 3:

- We introduce the notion of *reward compatibility* to permit sample-efficient IRL without additional structural assumptions. This framework replaces the feasible set as solution concept for IRL.
- We reformulate IRL as a *classification problem* to allow the practical development of efficient algorithms. An algorithm will take in input a single reward and will output a single boolean.

4.1 Reward Compatibility

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$ be an MDP without reward and let π^E be the expert's policy.[5] Let $U \geqslant L \geqslant 0$ be arbitrary. As described in the previous section, the feasible set for optimal expert $\mathcal{R}_{\mathcal{M},\pi^E}$ (Definition 1) and the feasible set for suboptimal expert $\mathcal{R}^{L,U}_{\mathcal{M},\pi^E}$ (Definition 2) contain *all and only* the reward functions that *exactly* comply with the given constraint, i.e., respectively, that make π^E optimal, and that make the suboptimality of π^E inside $[L, U]$. In other words, we can interpret the feasible set as carrying out a *binary classification* of rewards based on whether they satisfy a given "hard" constraint, as shown in Figure 1 (Left). We say that the rewards inside the feasible set are *compatible* with

5. For simplicity, we consider a tabular MDP. Nevertheless, the presentation is independent of structural assumptions of the MDP (e.g., Linear MDP).

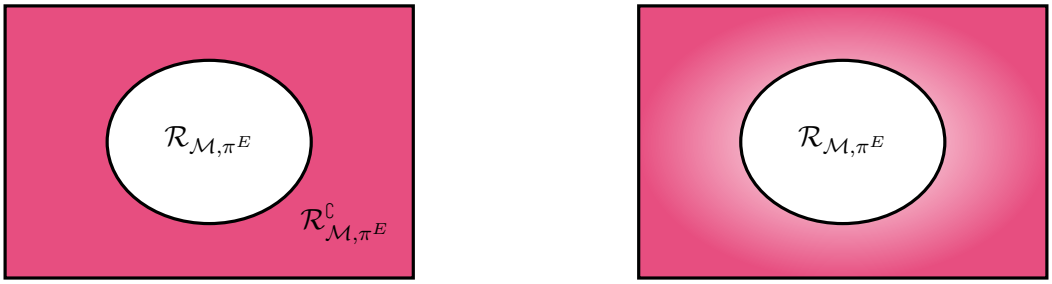


Figure 1: (Left) In the framework of the feasible set, all the rewards in $\mathcal{R}^{\complement}_{\mathcal{M},\pi^E}$ (in pink), i.e., outside the feasible set $\mathcal{R}_{\mathcal{M},\pi^E}$ (in white), are considered equally wrong. (Right) In the framework of the reward compatibility, the rewards outside the feasible set $\mathcal{R}_{\mathcal{M},\pi^E}$ suffer from different errors (scale of pink, with darker colors corresponding to larger non-compatibility values).

the demonstrations of π^E (Metelli et al., 2023). However, our insight is that some rewards outside the feasible set are *"more" compatible* with π^E than others, as shown in Figure 1 (Right) and in the following example (a similar example may be constructed also for the suboptimal expert):

Example 1. *Consider an MDP without reward $\mathcal{M}$ with one state and horizon one in which the expert has three actions: eating a muffin (M), a cake (C), or a soup (S). Assume that the expert, assumed optimal, demonstrates action $\pi^E = M$, i.e., it eats the muffin. Clearly, the rewards r_1, r_2 defined as follows:*

$$r_1(a) = \begin{cases} +0.99 & \text{if } a = M \\ +1 & \text{if } a = C \\ 0 & \text{if } a = S \end{cases}, \qquad r_2(a) = \begin{cases} 0 & \text{if } a = M \\ +1 & \text{if } a = C \\ 0 & \text{if } a = S \end{cases},$$

do not belong to the feasible set $\mathcal{R}_{\mathcal{M},\pi^E}$, because they do not make "eating the muffin" the optimal strategy. However, intuitively, r_1 is "more" compatible *with π^E than r_2, because it makes M a very good action, while reward r_2 makes it very bad. Clearly, we make a larger error if we model the preferences of the expert with r_2 instead of r_1. However, the feasible set $\mathcal{R}_{\mathcal{M},\pi^E}$ is blind to the difference between r_1 and r_2, and it "refuses" both of them.*

In the setting with optimal expert, we propose the following "soft" definition of compatibility of a reward with demonstrations to capture this intuition. To be precise, since the larger the quantity the smaller the compatibility, we talk of (non)compatibility.

Definition 6 (Reward (non)Compatibility - Optimal Expert)**.** *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$ be an MDP without reward and let π^E be the expert's policy. For any reward $r \in \mathfrak{R}$, we define the* (non)compatibility $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r)$ *of r with π^E in $\mathcal{M}$ as*

$$\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r) \coloneqq J^*(r;p) - J^{\pi^E}(r;p). \tag{5}$$

In words, $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r)$ quantifies the *suboptimality* of π^E in the MDP obtained by $\mathcal{M}$ and r. Since, by definition of the setting, the true expert's reward r^E satisfies $J^*(r^E;p) - J^{\pi^E}(r^E;p) = 0$, then it is clear that $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r)$ measures the degree to which reward r fulfils

this constraint. Observe that the feasible set $\mathcal{R}_{\mathcal{M},\pi^E}$ can be seen as the set of rewards with zero (non)compatibility, i.e., $\mathcal{R}_{\mathcal{M},\pi^E} = \{r \in \mathfrak{R} \,|\, \overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r) = 0\}$. Since the smaller the $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r)$ the more r is compatible with π^E, and since the larger the $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r)$ the less r is compatible with π^E, we say that $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}$ quantifies the *(non)*compatibility. Visually, Figure 1 (Right) represents function $\overline{\mathcal{C}}_{\mathcal{M},\pi^E} : \mathfrak{R} \to [0, +\infty]$ using different magnitudes of the pink color. Rewards r with zero $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r)$ (i.e., inside the feasible set) are pictured in white, while rewards r with larger values of $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r)$ are pictured with larger magnitudes of pink.

Example 1 (continuing from p. 13)**.** *Reward (non)compatibility permits to discriminate between r_1 and r_2. Indeed, we have that $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r_1) = 0.01$, $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r_2) = 1$. Since reward r_1 suffers from a smaller (non)compatibility than r_2, i.e., $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r_1) < \overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r_2)$, then we have that r_1 is more compatible with π^E than r_2, as expected.*

By definition of IRL, the observed expert's policy π^E does not reveal any information about the other policies. Thus, it is meaningful that $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}$ considers the suboptimality of π^E only, as illustrated below.

Example 2. *Consider now reward r_1', defined below, and compare it with r_1.*

$$r_1'(a) = \begin{cases} +0.99 & \text{if } a = M \\ 0 & \text{if } a = C \\ +1 & \text{if } a = S \end{cases}.$$

The optimal policy under r_1' is to play S, while under r_1 we would play C. However, demonstrations from π^E alone do not provide information on C or S, but only about $\pi^E = M$, therefore it is meaningful that r_1 and r_1' are equally compatible $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r_1') = \overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r_1) = 0.01$ with the given demonstrations.

In an analogous manner, we can define a notion of reward compatibility for the setting with suboptimal expert:

Definition 7 (Reward (non)Compatibility - Suboptimal Expert)**.** *Let $U \geqslant L \geqslant 0$ be arbitrary. Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$ be an MDP without reward and let π^E be the expert's policy. For any reward $r \in \mathfrak{R}$, we define the* (non)compatibility $\overline{\mathcal{C}}^{L,U}_{\mathcal{M},\pi^E}(r)$ *of r with π^E in $\mathcal{M}$ based on L, U as*

$$\overline{\mathcal{C}}^{L,U}_{\mathcal{M},\pi^E}(r) \coloneqq \min_{x \in [L,U]} \left| \left(J^*(r;p) - J^{\pi^E}(r;p) \right) - x \right|. \tag{6}$$

Simply put, $\overline{\mathcal{C}}^{L,U}_{\mathcal{M},\pi^E}(r)$ measures how far from the interval $[L, U]$ lies the suboptimality of the expert's policy π^E w.r.t. r in $\mathcal{M}$. In other words, $\overline{\mathcal{C}}^{L,U}_{\mathcal{M},\pi^E}(r)$ quantifies the extent to which reward r fulfils the constraint $J^*(\cdot;p) - J^{\pi^E}(\cdot;p) \in [L, U]$, that defines the setting. Note that, again, the feasible set can be seen as the set of rewards with zero (non)compatibility $\mathcal{R}^{L,U}_{\mathcal{M},\pi^E} = \{r \in \mathfrak{R} \,|\, \overline{\mathcal{C}}^{L,U}_{\mathcal{M},\pi^E}(r) = 0\}$.

4.2 The IRL Classification Formulation

In practical applications, we are often interested in understanding whether some designed or computed reward function represents the preferences of the observed expert agent. Beyond not being practically implementable, as mentioned in Section 3, an algorithm that outputs the feasible set would not even permit to search efficiently for specific rewards inside the set. For these reasons, we reformulate IRL as a classification problem as follows.

Definition 8 (IRL Classification Problem)**.** *An* IRL Classification Problem *instance is a tuple* $(\mathcal{M}, \pi^E, \mathcal{R}, \Delta)$*, where* $\mathcal{M}$ *is an MDP without reward,* π^E *is the expert's policy,* $\mathcal{R} \subseteq \mathfrak{R}$ *is the set of rewards to classify, and* $\Delta \in \mathbb{R}_{\geqslant 0}$ *is a threshold. The goal is to classify the rewards* $r \in \mathcal{R}$ *based on a given notion of (non)compatibility* $\overline{\mathcal{C}} : \mathfrak{R} \to \mathbb{R}_{\geqslant 0}$*:*

$$\forall r \in \mathcal{R}: \quad \textbf{if } \overline{\mathcal{C}}(r) \leqslant \Delta \textbf{ then } \quad \textbf{return } \text{True}, \quad \textbf{else return } \text{False}.$$

Definition 9 (IRL Algorithm)**.** *An* IRL algorithm *takes in input a reward* $r \in \mathcal{R}$ *and outputs a boolean saying whether* $\overline{\mathcal{C}}(r) \leqslant \Delta$*.*

In words, in an IRL classification problem we aim to classify a given set of reward functions $\mathcal{R}$ in two classes, depending on their (non)compatibility with a given expert's policy π^E in an environment $\mathcal{M}$. Specifically, one class represents rewards that are compatible with π^E, while the other class describes rewards that are not. Whether a (non)compatibility value is considered small or large depends on the classification threshold Δ. The role of the threshold Δ is that of a problem parameter, whose choice, in practice, may be dictated by how much we perceive the model of (sub-)optimal expert as correct or not—we can choose its intensity based on the amount of misspecification of the model. Moreover, note that the definitions above are general, and the notion of (non)compatibility $\overline{\mathcal{C}}$ adopted depends on the specific IRL setting considered. In particular, they can be applied to settings other than IRL with optimal (Ng and Russell, 2000) or suboptimal (Poiani et al., 2024) expert, like maximum entropy IRL (Ziebart et al., 2008) or Bayesian IRL (Ramachandran and Amir, 2007), as long as a suitable definition of (non)compatibility is provided. Observe that we allow for $\mathcal{R} \neq \mathfrak{R}$ to manage scenarios in which, for instance, we have some prior knowledge on r^E, i.e., $r^E \in \mathcal{R} \subset \mathfrak{R}$, and therefore we want to "update" our knowledge with the observed demonstrations of π^E without wasting time or computational power with the non-interesting rewards outside $\mathcal{R}$.

Relation with the feasible set. How does the solution concept/learning target of the IRL classification formulation relate to the feasible set? Let us denote by $\mathcal{R}^\star := \{r \in \mathfrak{R} \,|\, \overline{\mathcal{C}}(r) = 0\}$ the feasible set for the considered IRL setting. Moreover, let $\mathcal{R} = \mathfrak{R}$ (i.e., we have to classify all the rewards) and define the set of rewards positively classified by an IRL algorithm as $\mathcal{R}_\Delta$, i.e., $\mathcal{R}_\Delta := \{r \in \mathfrak{R} \,|\, \overline{\mathcal{C}}(r) \leqslant \Delta\}$. Then, it is clear that, for any $\Delta \geqslant 0$:

$$\mathcal{R}^\star = \mathcal{R}_0 \subseteq \mathcal{R}_\Delta.$$

In words, if we choose $\Delta = 0$, then we are simply classifying as positive all and only the rewards inside the feasible set, while if we consider $\Delta > 0$ strictly, then we are enlarging the feasible set. The situation is exemplified in Figure 2.

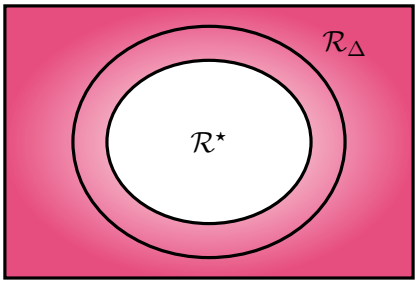


Figure 2: The set of rewards positively classified by an IRL algorithm $\mathcal{R}_\Delta$ with $\Delta > 0$ represents an enlargement of the feasible set $\mathcal{R}^\star$, i.e., $\mathcal{R}^\star \subseteq \mathcal{R}_\Delta$. Visually, $\mathcal{R}_\Delta$ contains rewards outside $\mathcal{R}^\star$ whose magnitude of pink is not too intense.

Differently from the framework of the feasible set, as we shall see in the next section, it is possible to practically implement the new notion of IRL algorithm (Definition 9), with guarantees of sample efficiency even when the state space is large, making the same structural assumptions considered for (forward) RL. Intuitively, the reward compatibility framework permits the development of sample-efficient algorithms at the price of an enlargement of the feasible set controlled by the classification threshold Δ.

4.3 A Concrete Example

Consider the 5x5 gridworld in Fig. 3 (a), where the initial state is **S**, the horizon is 4, and there are four deterministic actions that bring the agent to its left/right/up/bottom cells. State **G** represents the goal state demonstrated by the expert π^E (i.e., the blue arrows represent the expert trajectory), while state **D** represents an alternative goal state. Consider a one-dimensional family of state-only stationary reward functions parameterized by $\alpha \in [0, 1]$:

$$r_\alpha(s) = (1 - \alpha)\mathbb{1}\{s = \mathbf{G}\} + \alpha\mathbb{1}\{s = \mathbf{D}\}.$$

Intuitively, the parameter α quantifies how much state **D** is preferred to **G**. Since the horizon is four timesteps and the dynamics is deterministic, we have that $J^{\pi^E}(r_\alpha) = 1 - \alpha$, $J^*(r_\alpha) = \max\{1 - \alpha, \alpha\}$, and $\overline{C}(r_\alpha) = \max\{0, 2\alpha - 1\}$. In this example, the feasible set (restricted to the family r_α) is made of all and only the rewards with $\alpha \leqslant 0.5$, as shown in Fig. 3 (b): intuitively, these are the values for which $J^*(r_\alpha) = J^{\pi^E}(r_\alpha)$, i.e., for which the expert is perfectly optimal (when state **G** is preferred to **D**). However, the "issue" with this formulation is that it treats $\alpha = 0.51$ and $\alpha = 1$ identically, i.e., both are rejected. Reward compatibility instead recognizes that the former makes the demonstrated behavior almost optimal, while the latter makes it maximally inappropriate. This is shown in Fig. 3 (c), where we see that the (non)compatibility increases (linearly in this case) as we escape from the feasible set region. In particular, note that by fixing Δ (to 0.1 in the example, see the dashed line), we can enlarge the feasible set in a controlled manner, by i.a. controlling the approximation error.

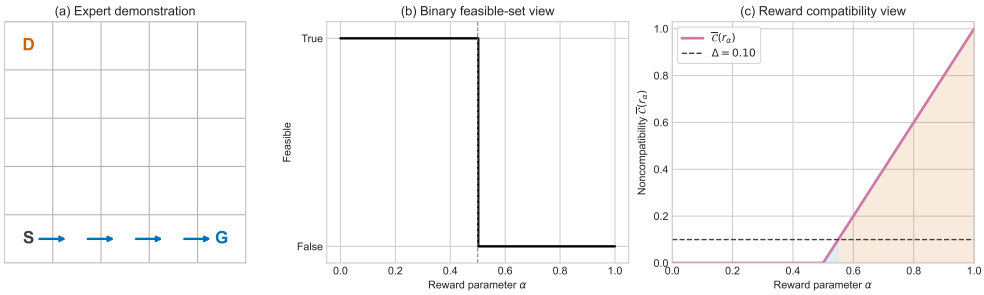


Figure 3: Illustrative example for reward compatibility.

5. Learning Compatible Rewards in the Online Setting

In this section, we first describe the online learning setting for IRL classification, and, then, we present an algorithm, CATY-IRL (CompATibilitY for IRL), that solves the task in a sample-efficient manner.

5.1 Problem Setting

In practical applications, the transition model p of the environment and the expert's policy π^E are not known, but have to be estimated from samples. We consider the following learning setting for the IRL classification problem, which is made of two phases: an exploration phase and a classification phase.

First, during the exploration phase, the learning algorithm receives as input an expert's data set $\mathcal{D}^E = \{(s_1^{E,i}, a_1^{E,i}, \dots, s_H^{E,i}, a_H^{E,i}, s_{H+1}^{E,i})\}_{i \in [\![\tau^E]\!]}$ of τ^E state-action trajectories collected by the expert's policy π^E in the (unknown) environment $\mathcal{M}$, and the set of rewards to classify $\mathcal{R}$, and it is allowed to explore the environment $\mathcal{M}$ at will to collect an exploration data set $\mathcal{D} = \{(s_1^i, a_1^i, \dots, s_H^i, a_H^i, s_{H+1}^i)\}_{i \in [\![\tau]\!]}$ of τ state-action trajectories.

Next, during the classification phase, the learning algorithm is not allowed to interact with the environment $\mathcal{M}$ anymore. It receives in input a reward $r \in \mathcal{R}$ and a threshold $\Delta \geqslant 0$, and it must classify r based on its (non)compatibility $\overline{\mathcal{C}}(r)$ and Δ. Since the transition model p of $\mathcal{M}$ and the expert's policy π^E are unknown, the learning algorithm has to use the data sets $\mathcal{D}$ and $\mathcal{D}^E$ to construct a meaningful estimate of (non)compatibility $\widehat{\mathcal{C}}(r) \approx \overline{\mathcal{C}}(r)$ for the classification.

5.2 Learning Framework

Observe that the considered learning setting represents the most common IRL setting in practical applications, where we are given a batch data set of expert demonstrations $\mathcal{D}^E$ and we can actively explore the environment to estimate its dynamics p. Intuitively, a learning algorithm is efficient in this context if it carries out an accurate classification of the input rewards with high probability, using the least amount of samples τ^E and τ possible. We formalize this concept as follows:

Definition 10 (PAC Algorithm - Online setting)**.** *Let $\epsilon > 0, \delta \in (0,1)$, and let $\mathcal{D}^E$ be a data set of τ^E expert's trajectories. An algorithm $\mathfrak{A}$ exploring for τ episodes is* (ϵ, δ)-PAC

for the IRL classification problem if:

$$\mathop{\mathbb{P}}_{\mathcal{M},\pi^E,\mathfrak{A}}\left(\sup_{r\in\mathcal{R}}\left|\overline{\mathcal{C}}(r)-\widehat{\mathcal{C}}(r)\right|\leqslant\epsilon\right)\geqslant 1-\delta,$$

where $\mathbb{P}_{\mathcal{M},\pi^E,\mathfrak{A}}$ *is the joint probability measure induced by* π^E *and* $\mathfrak{A}$ *in* $\mathcal{M}$*, and* $\widehat{\mathcal{C}}$ *is the estimate of some (non)compatibility notion* $\overline{\mathcal{C}}$ *computed by* $\mathfrak{A}$ *using* τ^E *and* τ *trajectories. The* sample complexity *is defined by the pair* (τ^E,τ)*.*

Again, observe that this definition is general and independent of the specific (non)compatibility notion $\overline{\mathcal{C}}$ adopted. We remark that, according to Definition 10, an algorithm is PAC depending on the accuracy with which it estimates the (non)compatibility of the various rewards in input, and not on the accuracy with which it classifies them. Thus, to understand why it is interesting to design PAC algorithms for the IRL classification problem based on Definition 10, we have to understand what is the accuracy of such an algorithm for that task.

To this aim, let us consider an IRL classification problem instance $(\mathcal{M},\pi^E,\mathcal{R},\Delta)$ with (non)compatibility $\overline{\mathcal{C}}$, and an (ϵ,δ)-PAC algorithm $\mathfrak{A}$ (for some $\epsilon>0,\delta\in(0,1)$), that computes estimates of (non)compatibility $\widehat{\mathcal{C}}$ and classifies the rewards based on some threshold $\eta\in\mathbb{R}$ potentially different from Δ.[6] We denote the set of rewards to classify with true (non)compatibility smaller than Δ as $\mathcal{R}_\Delta\coloneqq\{r\in\mathcal{R}\,|\,\overline{\mathcal{C}}(r)\leqslant\Delta\}$, and the set of rewards positively classified by algorithm $\mathfrak{A}$ as $\widehat{\mathcal{R}}_\eta\coloneqq\{r\in\mathcal{R}\,|\,\widehat{\mathcal{C}}(r)\leqslant\eta\}$. The main question is: *what is the relationship between* $\mathcal{R}_\Delta$ *and* $\widehat{\mathcal{R}}_\eta$?

By Definition 10, it is easy to see that, with probability at least $1-\delta$, it holds that

$$\widehat{\mathcal{R}}_{\Delta-\epsilon}\subseteq\mathcal{R}_\Delta\subseteq\widehat{\mathcal{R}}_{\Delta+\epsilon},$$

namely, with a careful choice of the classification threshold η adopted by the algorithm $\mathfrak{A}$, we can guarantee that, under the good event (i.e., with probability at least $1-\delta$), either *all* the rewards inside $\mathcal{R}_\Delta$, i.e., rewards with true (non)compatibility smaller than Δ, are classified correctly ($\eta\geqslant\Delta+\epsilon$), or *only* the rewards inside $\mathcal{R}_\Delta$, i.e., rewards with true (non)compatibility smaller than Δ, are classified correctly ($\eta\leqslant\Delta-\epsilon$). Thus, we can trade-off the amount of "false negatives"/"false positives" by choosing the threshold η. In particular, note that, if we choose $\eta=\Delta-\epsilon$ (to minimize the "false positives"), then we have the guarantee that $\widehat{\mathcal{R}}_\eta$ is not "too small":

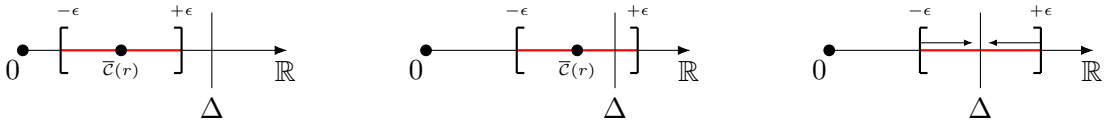
$$\mathcal{R}_{\Delta-2\epsilon}\subseteq\widehat{\mathcal{R}}_\eta\subseteq\mathcal{R}_\Delta.$$

Analogously, if we choose $\eta=\Delta+\epsilon$ (to minimize the "false negatives"), then we have the guarantee that $\widehat{\mathcal{R}}_\eta$ is not "too large":

$$\mathcal{R}_\Delta\subseteq\widehat{\mathcal{R}}_\eta\subseteq\mathcal{R}_{\Delta+2\epsilon}.$$

To sum up, the notion of PAC algorithm in Definition 10 is meaningful because it permits to manage the amount of "false negatives" and "false positives" in a simple yet effective manner. Moreover, it guarantees that, if $\eta\in[\Delta-\epsilon,\Delta+\epsilon]$, then all the rewards $r\in\mathcal{R}$ with

6. Threshold Δ is settled by the problem, but we can implement our algorithm using any other value of threshold.



(a) Reward r is classified correctly.

(b) Reward r can be mis-classified.

(c) Range of uncertain (non)compatibility values.

Figure 4: The axis represents (non)compatibility values $\overline{\mathcal{C}}(\cdot)$ and we consider threshold $\eta = \Delta$. (a) Rewards r with $\overline{\mathcal{C}}(r) \leqslant \Delta - \epsilon$ or $\overline{\mathcal{C}}(r) \geqslant \Delta + \epsilon$ are correctly classified by an (ϵ, δ)-PAC with high probability, while (b) in the opposite case, r can be mis-classified. (c) The red interval $[\Delta - \epsilon, \Delta + \epsilon]$ exemplifies the set of rewards $\{r \in \mathcal{R} \,|\, |\overline{\mathcal{C}}(r) - \Delta| \leqslant \epsilon\}$ that are (potentially) mis-classified. The length of the interval reduces with ϵ.

true (non)compatibility $\overline{\mathcal{C}}(r) \leqslant \Delta - 2\epsilon \vee \overline{\mathcal{C}}(r) \geqslant \Delta + 2\epsilon$ are correctly classified with high probability, as shown in Figure 4.

5.3 Algorithm

In this section, we present CATY-IRL (CompATibilitY for IRL), a learning algorithm for solving the IRL classification problem in the online setting presented earlier. We begin by describing CATY-IRL as a general algorithmic template that can be applied to a variety of function approximation settings (i.e., classes of MDPs, like tabular MDPs or linear MDPs) that satisfy certain properties,[7] and to two different kinds of reward (non)compatibility (optimal expert in Definition 6, and suboptimal expert in Definition 7). Later on, we "instantiate" CATY-IRL in three classes of MDPs (tabular MDPs, tabular MDPs with linear rewards, Linear MDPs), and we demonstrate that CATY-IRL is a PAC algorithm (see Definition 5) for all these settings by providing explicit sample complexity bounds.

An IRL algorithm for the online setting is made of two phases:

Exploration phase. During the *exploration* phase (see Algorithm 1), CATY-IRL collects a data set $\mathcal{D}$ of trajectories by executing a reward-free exploration (RFE) algorithm $\mathfrak{A}$ (Jin et al., 2020a) for the considered function approximation setting. In this way, we are guaranteed that, whatever reward $r \in \mathcal{R}$ will be provided in input to the classification phase, we will be able to compute an estimate $\widehat{J}^*(r)$ "sufficiently" close to $J^*(r; p)$ using a reasonably small amount of samples (i.e., polynomial in the problem dimensions). Formally, by definition, any RFE algorithm $\mathfrak{A}$ guarantees that, for any $\epsilon > 0, \delta \in (0, 1)$, there exists (a reasonably small) $N > 0$ such that, if $\mathfrak{A}$ is executed for $\tau \geqslant N$ times, then it holds that

$$\mathop{\mathbb{P}}_{\mathcal{M}, \pi^E, \mathfrak{A}} \left(\sup_{r \in \mathfrak{R}} \left| J^*(r; p) - \widehat{J}^*(r) \right| \leqslant \frac{\epsilon}{2} \right) \geqslant 1 - \frac{\delta}{2}. \tag{7}$$

The reason why Algorithm 1 receives in input also the set $\mathcal{R}$ will be clear later.

Classification phase. During the *classification* phase (see Algorithm 2), CATY-IRL performs the estimation $\widehat{\mathcal{C}}(r)$ of the (non)compatibility term for the input reward $r \in \mathcal{R}$ by splitting it into two *independent* estimates: $\widehat{J}^E(r) \approx J^{\pi^E}(r; p)$, which is computed with $\mathcal{D}^E$ only,

7. As we will see, if two properties are satisfied, then CATY-IRL is a PAC algorithm for the considered setting.

Algorithm 1: CATY-IRL- Exploration phase

Data: Rewards to classify $\mathcal{R}$, number of episodes τ

`// Explore the environment with a RFE algorithm:`

1 $\mathcal{D} \leftarrow$ RFE_Exploration(τ)

2 Return $\mathcal{D}$

and $\widehat{J}^*(r) \approx J^*(r;p)$, which is computed with $\mathcal{D}$ only. We have already mentioned that the usage of a RFE algorithm guarantees that $\widehat{J}^*(r)$ is close to $J^*(r;p)$. Concerning the estimate $\widehat{J}^E(r)$ of the expert's performance $J^{\pi^E}(r;p)$, as shown in Line 1 of Algorithm 2, we use the empirical estimate (i.e., sample mean). We want $\widehat{J}^E(r)$ to be "close" to $J^{\pi^E}(r;p)$ for any $r \in \mathcal{R}$. As such, we require the function approximation setting considered to guarantee this closeness using a finite and reasonably small (i.e., polynomial in the problem dimensions) number of samples. Formally, we want that, for any $\epsilon > 0, \delta \in (0,1)$, whatever the input reward $r \in \mathcal{R}$, there exists (a reasonably small) $N^E > 0$ such that, if the data set $\mathcal{D}^E$ contains at least $\tau^E \geqslant N^E$ trajectories, then it holds that

$$\mathop{\mathbb{P}}_{\mathcal{M},\pi^E}\left(\sup_{r\in\mathfrak{R}}\left|J^{p,\pi^E}(r;p) - \widehat{J}^E(r)\right| \leqslant \frac{\epsilon}{2}\right) \geqslant 1 - \frac{\delta}{2}. \tag{8}$$

Then, CATY-IRL combines the estimates $\widehat{J}^E(r) \approx J^{\pi^E}(r;p)$ and $\widehat{J}^*(r) \approx J^*(r;p)$ to obtain the estimate of compatibility $\widehat{\mathcal{C}}(r)$ in a way dependent on the specific notion of (non)compatibility considered. Specifically, in the setting with optimal expert, in which the (non)compatibility $\overline{\mathcal{C}}$ considered is $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}$, defined in Eq. (5), then CATY-IRL computes estimate $\widehat{\mathcal{C}}(r)$ for input reward $r \in \mathcal{R}$ as

$$\widehat{\mathcal{C}}(r) = \widehat{J}^*(r) - \widehat{J}^E(r), \tag{9}$$

irrespective of the class of MDP. By using a union bound and a triangle inequality, if Eq. (7) and Eq. (8) hold, then it is easy to see that CATY-IRL is a PAC algorithm, that is,

$$\mathop{\mathbb{P}}_{\mathcal{M},\pi^E,\mathfrak{A}}\left(\sup_{r\in\mathcal{R}}\left|\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r) - \widehat{\mathcal{C}}(r)\right| \leqslant \epsilon\right) \geqslant 1 - \delta.$$

Similarly, in the setting with $[L,U]$-suboptimal expert (for arbitrary $U \geqslant L \geqslant 0$) where the (non)compatibility $\overline{\mathcal{C}}$ considered is $\overline{\mathcal{C}}^{L,U}_{\mathcal{M},\pi^E}$ defined in Eq. (6), CATY-IRL computes estimate $\widehat{\mathcal{C}}(r)$ for input reward $r \in \mathcal{R}$ as

$$\widehat{\mathcal{C}}(r) = \min_{x\in[L,U]}\left|x - (\widehat{J}^*(r) - \widehat{J}^E(r))\right|, \tag{10}$$

irrespective of the class of MDP. Note that this quantity can be computed in constant time since, for any $y \in \mathbb{R}$

$$\min_{x\in[L,U]}|x-y| = \begin{cases} L-y & \text{if } y < L \\ 0 & \text{if } y \in [L,U] \\ y-U & \text{if } y > U \end{cases}. \tag{11}$$

Again, if Eq. (7) and Eq. (8) hold, by using a union bound and the fact that the difference of two minima can be upper bounded by the maximum of the difference, i.e., that for any

Algorithm 2: CATY-IRL- Classification phase

Data: Expert data $\mathcal{D}^E$, exploration data $\mathcal{D}$, reward to classify $r \in \mathcal{R}$, threshold η

// Estimate the expert's performance $\widehat{J}^E(r)$:

1 $\widehat{J}^E(r) \leftarrow \frac{1}{\tau^E} \sum_{i \in [\![\tau^E]\!]} \sum_{h \in [\![H]\!]} r_h(s_h^{E,i}, a_h^{E,i})$

// Estimate the optimal performance $\widehat{J}^*(r)$:

2 $\widehat{J}^*(r) \leftarrow$ RFE_Planning$(\mathcal{D}, r)$

// Classify the reward:

optimal expert

3 $\widehat{\mathcal{C}}(r) \leftarrow \widehat{J}^*(r) - \widehat{J}^E(r)$

suboptimal expert

4 $\widehat{\mathcal{C}}(r) \leftarrow \min_{x \in [L,U]} |x - (\widehat{J}^*(r) - \widehat{J}^E(r))|$

5 $class \leftarrow True$ if $\widehat{\mathcal{C}}(r) \leqslant \eta$ else $False$

6 Return $class$

$y, y' \in \mathbb{R}$

$$\left| \min_{x \in [L,U]} |x - y| - \min_{x \in [L,U]} |x - y'| \right| \leqslant |y - y'|,$$

then it is easy to see that CATY-IRL is a PAC algorithm, that is,

$$\mathbb{P}_{\mathcal{M}, \pi^E, \mathfrak{A}} \left(\sup_{r \in \mathcal{R}} \left| \overline{\mathcal{C}}^{L,U}_{\mathcal{M}, \pi^E}(r) - \widehat{\mathcal{C}}(r) \right| \leqslant \epsilon \right) \geqslant 1 - \delta.$$

Finally, CATY-IRL applies the potentially negative threshold η to the estimate of (non)compatibility $\widehat{\mathcal{C}}(r)$ to perform the classification. Observe that $\Delta \in \mathbb{R}_{\geqslant 0}$ is the threshold in the definition of IRL classification problem of Definition 8, while $\eta \in \mathbb{R}$ is the actual threshold applied by CATY-IRL. As explained in the previous section, η can be different from Δ to trade-off the amount of false negative/positive.

Remark 1. *To prove that CATY-IRL is a PAC algorithm, the function approximation setting considered must satisfy two properties. First, it must admit a RFE algorithm for which Eq. (7) holds for a "small" N. Next, it must guarantee that Eq. (8) holds with a finite value of N^E at most polynomial in the dimensions of the problem. If both these conditions are satisfied, then, as explained in this section, CATY-IRL can be shown to be a PAC algorithm in this function approximation setting, with a sample complexity of N, N^E.*

5.3.1 Tabular MDPs and Tabular MDPs with Linear Rewards

In tabular MDPs and tabular MDPs with linear rewards, for exploration (Line 1 of Algorithm 1), we let CATY-IRL execute two different RFE algorithms depending on the amount of rewards to classify $\mathcal{R}$. Specifically, if $|\mathcal{R}|$ is a "small" constant w.r.t. to the size of the MDP, i.e., if it holds that $S + \log \frac{1}{\delta} \geqslant |\mathcal{R}| \log \frac{|\mathcal{R}|}{\delta}$ (see Theorem 4), then we run algorithm BPI-UCBVI (Menard et al., 2021) for each reward $r \in \mathcal{R}$. Otherwise we execute algorithm

RF-Express (Menard et al., 2021) once. Consequently, the computation of $\widehat{J}^*(r)$ (Line 2 of Algorithm 2) depends on the exploration algorithm adopted. Concerning BPI-UCBVI, CATY-IRL considers:[8]

$$\widehat{J}^*(r) = \max_{a \in \mathcal{A}} \widetilde{Q}_1^{\tau+1}(s_0, a), \tag{12}$$

where $\widetilde{Q}$ is the upper confidence bound on the optimal Q-function constructed by BPI-UCBVI (see Algorithm 2 of Menard et al. (2021)), $\mathcal{A}$ is the action space of the considered MDP without reward $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$, s_0 such that $d_0(s_0) = 1$ is the initial state, and τ is the number of iterations of algorithm BPI-UCBVI. With regards to RF-Express, CATY-IRL executes the *Backward Induction* algorithm (see Section 4.5 of Puterman (1994)) to compute the optimal V-function $V_h^*(\cdot; p, r)$ at all h in the MDP $(\mathcal{S}, \mathcal{A}, H, d_0, \widehat{p}, r)$, where $\widehat{p} \approx p$ is the empirical estimate of the transition model of $\mathcal{M}$ constructed with the data $\mathcal{D}$ gathered at exploration phase by RF-Express (see Algorithm 1 of Menard et al. (2021)). Then, CATY-IRL takes

$$\widehat{J}^*(r) = V_1^*(s_0; \widehat{p}, r), \tag{13}$$

where s_0 such that $d_0(s_0) = 1$ is the initial state.

5.3.2 Linear MDPs

In Linear MDPs, CATY-IRL uses algorithm RFLin (Wagenmaker et al., 2022) to explore the environment and to construct an estimate of the optimal utility $\widehat{J}^*(r)$ for any input reward $r \in \mathfrak{R}$:

$$\widehat{J}^*(r) = V_1(s_0), \tag{14}$$

where function V is the upper bound to the optimal value function defined at line 9 of Algorithm 2 of Wagenmaker et al. (2022) (RFLin-Plan), and s_0 is the initial state.

5.4 Sample Complexity

The following result shows that CATY-IRL is a PAC algorithm for IRL classification in the online setting (Definition 10) for all the classes of MDP and definitions of (non)compatibility mentioned earlier.

Theorem 4 (Sample Complexity of CATY-IRL)**.** *Assume that there is a single initial state. Let $U \geqslant L \geqslant 0$ be arbitrary and let $\epsilon > 0, \delta \in (0, 1)$. Then CATY-IRL executed with $\eta = \Delta$ is (ϵ, δ)-PAC for IRL classification in the online setting for both the optimal expert and the $[L, U]$-suboptimal expert settings, with a sample complexity upper bounded by*

$$\begin{aligned}
&\textit{Tabular MDPs:} && \tau^E \leqslant \widetilde{\mathcal{O}}\Big(\frac{H^3SA}{\epsilon^2}\log\frac{1}{\delta}\Big), && \tau \leqslant \widetilde{\mathcal{O}}\Big(\frac{H^3SA}{\epsilon^2}\Big(S + \log\frac{1}{\delta}\Big)\Big),\\
&\textit{Tabular MDPs with linear rewards:} && \tau^E \leqslant \widetilde{\mathcal{O}}\Big(\frac{H^3d}{\epsilon^2}\log\frac{1}{\delta}\Big), && \tau \leqslant \widetilde{\mathcal{O}}\Big(\frac{H^3SA}{\epsilon^2}\Big(S + \log\frac{1}{\delta}\Big)\Big),\\
&\textit{Linear MDPs:} && \tau^E \leqslant \widetilde{\mathcal{O}}\Big(\frac{H^3d}{\epsilon^2}\log\frac{1}{\delta}\Big), && \tau \leqslant \widetilde{\mathcal{O}}\Big(\frac{H^5d}{\epsilon^2}\Big(d + \log\frac{1}{\delta}\Big)\Big).
\end{aligned}$$

8. In case CATY-IRL executes BPI-UCBVI as many times as there are rewards in $r \in \mathcal{R}$, for the estimate of $\widehat{J}^*(r)$, we, of course, consider the $\widetilde{Q}$ obtained by exploring with the specific reward r.

If $|\mathcal{R}|\log(|\mathcal{R}|/\delta) \leqslant S + \log(1/\delta)$*, then* CATY-IRL *achieves the following improved rate in both tabular MDPs and tabular MDPs with linear rewards, in both the optimal and suboptimal expert settings:*

$$\tau^E \leqslant \widetilde{\mathcal{O}}\Big(\frac{H^2}{\epsilon^2}\log\frac{|\mathcal{R}|}{\delta}\Big), \quad \tau \leqslant \widetilde{\mathcal{O}}\Big(\frac{H^3SA}{\epsilon^2}|\mathcal{R}|\log\frac{|\mathcal{R}|}{\delta}\Big).$$

Some observations are in order. First, we remark that the assumption of a single initial state is common and not restrictive (Menard et al., 2021). The rate dependent on $|\mathcal{R}|$ that permits to improve over $S + \log(1/\delta)$, i.e., over a S^2 dependency, is possible in tabular MDPs and tabular MDPs with linear rewards, where executing $|\mathcal{R}|$ times the algorithm BPI-UCBVI requires less samples than running RF-Express once, i.e., as long as $|\mathcal{R}|\log(|\mathcal{R}|/\delta) \leqslant S + \log(1/\delta)$. Instead, we conjecture that the d^2 dependence is unavoidable in Linear MDPs because of the lower bound for BPI in Wagenmaker et al. (2022). Observe that the bounds do not depend on the classification threshold Δ as long as we set $\eta = \Delta$. Moreover, it is remarkable that both the settings with optimal and $[L, U]$-suboptimal expert enjoy the same sample complexity, which does not depend on L, U. In tabular MDPs with deterministic expert, one might use the results in Xu et al. (2023) to reduce the rate of τ^E at the price of increasing τ with additional logarithmic factors. However, in Lazzati et al. (2024b), we present a lower bound to τ for tabular MDPs in the setting with optimal expert that demonstrates that the upper bound provided by CATY-IRL is unimprovable up to logarithmic factors when $|\mathcal{R}|\log(|\mathcal{R}|/\delta) > S + \log(1/\delta)$, thus showing that CATY-IRL is *minimax optimal* in tabular MDPs with optimal expert.

In light of the result in Theorem 4, we conclude that the *reward compatibility* framework allows the *practical* development of *sample efficient* algorithms (e.g., CATY-IRL) in Linear MDPs with large/continuous state spaces.[9]

Proof of Theorem 4 As explained in Section 5.3, to show that CATY-IRL is PAC for the three function approximation settings considered in the statement of the theorem, we have to show that Eq. (7) and Eq. (8) hold with the mentioned number of samples N, N^E.

Let us begin by analyzing N^E for tabular MDPs. When $|\mathcal{R}|$ is finite, we have

$$\begin{aligned}
\mathop{\mathbb{P}}_{\mathcal{M},\pi^E}\Big(\exists r \in \mathcal{R} : \Big|J^{\pi^E}(r;p) - \widehat{J}^E(r)\Big| > \frac{\epsilon}{2}\Big) &\overset{(1)}{\leqslant} \sum_{r\in\mathcal{R}} \mathop{\mathbb{P}}_{\mathcal{M},\pi^E}\Big(\Big|J^{\pi^E}(r;p) - \widehat{J}^E(r)\Big| > \frac{\epsilon}{2}\Big) \\
&\overset{(2)}{=} \sum_{r\in\mathcal{R}} \mathop{\mathbb{P}}_{\mathcal{M},\pi^E}\Big(\Big|\mathbb{E}[\widehat{J}^E(r)] - \widehat{J}^E(r)\Big| > \frac{\epsilon}{2}\Big) \\
&\overset{(3)}{\leqslant} \sum_{r\in\mathcal{R}} 2e^{\frac{-\tau^E\epsilon^2}{2H^2}} \\
&\overset{(4)}{=} \frac{\delta}{2},
\end{aligned}$$

where at (1) we apply a union bound, at (2) we note that $J^{\pi^E}(r;p) = \mathbb{E}_{\mathcal{M},\pi^E}[\widehat{J}^E(r)] =: \mathbb{E}[\widehat{J}^E(r)]$ is the expected value of $\widehat{J}^E(r)$, at (3) we apply the Hoeffding's inequality, and at

9. Clearly, CATY-IRL can be implemented in practice, since it considers a single reward at a time instead of computing the full feasible set.

(4) we set $\delta/2 = |\mathcal{R}| 2e^{\frac{-\tau^E \epsilon^2}{2H^2}}$, so that

$$\tau^E \geqslant \frac{2H^2}{\epsilon^2} \ln \frac{4|\mathcal{R}|}{\delta}.$$

For any choice of $\mathcal{R}$ we have

$$\begin{aligned}
\mathop{\mathbb{P}}_{\mathcal{M},\pi^E}\left(\exists r \in \mathcal{R} : \left|J^{\pi^E}(r;p) - \widehat{J}^E(r)\right| > \frac{\epsilon}{2}\right) &\leqslant \mathop{\mathbb{P}}_{\mathcal{M},\pi^E}\left(\exists r \in \mathfrak{R} : \left|J^{\pi^E}(r;p) - \widehat{J}^E(r)\right| > \frac{\epsilon}{2}\right) \\
&\overset{(5)}{=} \mathop{\mathbb{P}}_{\mathcal{M},\pi^E}\left(\exists r \in \mathfrak{R}_{\text{vertices}} : \left|J^{\pi^E}(r;p) - \widehat{J}^E(r)\right| > \frac{\epsilon}{2}\right) \\
&\overset{(6)}{\leqslant} \sum_{r \in \mathfrak{R}_{\text{vertices}}} \mathop{\mathbb{P}}_{\mathcal{M},\pi^E}\left(\left|J^{\pi^E}(r;p) - \widehat{J}^E(r)\right| > \frac{\epsilon}{2}\right) \\
&\overset{(7)}{=} \sum_{r \in \mathfrak{R}_{\text{vertices}}} \mathop{\mathbb{P}}_{\mathcal{M},\pi^E}\left(\left|\mathbb{E}[\widehat{J}^E(r)] - \widehat{J}^E(r)\right| > \frac{\epsilon}{2}\right) \\
&\overset{(8)}{\leqslant} \sum_{r \in \mathfrak{R}_{\text{vertices}}} 2e^{\frac{-\tau^E \epsilon^2}{2H^2}} \\
&\overset{(9)}{=} \frac{\delta}{2},
\end{aligned}$$

where at (5) we make the same observation as in the proof of Lemma 6 of Shani et al. (2022): $\mathfrak{R}$ is the SAH-dimensional hypercube, thus all the rewards $r \in \mathfrak{R}$ can be written as a convex combination of rewards in $\mathfrak{R}_{\text{vertices}} \coloneqq \{r \in \mathfrak{R} \,|\, \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!] : \; r_h(s,a) \in \{-1,+1\}\}$. Thus, to upper bound the difference:

$$\sup_{r \in \mathfrak{R}} \left|J^{\pi^E}(r;p) - \widehat{J}^E(r)\right| = \sup_{r \in \mathfrak{R}} \left| \sum_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]} \left(d_h^{p,\pi^E}(s,a) - \widehat{d}_h^E(s,a)\right) r_h(s,a) \right|,$$

where $\widehat{d}_h^E(s,a)$ is the sample mean of $d_h^{p,\pi^E}(s,a)$ using $\mathcal{D}^E$, it suffices to consider just the rewards in $\mathfrak{R}_{\text{vertices}}$, since the quantity is maximized by $r_h(s,a) = +1$ when the difference $d_h^{p,\pi^E}(s,a) - \widehat{d}_h^E(s,a) \geqslant 0$, and by $r_h(s,a) = -1$ otherwise. At (6) we apply a union bound, at (7) we recognize the expectation, at (8) we apply Hoeffding's inequality, and at (9) we set $\delta/2 = 2^{SAH} 2e^{\frac{-\tau^E \epsilon^2}{2H^2}}$, since $|\mathfrak{R}_{\text{vertices}}| = 2^{SAH}$, i.e., the number of vertices in the n-dimensional hypercube is 2^n. Thus:

$$\tau^E \geqslant \frac{2SAH^3}{\epsilon^2} \ln \frac{4}{\delta}.$$

Concerning tabular MDPs with linear rewards and Linear MDPs, note that the derivation is exactly the same, with the only difference that the reward functions are now d-dimensional for all $h \in [\![H]\!]$, thus the union bound has to be computed over a d-dimensional hypercube, obtaining:

$$\tau^E \geqslant \frac{2dH^3}{\epsilon^2} \ln \frac{4}{\delta}.$$

Let us now analyze the number of samples N. In tabular MDPs and tabular MDPs with linear rewards, when $\mathcal{R}$ is finite and satisfies $|\mathcal{R}| \log(|\mathcal{R}|/\delta) \leqslant S + \log(1/\delta)$, then CATY-IRL

executes algorithm BPI-UCBVI of Menard et al. (2021) for every $r \in \mathcal{R}$, and it sets:

$$\widehat{J}^*(r) = \max_{a \in \mathcal{A}} \widetilde{Q}_1^{\tau+1}(s_0, a),$$

as explained in Eq. (12). As shown by the authors (Menard et al., 2021) in their proof of Lemma 2, it holds that

$$J^*(r; p) - J^{\widehat{\pi}}(r; p) \leqslant \max_{a \in \mathcal{A}} \widetilde{Q}_1^{\tau+1}(s_0, a) - J^{\widehat{\pi}}(r; p),$$

where $\widehat{\pi}$ is the actual output of BPI-UCBVI. Thus, for any $\epsilon' > 0$:

$$J^*(r; p) - J^{\widehat{\pi}}(r; p) \leqslant \epsilon' \implies \max_{a \in \mathcal{A}} \widetilde{Q}_1^{\tau+1}(s_0, a) - J^*(r; p) \leqslant \epsilon',$$

since $\max_{a \in \mathcal{A}} \widetilde{Q}_1^{\tau+1}(s_0, a) \geqslant J^*(r; p)$ under their good event. BPI-UCBVI guarantees that, at each execution with a reward $r \in \mathcal{R}$, with:

$$\tau \leqslant \widetilde{\mathcal{O}}\Big(\frac{H^3 SA}{\epsilon^2} \ln \frac{|\mathcal{R}|}{\delta}\Big),$$

it holds that (see Theorem 2 of Menard et al. (2021)):

$$\mathop{\mathbb{P}}_{\mathcal{M}, \pi^E, \mathfrak{A}} \left(\left| J^*(r; p) - \widehat{J}^*(r) \right| \leqslant \frac{\epsilon}{2} \right) \geqslant 1 - \frac{\delta}{2|\mathcal{R}|}.$$

Through a union bound, we obtain:

$$\begin{aligned} \mathop{\mathbb{P}}_{\mathcal{M}, \pi^E, \mathfrak{A}} \left(\exists r \in \mathcal{R} : \left| J^*(r; p) - \widehat{J}^*(r) \right| > \frac{\epsilon}{2} \right) &\leqslant \sum_{r \in \mathcal{R}} \mathop{\mathbb{P}}_{\mathcal{M}, \pi^E, \mathfrak{A}} \left(\left| J^*(r; p) - \widehat{J}^*(r) \right| > \frac{\epsilon}{2} \right) \\ &\leqslant \frac{\delta}{2}. \end{aligned}$$

In tabular MDPs and tabular MDPs with linear rewards, when `CATY-IRL` executes RF-Express as subroutine, and it sets, for all $r \in \mathcal{R}$ (Eq. (13)):

$$\widehat{J}^*(r) = V_1^*(s_0; \widehat{p}, r),$$

since $V_1^*(s_0; \widehat{p}, r)$ is the exact optimal performance in the MDP with dynamics $\widehat{p}$ since it has been compute through Backward Induction (Puterman, 1994), then, thanks to Theorem 1 of Menard et al. (2021), we have that Eq. (7) holds with:

$$\tau \leqslant \widetilde{\mathcal{O}}\Big(\frac{H^3 SA}{\epsilon^2} \Big(S + \log \frac{1}{\delta}\Big)\Big).$$

Therefore, we prefer to execute BPI-UCBVI for $|\mathcal{R}|$ times instead of RF-Express once if $\mathcal{R}$ satisfies (modulo some constants):

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

Finally, in Linear MDPs, by using algorithm RFLin (Wagenmaker et al., 2022), and making the estimate (Eq. (14)):

$$\widehat{J}^*(r) = V_1(s_0),$$

then, Theorem 1 of Wagenmaker et al. (2022) guarantees that Eq. (7) holds with (we omit linear terms in $1/\epsilon$):

$$\tau \leqslant \widetilde{\mathcal{O}}\Big(\frac{H^5 d}{\epsilon^2}\big(d + \log\frac{1}{\delta}\big)\Big).$$

This concludes the proof. ■

6. Learning Compatible Rewards in the Offline Setting

In this section, we analyse the reward compatibility framework in the offline setting for tabular MDPs.

6.1 Problem Setting

In many applications, IRL is better framed as an *offline* problem, in which there is no possibility to actively exploring the environment to improve our estimates. For this reason, in this section, we consider the offline scenario in which we are given a batch expert's data set $\mathcal{D}^E = \{(s_1^{E,i}, a_1^{E,i}, \dots, s_H^{E,i}, a_H^{E,i}, s_{H+1}^{E,i})\}_{i \in [\![\tau^E]\!]}$ of τ^E state-action trajectories collected by the expert's policy π^E in an MDP without reward $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$ (with unknown dynamics p), and an additional batch data set $\mathcal{D}^b = \{(s_1^{b,i}, a_1^{b,i}, \dots, s_H^{b,i}, a_H^{b,i}, s_{H+1}^{b,i})\}_{i \in [\![\tau^b]\!]}$ of τ^b state-action trajectories collected by executing a behavioral policy π^b in the same MDP without reward of the expert $\mathcal{M}$.

Comparing with the *online* setting presented in Section 5.1, we still have a batch data set $\mathcal{D}^E$ that gives us information on π^E (and its occupancy measure d^{p,π^E}). However, instead of being allowed to explore the environment at will to construct an estimate of the transition model p, we now have to estimate p using the trajectories in the new batch data set $\mathcal{D}^b$.

We make two remarks to clarify the importance of using two data sets.

Remark 2. *As we did in the* online *setting, we will* not *use the data in $\mathcal{D}^E$ to improve our estimate of the transition model p. The reason is that mixing the data would* unnecessarily *complicate the theoretical analysis without any significant advantage. Nevertheless, note that, in practice, using all the samples in $\mathcal{D}^E \cup \mathcal{D}^b$ to estimate p might improve the performance of the algorithms.*

Remark 3. *The requirement of* two data sets *is* not necessary*, although it can be useful. In fact, in many applications, the expert's policy π^E is deterministic or moderately stochastic. Thus, it provides a limited coverage of the state-action space, preventing us from constructing estimates of the transition model p in portions of the space that are not reached by π^E. To avoid this, we can consider an additional data set $\mathcal{D}^b$ collected by a potentially more explorative (stochastic) policy π^b. Note that this is a generalization of the common setting with only $\mathcal{D}^E$, that we recover if we take $\pi^b = \pi^E$. In the following, no restriction is made on π^b.*

6.2 Reward Compatibility

We now extend the reward compatibility approach to this setting.

Non-learnability of the (non)compatibility. The major difficulty of the offline setting is that, even in the limit of infinite trajectories in the batch data set $\mathcal{D}^b$, the behavioral policy may cover only a portion $\mathcal{Z}^{p,\pi^b} \subset \mathcal{S} \times \mathcal{A} \times [\![H]\!]$ of the space, preventing the estimation of the transition model in triples $(s,a,h) \notin \mathcal{Z}^{p,\pi^b}$. Observe that, whatever reward $r \in \mathfrak{R}$ we consider, both notions of (non)compatibility $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r)$ (Definition 6) and $\overline{\mathcal{C}}^{L,U}_{\mathcal{M},\pi^E}(r)$ (Definition 7) depend on the optimal performance $J^*(r;p)$ that can be achieved under r. Intuitively, since $J^*(r;p)$ depends on the transition model p at all the triples $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!]$ (reachable by at least one policy from the initial state s_0), and since we do not have information on the transition model in some of these triples because of the partial coverage of π^b, then estimating the (non)compatibility of r is not feasible in the offline setting:[10]

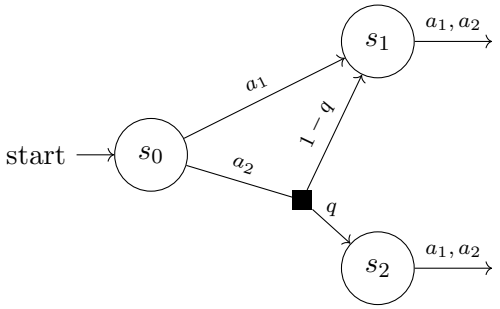
Theorem 5 (Non-learnability of $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}$ and $\overline{\mathcal{C}}^{L,U}_{\mathcal{M},\pi^E}$ in the offline setting)**.** *Let $U \geqslant L \geqslant 0$ be arbitrary. Let $\rho : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ be an arbitrary metric between scalars. Let $\mathfrak{A}$ be any algorithm that aims to estimate the (non)compatibilities in Definition 6 and 7. Then, there exists an MDP without reward $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$, an expert's policy π^E, and a behavioral policy π^b, such that, even if the batch data sets $\mathcal{D}^E \sim \pi^E$ and $\mathcal{D}^b \sim \pi^b$ contain an infinite amount of trajectories, there are $\epsilon, \delta \in (0,1)$ and $r \in \mathfrak{R}$ for which:*

$$\mathop{\mathbb{P}}_{\mathcal{M},\pi^E,\pi^b}\Big(\rho\Big(\overline{\mathcal{C}}(r), \widehat{\mathcal{C}}(r)\Big) \geqslant \epsilon\Big) \geqslant \delta,$$

where $\mathbb{P}_{\mathcal{M},\pi^E,\pi^b}$ denotes the probability measure induced by π^E and π^b in $\mathcal{M}$, $\overline{\mathcal{C}}(r)$ is any of $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r)$ or $\overline{\mathcal{C}}^{L,U}_{\mathcal{M},\pi^E}(r)$, and $\widehat{\mathcal{C}}(r)$ is the estimate of $\overline{\mathcal{C}}(r)$ computed by $\mathfrak{A}$ using only the data in $\mathcal{D}^E$ and $\mathcal{D}^b$.

Proof We begin by considering the setting with optimal expert.

Consider the MDP without reward pictured below, where s_0 is the initial state, there are two actions a_1, a_2 in each state, and the horizon is $H = 2$. From state s_0, action a_1 brings deterministically to s_1, while action a_2 brings to s_2 with probability $q \in [0,1]$, and to s_1 with probability $1-q$:



10. Observe that also the notion of feasible set is not "learnable" in this setting, as we have shown in Theorem C.1 of Lazzati et al. (2024a).

We denote as $\mathcal{M}_0, \mathcal{M}_1$ the MDPs without reward corresponding to, respectively, $q = 0$ and to $q = 1$. Let $\pi^E = \pi^b$ be the policies that play action a_1 at every state.

We now show that there exists at least a reward $r \in \mathfrak{R}$ for which no algorithm $\mathfrak{A}$ can estimate the (non)compatibility $\overline{\mathcal{C}}_{\mathcal{M}_0,\pi^E}(r)$ of reward r in $\mathcal{M}_0$ and $\mathcal{M}_1$ with arbitrary accuracy and failure probability $\epsilon, \delta \in (0,1)$ using only $\mathcal{D}^E \sim \pi^E, \mathcal{D}^b \sim \pi^b$, even if they contain an infinite amount of samples.

Let $r \in \mathfrak{R}$ be the reward function that assigns reward 1 to all actions played in states s_0 and s_2, and reward 0 to actions played in state s_1. The (non)compatibilities of r with π^E in $\mathcal{M}_0$ and $\mathcal{M}_1$ are:

$$\overline{\mathcal{C}}_{\mathcal{M}_0,\pi^E}(r) = 0, \qquad \overline{\mathcal{C}}_{\mathcal{M}_1,\pi^E}(r) = 1.$$

Since $\overline{\mathcal{C}}_{\mathcal{M}_0,\pi^E}(r) \neq \overline{\mathcal{C}}_{\mathcal{M}_1,\pi^E}(r)$, and since ρ is a metric, then there exists a value $k > 0$ such that

$$\rho\Big(\overline{\mathcal{C}}_{\mathcal{M}_0,\pi^E}(r), \overline{\mathcal{C}}_{\mathcal{M}_1,\pi^E}(r)\Big) = k.$$

Since $\pi^E = \pi^b$ always play action a_1, then even with an infinite amount of samples, $\mathcal{D}^E, \mathcal{D}^b$ do not reveal any information on the transition model of action a_2 in s_0, thus algorithm $\mathfrak{A}$ cannot discriminate between environments $\mathcal{M}_0, \mathcal{M}_1$. By choosing $\epsilon < k/2, \delta > 1/2$, we have that either the output $\widehat{\mathcal{C}}(r)$ of $\mathfrak{A}$ satisfies, for any $\mathcal{M} \in \{\mathcal{M}_0, \mathcal{M}_1\}$, $\mathbb{P}_{\mathcal{M},\pi^E,\pi^b}(\rho(\overline{\mathcal{C}}_{\mathcal{M}_0,\pi^E}, \widehat{\mathcal{C}}(r)) < k/2) > 1/2$ or $\mathbb{P}_{\mathcal{M},\pi^E,\pi^b}(\rho(\overline{\mathcal{C}}_{\mathcal{M}_1,\pi^E}, \widehat{\mathcal{C}}(r)) < k/2) > 1/2$, but not both since ρ satisfies the triangle inequality. Thus, in at least one of the two instances $\mathcal{M}_0, \mathcal{M}_1$ any algorithm $\mathfrak{A}$ satisfies the statement of the theorem. This concludes the proof for the setting with optimal expert.

An analogous construction can be constructed for the setting with suboptimal expert. ■

New notions of reward compatibility. Theorem 5 shows that, because of the partial coverage induced by π^b, the notions of (non)compatibility $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}$ and $\overline{\mathcal{C}}^{L,U}_{\mathcal{M},\pi^E}$ do not represent reasonable learning targets in the offline setting. For this reason, the best we can do is to define an "optimistic" and a "pessimistic" extension of (non)compatibility, that represent, respectively, the best and the worst possible value of (non)compatibility of the considered reward given the available information:

Definition 11 (Best and worst (non)compatibility)**.** *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$ be an MDP without reward, π^E the expert's policy, and π^b the behavioral policy. Let $\mathcal{Z} := \mathcal{Z}^{p,\pi^b}$ be the portion of space covered by π^b. Then, given a notion of (non)compatibility $\overline{\mathcal{C}}$, for any $r \in \mathfrak{R}$, we define the* best (non)compatibility $\overline{\mathcal{C}}^b(r)$ *of reward r given partial coverage $\mathcal{Z}$ as*

$$\overline{\mathcal{C}}^b(r) := \min_{p' \in [p]_{\equiv_{\mathcal{Z}}}} \overline{\mathcal{C}}(r). \tag{15}$$

Similarly, we define the worst (non)compatibility $\overline{\mathcal{C}}^w(r)$ *of reward r given partial coverage $\mathcal{Z}$ as*

$$\overline{\mathcal{C}}^w(r) := \max_{p' \in [p]_{\equiv_{\mathcal{Z}}}} \overline{\mathcal{C}}(r). \tag{16}$$

Intuitively, at best, the (non)compatibility of r is $\overline{\mathcal{C}}^b(r) \leqslant \overline{\mathcal{C}}(r)$, and, at worst, it is $\overline{\mathcal{C}}^w(r) \geqslant \overline{\mathcal{C}}(r)$. In other words, we are proposing a best-/worst-case approach to cope with the missing knowledge of the dynamics p outside $\mathcal{Z}^{p,\pi^b}$. From now on, we use $\mathcal{Z} := \mathcal{Z}^{p,\pi^b}$ to denote the portion of space covered by π^b. Given any r, we denote respectively as $\overline{\mathcal{C}}^b_{\mathcal{M},\pi^E,\mathcal{Z}}(r) := \min_{p' \in [p]_{\equiv_{\mathcal{Z}}}} \overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r)$ and $\overline{\mathcal{C}}^w_{\mathcal{M},\pi^E,\mathcal{Z}}(r) := \max_{p' \in [p]_{\equiv_{\mathcal{Z}}}} \overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r)$ the best and worst (non)compatibility of r in the setting with optimal expert. Analogously, for arbitrary $U \geqslant L \geqslant 0$, we denote respectively as $\overline{\mathcal{C}}^{L,U,b}_{\mathcal{M},\pi^E,\mathcal{Z}}(r) := \min_{p' \in [p]_{\equiv_{\mathcal{Z}}}} \overline{\mathcal{C}}^{L,U}_{\mathcal{M},\pi^E}(r)$ and $\overline{\mathcal{C}}^{L,U,w}_{\mathcal{M},\pi^E,\mathcal{Z}}(r) := \max_{p' \in [p]_{\equiv_{\mathcal{Z}}}} \overline{\mathcal{C}}^{L,U}_{\mathcal{M},\pi^E}(r)$ the best and worst (non)compatibility of r in the setting with $[L,U]$-suboptimal expert.

Each reward compatibility defines a feasible set. We observe that, analogously to what we have done in Section 4, we can define two new notions of "feasible set" as the sets of rewards with, respectively, zero best and worst (non)compatibility:

$$\mathcal{R}^b := \Big\{ r \in \mathfrak{R} \,\Big|\, \overline{\mathcal{C}}^b(r) = 0 \Big\},$$
$$\mathcal{R}^w := \Big\{ r \in \mathfrak{R} \,\Big|\, \overline{\mathcal{C}}^w(r) = 0 \Big\}.$$

These sets satisfy the inclusion monotonicity property (Lazzati et al., 2024a):

$$\mathcal{R}^b \subseteq \mathcal{R} \subseteq \mathcal{R}^w,$$

where $\mathcal{R} := \big\{ r \in \mathfrak{R} \,\big|\, \overline{\mathcal{C}}(r) = 0 \big\}$ is the feasible set. If we consider the setting with optimal expert $\overline{\mathcal{C}} = \overline{\mathcal{C}}_{\mathcal{M},\pi^E}$, then we recover the notions of sub-/super-feasible set that we analysed in Lazzati et al. (2024a).

Observe that the notions of IRL classification problem (Definition 8) and IRL algorithm (Definition 9) comply with the new definitions. The only difference is that, now, we expect our IRL algorithm to output two booleans, representing the classification carried out based on both the best and worst (non)compatibilities (we use a single threshold Δ for both).

We conclude this section with a remark. As explained in Lazzati et al. (2024a) (Proposition 8.1) under the name of "bitter lesson", in the setting with optimal expert, when we have only data collected by a deterministic expert's policy π^E, i.e., $\mathcal{D}^b = \mathcal{D}^E$, then the sub-feasible set $\mathcal{R}^w_{\mathcal{M},\pi^E} := \big\{ r \in \mathfrak{R} \,\big|\, \overline{\mathcal{C}}^w_{\mathcal{M},\pi^E,\mathcal{Z}}(r) = 0 \big\}$ exhibits some degeneracy, since it contains only reward functions whose greedy optimal action at all $(s,h) \in \mathcal{S}^{p,\pi^E}$ is the expert's action $\pi^E_h(s)$. However, reward compatibility can overcome this limitation, in the following manner. Let $\mathcal{R}^w_{\mathcal{M},\pi^E,\Delta} := \big\{ r \in \mathfrak{R} \,\big|\, \overline{\mathcal{C}}^w_{\mathcal{M},\pi^E,\mathcal{Z}}(r) \leqslant \Delta \big\}$ be the set of rewards that should be positively classified in the IRL classification problem with threshold Δ. Clearly, $\mathcal{R}^w_{\mathcal{M},\pi^E} \subseteq \mathcal{R}^w_{\mathcal{M},\pi^E,\Delta}$, and it is not difficult to see that, if $\Delta > 0$ strictly, then $\mathcal{R}^w_{\mathcal{M},\pi^E,\Delta}$ contains rewards that do not suffer from the "bitter lesson", i.e., do not make π^E the greedy policy w.r.t. the immediate reward r. The price to pay for the increased expressivity of the solution concept is an increased (non)compatibility (i.e., error) in the learned rewards. In other words, the larger the Δ, the more the expressivity of the learned rewards, but, at the same time, the larger their (non)compatibility.

6.3 Learning Framework

We are interested in solving the IRL classification problem in the offline setting. For this purpose, we define an algorithm to be efficient if it provides "good" estimates for both the best and worst (non)compatibilities:

Definition 12 (PAC Algorithm - Offline setting)**.** *Let $\epsilon > 0, \delta \in (0,1)$, and let $\mathcal{D}^E$ be a data set of τ^E expert's trajectories and $\mathcal{D}^b$ a data set of τ^b behavioral trajectories. An algorithm $\mathfrak{A}$ is (ϵ, δ)-PAC for the IRL classification problem if:*

$$\mathop{\mathbb{P}}_{\mathcal{M},\pi^E,\pi^b}\left(\sup_{r\in\mathcal{R}}\left|\overline{\mathcal{C}}^b(r) - \widehat{\mathcal{C}}^b(r)\right| \leqslant \epsilon \ \wedge \ \sup_{r\in\mathcal{R}}\left|\overline{\mathcal{C}}^w(r) - \widehat{\mathcal{C}}^w(r)\right| \leqslant \epsilon\right) \geqslant 1-\delta,$$

where $\mathbb{P}_{\mathcal{M},\pi^E,\pi^b}$ is the joint probability measure induced by π^E and π^b in $\mathcal{M}$, and $\widehat{\mathcal{C}}^b, \widehat{\mathcal{C}}^w$ are, respectively, the estimates of the best and worst (non)compatibilities $\overline{\mathcal{C}}^b, \overline{\mathcal{C}}^w$ computed by $\mathfrak{A}$. The sample complexity *is defined by the pair (τ^E, τ^b).*

Observe that the PAC framework is general, and it can be instantiated for both the optimal and suboptimal expert settings. Let η_b, η_w be the thresholds used by our learning algorithm $\mathfrak{A}$ to classify rewards,[11] and define the sets of rewards positively classified as

$$\widehat{\mathcal{R}}^b_{\eta_b} := \left\{r \in \mathfrak{R} \,|\, \widehat{\mathcal{C}}^b(r) \leqslant \eta_b\right\},$$
$$\widehat{\mathcal{R}}^w_{\eta_w} := \left\{r \in \mathfrak{R} \,|\, \widehat{\mathcal{C}}^w(r) \leqslant \eta_w\right\}.$$

Similarly, define the sets of rewards that should actually be positively classified as

$$\mathcal{R}^b_{\Delta} := \left\{r \in \mathfrak{R} \,|\, \overline{\mathcal{C}}^b(r) \leqslant \Delta\right\},$$
$$\mathcal{R}^w_{\Delta} := \left\{r \in \mathfrak{R} \,|\, \overline{\mathcal{C}}^w(r) \leqslant \Delta\right\}.$$

It is not difficult to see that a PAC algorithm for the offline setting provides the same guarantees highlighted in Section 5.2 between sets $\widehat{\mathcal{R}}^b_{\eta_b}, \mathcal{R}^b_{\Delta}$ and $\widehat{\mathcal{R}}^w_{\eta_w}, \mathcal{R}^w_{\Delta}$. In other words, accurately tuning the thresholds η_b, η_w permits to trade-off the amount of "false negatives" / "false positives".

6.4 Algorithm

In this section, we present CATY-OFF-IRL (CompATibilitY for OFFline IRL), a learning algorithm for solving the IRL classification problem in the offline setting in *tabular MDPs*, in both the settings with optimal and suboptimal expert. The pseudocode of the algorithm is reported in Algorithm 3.

Description of the algorithm. CATY-OFF-IRL takes in input a reward function $r \in \mathcal{R}$ and a threshold Δ, and it aims to output two booleans meant to classify the reward based on its best and worst (non)compatibilities. In both the optimal and suboptimal expert settings, to compute an estimate of the best and worst (non)compatibilities, CATY-OFF-IRL first estimates three different quantities: the expert's performance $\widehat{J}^E(r) \approx J^{\pi^E}(r;p)$, the best

11. Even though we have only one Δ, we can use different thresholds η_b, η_w.

Algorithm 3: CATY-OFF-IRL

Data: Expert data set $\mathcal{D}^E$, behavioral data set $\mathcal{D}^b$, classification threshold η, reward to classify $r \in \mathcal{R}$

// Estimate the expert's performance $\widehat{J}^E(r)$:

$\widehat{J}^E(r) \leftarrow \frac{1}{\tau^E} \sum_{i \in [\![\tau^E]\!]} \sum_{h \in [\![H]\!]} r_h(s_h^{E,i}, a_h^{E,i})$

// Estimate the support $\mathcal{Z}$ and the transition model p:

if $|\mathcal{R}| = 1$ **then**

  Randomly split $\mathcal{D}^b = \{\omega_i^b\}_{i \in [\![\tau^b]\!]}$ into H data sets $\{\mathcal{D}_h^b\}_{h \in [\![H]\!]}$ with $|\mathcal{D}_h^b| = \lfloor \tau^b / H \rfloor$

  $\widehat{\mathcal{Z}} \leftarrow \{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!] \,|\, \exists i \in [\![\tau^b]\!] : \omega_i^b \in \mathcal{D}_h^b \wedge (s_h^{b,i}, a_h^{b,i}) = (s,a)\}$

  $\widehat{p}_h(\cdot|s,a) \leftarrow$ empirical estimate of p from $\mathcal{D}_h^b$ for all $(s,a,h) \in \widehat{\mathcal{Z}}$ through Eq. (17)

else

  $\widehat{\mathcal{Z}} \leftarrow \{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [\![H]\!] \,|\, \exists i \in [\![\tau^b]\!] : (s_h^{b,i}, a_h^{b,i}) = (s,a)\}$

  $\widehat{p}_h(\cdot|s,a) \leftarrow$ empirical estimate of p from $\mathcal{D}^b$ for all $(s,a,h) \in \widehat{\mathcal{Z}}$ through Eq. (18)

// Estimate the optimal performances $\widehat{J}_m^*(r)$ and $\widehat{J}_M^*(r)$:

$\widehat{Q}_H^m(s,a), \widehat{Q}_H^M(s,a) \leftarrow r_H(s,a)\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}$

for $h = H-1$ **to** 1 **do**

  for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**

    $\widehat{Q}_h^M(s,a) \leftarrow r_h(s,a) + \max_{p' \in [\widehat{p}]_{\equiv_{\widehat{\mathcal{Z}}}}} \sum_{s' \in \mathcal{S}} p'_h(s'|s,a) \max_{a' \in \mathcal{A}} \widehat{Q}_{h+1}^M(s',a')$

    $\widehat{Q}_h^m(s,a) \leftarrow r_h(s,a) + \min_{p' \in [\widehat{p}]_{\equiv_{\widehat{\mathcal{Z}}}}} \sum_{s' \in \mathcal{S}} p'_h(s'|s,a) \max_{a' \in \mathcal{A}} \widehat{Q}_{h+1}^m(s',a')$

  end

end

$\widehat{J}_m^*(r) \leftarrow \max_{a \in \mathcal{A}} \widehat{Q}_1^m(s_0,a)$

$\widehat{J}_M^*(r) \leftarrow \max_{a \in \mathcal{A}} \widehat{Q}_1^M(s_0,a)$

// Classify the reward:

$\widehat{\Delta}_m(r) \leftarrow \widehat{J}_m^*(r) - \widehat{J}^E(r)$

$\widehat{\Delta}_M(r) \leftarrow \widehat{J}_M^*(r) - \widehat{J}^E(r)$

optimal expert

$\widehat{\mathcal{C}}^b(r) \leftarrow \widehat{\Delta}_m(r)$

$\widehat{\mathcal{C}}^w(r) \leftarrow \widehat{\Delta}_M(r)$

suboptimal expert

$\widehat{\mathcal{C}}^b(r) \leftarrow \max\{\mathbb{1}\{\widehat{\Delta}_M(r) < L\}(L - \widehat{\Delta}_M(r)), \mathbb{1}\{\widehat{\Delta}_m(r) > U\}(\widehat{\Delta}_m(r) - U)\}$

$\widehat{\mathcal{C}}^w(r) \leftarrow \max\{\mathbb{1}\{\widehat{\Delta}_m(r) < L\}(L - \widehat{\Delta}_m(r)), \mathbb{1}\{\widehat{\Delta}_M(r) > U\}(\widehat{\Delta}_M(r) - U)\}$

$class^b \leftarrow True$ if $\widehat{\mathcal{C}}^b(r) \leqslant \eta$ else $False$

$class^w \leftarrow True$ if $\widehat{\mathcal{C}}^w(r) \leqslant \eta$ else $False$

Return $class^b, class^w$

optimal performance $\widehat{J}_M^*(r) \approx \max_{p' \in [p]_{\equiv_{\mathcal{Z}}}} J^*(r;p')$, and the worst optimal performance $\widehat{J}_m^*(r) \approx \min_{p' \in [p]_{\equiv_{\mathcal{Z}}}} J^*(r;p')$ of the considered reward $r \in \mathcal{R}$.

CATY-OFF-IRL computes $\widehat{J}^E(r)$ as the empirical estimate (sample mean) of the expert's performance $J^{\pi^E}(r;p)$ using data set $\mathcal{D}^E$ and reward r (see Line 1), analogously to CATY-IRL. To compute estimates $\widehat{J}^*_m(r)$ and $\widehat{J}^*_M(r)$, CATY-OFF-IRL uses the empirical estimates of the support of the behavioral policy distribution $\widehat{\mathcal{Z}} \approx \mathcal{Z}$ (Lines 4 and 7),[12] and of the transition model $\widehat{p} \approx p$ (Lines 5 and 8). Specifically, the estimates of transition model at Line 5 are computed as

$$\widehat{p}_h(s'|s,a) := \frac{\sum_{i \in [\![\tau^b]\!]} \mathbb{1}\{(s_h^{b,i}, a_h^{b,i}, s_{h+1}^{b,i}) = (s,a,s') \wedge \omega_i^b \in \mathcal{D}_h^b\}}{\sum_{i \in [\![\tau^b]\!]} \mathbb{1}\{(s_h^{b,i}, a_h^{b,i}) = (s,a) \wedge \omega_i^b \in \mathcal{D}_h^b\}} \quad \forall (s,a,h) \in \widehat{\mathcal{Z}}, \forall s' \in \mathcal{S}, \tag{17}$$

where we use $\{\omega_i^b\}_{i \in [\![\tau^b]\!]}$ to denote the τ^b state-action trajectories contained into $\mathcal{D}^b$, and splitted into the H data sets $\{\mathcal{D}_h^b\}_h$. Instead, the estimates of transition model at Line 8 are computed as

$$\widehat{p}_h(s'|s,a) := \frac{\sum_{i \in [\![\tau^b]\!]} \mathbb{1}\{(s_h^{b,i}, a_h^{b,i}, s_{h+1}^{b,i}) = (s,a,s')\}}{\sum_{i \in [\![\tau^b]\!]} \mathbb{1}\{(s_h^{b,i}, a_h^{b,i}) = (s,a)\}} \qquad \forall (s,a,h) \in \widehat{\mathcal{Z}}, \forall s' \in \mathcal{S}. \tag{18}$$

Then, CATY-OFF-IRL computes:

$$\widehat{J}^*_m(r) = \min_{p' \in [\widehat{p}]_{\equiv_{\widehat{\mathcal{Z}}}}} J^*(r;p'),$$
$$\widehat{J}^*_M(r) = \max_{p' \in [\widehat{p}]_{\equiv_{\widehat{\mathcal{Z}}}}} J^*(r;p'),$$

by applying the Extended Value Iteration (EVI) algorithm (Auer et al., 2008) (Lines 9-17). Finally, CATY-OFF-IRL estimates the best and worst (non)compatibilities $\widehat{\mathcal{C}}^b(r) \approx \overline{\mathcal{C}}^b_{\mathcal{M},\pi^E,\mathcal{Z}}(r)$ and $\widehat{\mathcal{C}}^w(r) \approx \overline{\mathcal{C}}^w_{\mathcal{M},\pi^E,\mathcal{Z}}(r)$ in the setting with optimal expert as (see Line 20):

$$\widehat{\mathcal{C}}^b(r) = \widehat{J}^*_m(r) - \widehat{J}^E(r),$$
$$\widehat{\mathcal{C}}^w(r) = \widehat{J}^*_M(r) - \widehat{J}^E(r).$$

In the setting with suboptimal expert, observe that the best and worst (non)compatibilities can be re-written in a clearer form. To this aim, we need some additional notation. For any $r \in \mathfrak{R}$, for any transition model $p' \in [p]_{\equiv_{\mathcal{Z}}}$, we define $\Delta(r;p') := J^*(r;p') - J^{\pi^E}(r;p')$, and also $\Delta_m(r) := \min_{p' \in [p]_{\equiv_{\mathcal{Z}}}} J^*(r;p') - J^{\pi^E}(r;p')$, $\Delta_M(r) := \max_{p' \in [p]_{\equiv_{\mathcal{Z}}}} J^*(r;p') - J^{\pi^E}(r;p')$. Then:

Proposition 6. *For any $U \geqslant L \geqslant 0$, for any $r \in \mathfrak{R}$, it holds that*

$$\overline{\mathcal{C}}^{L,U,w}_{\mathcal{M},\pi^E,\mathcal{Z}}(r) = \max\Big\{\mathbb{1}\{\Delta_m(r) < L\}\big(L - \Delta_m(r)\big), \mathbb{1}\{\Delta_M(r) > U\}\big(\Delta_M(r) - U\big)\Big\},$$
$$\overline{\mathcal{C}}^{L,U,b}_{\mathcal{M},\pi^E,\mathcal{Z}}(r) = \max\Big\{\mathbb{1}\{\Delta_M(r) < L\}\big(L - \Delta_M(r)\big), \mathbb{1}\{\Delta_m(r) > U\}\big(\Delta_m(r) - U\big)\Big\}.$$

Proof We begin with the worst (non)compatibility:

$$\overline{\mathcal{C}}^{L,U,w}_{\mathcal{M},\pi^E,\mathcal{Z}}(r) := \max_{p' \in [p]_{\equiv_{\mathcal{Z}}}} \overline{\mathcal{C}}^{L,U}_{\mathcal{M},\pi^E}(r)$$

12. Observe that we distinguish between the case in which $|\mathcal{R}| = 1$ and $|\mathcal{R}| > 1$ through the splitting of the data set $\mathcal{D}^b$ carried out at line 9. This passage is needed to obtain the sample complexity guarantee in Theorem 7.

$$
\begin{aligned}
&= \max_{p' \in [p]_{\equiv_{\mathcal{Z}}}} \min_{x \in [L,U]} \left| x - \left(J^*(r;p') - J^{\pi^E}(r;p') \right) \right| \\
&= \max_{p' \in [p]_{\equiv_{\mathcal{Z}}}} \min_{x \in [L,U]} \left| x - \Delta(r;p') \right| \\
&\overset{(1)}{=} \max_{p' \in [p]_{\equiv_{\mathcal{Z}}}} \begin{cases} L - \Delta(r;p'), & \text{if } \Delta(r;p') < L \\ 0, & \text{if } \Delta(r;p') \in [L,U] \\ \Delta(r;p') - U, & \text{if } \Delta(r;p') > U \end{cases} \\
&\overset{(2)}{=} \begin{cases} L - \Delta_m(r), & \text{if } \Delta_m(r) < L \\ 0, & \text{if } \forall p' \in [p]_{\equiv_{\mathcal{Z}}} : \Delta(r;p') \in [L,U] \\ \Delta_M(r) - U, & \text{if } \Delta_M(r) > U \end{cases} \\
&= \max\Big\{ \mathbb{1}\{\Delta_m(r) < L\}\big(L - \Delta_m(r)\big), \mathbb{1}\{\Delta_M(r) > U\}\big(\Delta_M(r) - U\big) \Big\},
\end{aligned}
$$

where at (1) we use the observation in Eq. (11), and at (2) we simply recognize the worst cases. An analogous derivation can be carried out for the best (non)compatibility. ■

Thus, at Line 21, CATY-OFF-IRL simply applies the formulas in Proposition 6 replacing the values of $\Delta_m(r), \Delta_M(r)$ with their estimates computed at lines 18-19. Finally, at lines 22-23, CATY-OFF-IRL performs the classification of the input reward.

Sample efficiency. The following result bounds the sample complexity of CATY-OFF-IRL:

Theorem 7 (Sample Complexity of CATY-OFF-IRL)**.** *Assume that there is a single initial state. Let $U \geqslant L \geqslant 0$ be arbitrary and let $\epsilon > 0, \delta \in (0,1)$. Then CATY-OFF-IRL executed with $\eta = \Delta$ is (ϵ,δ)-PAC for IRL classification in the offline (tabular) setting for both the optimal expert and the $[L,U]$-suboptimal expert settings, with a sample complexity upper bounded by*

$$
\text{if } |\mathcal{R}| = 1: \quad \tau^E \leqslant \widetilde{\mathcal{O}}\Big(\frac{H^2}{\epsilon^2} \log \frac{1}{\delta}\Big), \qquad \tau^b \leqslant \widetilde{\mathcal{O}}\Big(\frac{H^5 \log^2 \frac{|\mathcal{Z}|}{\delta}}{\epsilon^2 d_{\min}^{p,\pi^b}} + \frac{\log \frac{1}{\delta}}{\log \frac{1}{1 - d_{\min}^{p,\pi^b}}}\Big),
$$

$$
\text{otherwise:} \quad \tau^E \leqslant \widetilde{\mathcal{O}}\Big(\frac{H^3 SA}{\epsilon^2} \log \frac{1}{\delta}\Big), \qquad \tau^b \leqslant \widetilde{\mathcal{O}}\Big(\frac{H^4 \log \frac{|\mathcal{Z}|}{\delta}}{\epsilon^2 d_{\min}^{p,\pi^b}} \Big(S + \log \frac{|\mathcal{Z}|}{\delta}\Big) + \frac{\log \frac{1}{\delta}}{\log \frac{1}{1 - d_{\min}^{p,\pi^b}}}\Big),
$$

where $d_{\min}^{p,\pi^b} := \min_{(s,a,h) \in \mathcal{Z}} d_h^{p,\pi^b}(s,a)$.

We make some comments. First, note that the expert sample complexity τ^E coincides with that provided in Theorem 4, since CATY-OFF-IRL computes the same estimate as CATY-IRL for $J^{\pi^E}(r;p)$. Next, observe that the upper bound to τ^b is made of two terms, one that displays a tight dependence on the desired accuracy ϵ and a dependence of order H^5 or H^4 on the horizon, and another term dependent on the minimum non-zero value of the visitation distribution $d_{\min}^{p,\pi^b} > 0$, that is needed to ensure $\widehat{\mathcal{Z}} = \mathcal{Z}$. More in detail, note that, since $1/d_{\min}^{p,\pi^b} \leqslant \mathcal{Z}$, and since $\mathcal{Z} = \mathcal{S} \times \mathcal{A} \times [\![H]\!]$ when π^b covers the entire space, then the dependence on SA is hidden inside $d_{\min}^{p,\pi^b}$. Also note that, when $|\mathcal{R}| = 1$, we need at most $\propto S$

samples, otherwise we require $\propto S^2$ data. We stress that CATY-OFF-IRL is computationally efficient since it just implements EVI (Auer et al., 2008), and thus it has a time complexity of order $\mathcal{O}(HS^2A)$. Nevertheless, notice that CATY-OFF-IRL does not scale to problems with large/infinite state spaces even under the structure imposed by Linear MDPs, because our definitions of best and worst (non)compatibility rely on the portion $\mathcal{Z}$ of the state-action space covered by π^b (through $d_{\min}^{p,\pi^b}$). We leave to future works the development of alternative (non)compatibility notions more suitable for large-scale settings, that are able to exploit the structure of the considered problem (e.g., Linear MDP). Simply put, these notions should be based on a concept of coverage other than $\mathcal{Z}$. Finally, we note that, differently from the results on the feasible set in this setting (see Theorem 5.1 of Lazzati et al. (2024a)), our Theorem 7 does *not* require the assumption that $\mathcal{Z}^{p,\pi^E} \subseteq \mathcal{Z}^{p,\pi^b}$, providing an additional advantage of the (non)compatibility framework over the feasible set.

To prove Theorem 7, we begin by showing that estimating $\widehat{\Delta}_m(\cdot) \approx \Delta_m(\cdot)$ and $\widehat{\Delta}_M(\cdot) \approx \Delta_M(\cdot)$ accurately suffices also for the setting with suboptimal expert. Indeed, in the setting with optimal expert, it is immediate.

Lemma 8. *For any problem instance $\mathcal{M}, \pi^E, \pi^b$, for any $U \geqslant L \geqslant 0$, for any $r \in \mathfrak{R}$, it holds that*

$$\left|\overline{\mathcal{C}}^{L,U,w}_{\mathcal{M},\pi^E,\mathcal{Z}}(r) - \widehat{\mathcal{C}}^{w}(r)\right| \leqslant \max\Big\{\big|\Delta_m(r) - \widehat{\Delta}_m(r)\big|, \big|\Delta_M(r) - \widehat{\Delta}_M(r)\big|\Big\},$$
$$\left|\overline{\mathcal{C}}^{L,U,b}_{\mathcal{M},\pi^E,\mathcal{Z}}(r) - \widehat{\mathcal{C}}^{b}(r)\right| \leqslant \max\Big\{\big|\Delta_m(r) - \widehat{\Delta}_m(r)\big|, \big|\Delta_M(r) - \widehat{\Delta}_M(r)\big|\Big\}.$$

Proof We prove the result only for the worst (non)compatibility. For the best (non)compatibility the proof is analogous. We can write

$$\begin{aligned}
\left|\overline{\mathcal{C}}^{L,U,w}_{\mathcal{M},\pi^E,\mathcal{Z}}(r) - \widehat{\mathcal{C}}^{w}(r)\right| &\overset{(1)}{=} \Big|\max\Big\{\mathbb{1}\{\Delta_m(r) < L\}\big(L - \Delta_m(r)\big), \mathbb{1}\{\Delta_M(r) > U\}\big(\Delta_M(r) - U\big)\Big\} \\
&\qquad - \max\Big\{\mathbb{1}\{\widehat{\Delta}_m(r) < L\}\big(L - \widehat{\Delta}_m(r)\big), \mathbb{1}\{\widehat{\Delta}_M(r) > U\}\big(\widehat{\Delta}_M(r) - U\big)\Big\}\Big| \\
&\overset{(2)}{\leqslant} \max\Big\{\Big|\mathbb{1}\{\Delta_m(r) < L\}\big(L - \Delta_m(r)\big) - \mathbb{1}\{\widehat{\Delta}_m(r) < L\}\big(L - \widehat{\Delta}_m(r)\big)\Big|, \\
&\qquad \Big|\mathbb{1}\{\Delta_M(r) > U\}\big(\Delta_M(r) - U\big) - \mathbb{1}\{\widehat{\Delta}_M(r) > U\}\big(\widehat{\Delta}_M(r) - U\big)\Big|\Big\} \\
&\overset{(3)}{\leqslant} \max\Big\{\Big|\big(L - \Delta_m(r)\big) - \big(L - \widehat{\Delta}_m\big)\Big|, \Big|\big(\Delta_M(r) - U\big) - \big(\widehat{\Delta}_M - U\big)\Big|\Big\} \\
&= \max\Big\{\Big|\Delta_m(r) - \widehat{\Delta}_m\Big|, \Big|\Delta_M(r) - \widehat{\Delta}_M\Big|\Big\},
\end{aligned}$$

where at (1) we use Proposition 6, and the estimate used in Line 21 of CATY-OFF-IRL, at (2) we use the Lipschitzianity of the maximum operator $|\max\{x, y\} - \max\{x', y'\}| \leqslant \max\{|x - x'|, |y - y'|\}$ for all $x, y, x', y' \in \mathbb{R}$, at (3) we recognize that, in all four cases obtained by comparing L with $\Delta_m(r)$, and L with $\widehat{\Delta}_m$, the upper bound holds; similarly also for U and $\Delta_M(r)$ and $\widehat{\Delta}_M$. ■

Next, we have to show that our estimates are close to the true quantities with high probability. Depending on the cardinality of $\mathcal{R}$ (the set of rewards to classify), we have two different lemmas. Before presenting the results, we need some additional notation. Specifically, given

any set $\mathcal{X} \subseteq \mathcal{S}\times\mathcal{A}\times[\![H]\!]$, and any pair of "partial" transition models $p' \in \Delta^{\mathcal{S}}_{\mathcal{X}}, p''\Delta^{\mathcal{S}}_{\mathcal{X}^{\complement}}$, i.e., one defined only on triples in $\mathcal{X}$, and the other on the remaining triples $\mathcal{X}^{\complement} = \mathcal{S}\times\mathcal{A}\times[\![H]\!]\backslash\mathcal{X}$, we define the V- and Q-functions and the expected utility under the transition model obtained by combining p' and p'' as $V^{\pi}_h(s;p',p'',r), Q^{\pi}_h(s,a;p',p'',r), J^{\pi}(r;p',p'')$ for any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[\![H]\!], r \in \mathfrak{R}, \pi \in \Pi$. We extend the notation analogously to the optimal V- and Q-functions and expected utility. In addition, we denote the visit distribution of π in this context as $d^{p',p'',\pi}$.

Lemma 9 (Concentration for $|\mathcal{R}| = 1$)**.** *Let $|\mathcal{R}| = 1$, and let $p' \in \Delta^{\mathcal{S}}_{\mathcal{Z}^{\complement}}$, $\pi \in \Pi$, and $r \in \mathfrak{R}$ be arbitrary. Let $\delta \in (0,1)$. Then, there exists a constant $c > 0$ for which the event $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ defined as the intersection of the events:*

$$\mathcal{E}_1 := \left\{ \left| J^{\pi^E}(r;p) - \widehat{J}^E(r) \right| \leqslant c\sqrt{\frac{H^2 \log\frac{6}{\delta}}{\tau^E}} \right\},$$

$$\mathcal{E}_2 := \left\{ \left| \mathbb{E}_{s'\sim p_h(\cdot|s,a)}[V^{\pi}_{h+1}(s';\widehat{p},p',r)] - \mathbb{E}_{s'\sim \widehat{p}_h(\cdot|s,a)}[V^{\pi}_{h+1}(s';\widehat{p},p',r)] \right| \leqslant c\sqrt{\frac{H^2 \ln\frac{12|\widehat{\mathcal{Z}}|}{\delta}}{N^b_h(s,a)}}, \right.$$
$$\left. \wedge \quad \frac{1}{N^b_h(s,a)} \leqslant c \cdot \frac{H \ln\frac{6|\widehat{\mathcal{Z}}|}{\delta}}{\tau^b d^{p,\pi^b}_h(s,a)} \ \forall (s,a,h) \in \widehat{\mathcal{Z}} \right\},$$

$$\mathcal{E}_3 := \left\{ N^b_h(s,a) \geqslant 1, \ \forall (s,a,h) \in \mathcal{Z} \quad \textit{when } \tau^b \geqslant H \frac{\ln\frac{3|\mathcal{Z}|}{\delta}}{\ln\frac{1}{1-d^{p,\pi^b}_{\min}}} \right\},$$

where $N^b_h(s,a)$ is the random variable that counts the number of samples inside data set $\mathcal{D}^b_h$ for triple (s,a,h), and $d^{p,\pi^b}_{\min} := \min_{(s,a,h)\in\mathcal{Z}} d^{p,\pi^b}_h(s,a)$, holds with probability at least $1-\delta$.

Proof The result follows by an application of the union bound after having observed that each event $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ holds with probability $1 - \delta/3$. Specifically, we have already shown in the proof of Theorem 4 that $\mathcal{E}_1$ holds with probability $1-\delta/3$. Concerning event $\mathcal{E}_2$, the result follows from Lemma B.1 of Xie et al. (2021). Finally, w.r.t. event $\mathcal{E}_3$, the result follows as shown in Lemma F.1 of Lazzati et al. (2024a), with an additional H term due to the splitting of the data sets carried out by CATY-OFF-IRL. ■

Lemma 10 (Concentration for $|\mathcal{R}| \geqslant 1$)**.** *Let $\delta \in (0,1)$. Then, there exists a constant $c > 0$ for which the event $\mathcal{E}' := \mathcal{E}'_1 \cap \mathcal{E}'_2 \cap \mathcal{E}'_3$ defined as the intersection of the events:*

$$\mathcal{E}'_1 := \left\{ \sup_{r\in\mathfrak{R}} \left| J^{\pi^E}(r;p) - \widehat{J}^E(r) \right| \leqslant c\sqrt{\frac{SAH^3 \log\frac{6}{\delta}}{2\tau^E}} \right\},$$

$$\mathcal{E}'_2 := \left\{ N^b_h(s,a) KL(\widehat{p}_h(\cdot|s,a) \| p_h(\cdot|s,a)) \leqslant \beta(N^b_h(s,a), \delta/6), \right.$$
$$\left. \wedge \quad \frac{1}{N^b_h(s,a)} \leqslant c \cdot \frac{\ln\frac{6|\widehat{\mathcal{Z}}|}{\delta}}{\tau^b d^{p,\pi^b}_h(s,a)} \ \forall (s,a,h) \in \widehat{\mathcal{Z}} \right\},$$

$$\mathcal{E}_3' := \left\{ N_h^b(s,a) \geqslant 1,\ \forall (s,a,h) \in \mathcal{Z} \quad \textit{when } \tau^b \geqslant \frac{\ln \frac{3|\mathcal{Z}|}{\delta}}{\ln \frac{1}{1-d_{\min}^{p,\pi^b}}} \right\},$$

where $N_h^b(s,a)$ *is the random variable that counts the number of samples inside data set* $\mathcal{D}^b$ *for triple* (s,a,h), $\beta(n,\delta) := \ln(|\widehat{\mathcal{Z}}|/\delta) + (S-1)\ln(e(1+n/(S-1)))$ *and* $d_{\min}^{p,\pi^b} := \min_{(s,a,h)\in\mathcal{Z}} d_h^{p,\pi^b}(s,a)$, *holds with probability at least* $1-\delta$.

Proof The result follows by an application of the union bound after having observed that each event $\mathcal{E}_1', \mathcal{E}_2', \mathcal{E}_3'$ holds with probability $1-\delta/3$. Specifically, we have already shown in the proof of Theorem 4 that $\mathcal{E}_1'$ holds with probability $1-\delta/3$. Concerning event $\mathcal{E}_2'$, the result is contained in Lemma F.2 of Lazzati et al. (2024a). In particular, the first part is proved through Lemma 10 in Kaufmann et al. (2021), while for the second part we simply apply Lemma A.1 of Xie et al. (2021). Finally, w.r.t. event $\mathcal{E}_3'$, the result follows as shown in Lemma F.1 of Lazzati et al. (2024a). ■

Lemma 9 and Lemma 10 are needed for the following tasks. Event $\mathcal{E}_1$ guarantees that CATY-OFF-IRL provides an accurate estimate of the expert's expected utility for a single reward, while $\mathcal{E}_1'$ provides the guarantee for all the bounded rewards. Event $\mathcal{E}_2$, intuitively, guarantees that the estimate of the transition model $\widehat{p}$ is close to the true one p at all the triples in $\widehat{\mathcal{Z}}$ for a single reward, while $\mathcal{E}_2'$ provides the guarantee in one-norm. Finally, events $\mathcal{E}_3$ and $\mathcal{E}_3'$ guarantee that $\widehat{\mathcal{Z}} = \mathcal{Z}$. Note that, due to the splitting of the data sets when $|\mathcal{R}| = 1$, the bound on τ^b for $\mathcal{E}_3$ is larger by a H term than the bound for $\mathcal{E}_3'$.

To prove Theorem 7, we need to prove one last lemma:

Lemma 11. *Let* s_0 *be the initial state and let* $p, \widehat{p} \in \Delta_{\mathcal{Z}}^{\mathcal{S}}$ *and* $p' \in \Delta_{\mathcal{Z}^{\mathbb{C}}}^{\mathcal{S}}$ *be arbitrary. Then, for any policy* π *and reward function* $r \in \mathfrak{R}$, *it holds that*

$$\left| J^\pi(r;p,p') - J^\pi(r;\widehat{p},p') \right| \leqslant \sum_{(s,a,h)\in\mathcal{Z}} d_h^{p,p',\pi}(s,a) \left| \sum_{s'\in\mathcal{S}} (p_h(s'|s,a) - \widehat{p}_h(s'|s,a)) V_{h+1}^\pi(s';\widehat{p},p',r) \right|.$$

Proof Let us denote by $\overline{p}, \breve{p} \in \Delta_{\mathcal{S}\times\mathcal{A}\times[\![H]\!]}^{\mathcal{S}}$ the transition models obtained by combining, respectively, p with p', and $\widehat{p}$ with p'. Then, we can write

$$\begin{aligned}
\left| J^\pi(r;\overline{p}) - J^\pi(r;\breve{p}) \right| &= \left| V_1^\pi(s_0;\overline{p},r) - V_1^\pi(s_0;\breve{p},r) \right| \\
&= \Bigg| \sum_{a\in\mathcal{A}} \pi_1(a|s_0) \sum_{s'\in\mathcal{S}} \overline{p}_1(s'|s_0,a) V_2^\pi(s';\overline{p},r) \\
&\qquad - \sum_{a\in\mathcal{A}} \pi_1(a|s_0) \sum_{s'\in\mathcal{S}} \breve{p}_1(s'|s_0,\pi_1(s_0)) V_2^\pi(s';\breve{p},r) \Bigg| \\
&\leqslant \sum_{a\in\mathcal{A}} \pi_1(a|s_0) \left| \sum_{s'\in\mathcal{S}} (\overline{p}_1(s'|s_0,a) - \breve{p}_1(s'|s_0,a)) V_2^\pi(s';\breve{p},r) \right| \\
&\qquad + \sum_{a\in\mathcal{A}} \pi_1(a|s_0) \sum_{s'\in\mathcal{S}} \overline{p}_1(s'|s_0,a) \left| V_2^\pi(s';\overline{p},r) - V_2^\pi(s';\breve{p},r) \right| \\
&\overset{(1)}{\leqslant} \dots
\end{aligned}$$

$$\leqslant \mathbb{E}_{\overline{p},\pi}\Bigg[\sum_{h\in[\![H]\!]}\Big|\sum_{s'\in\mathcal{S}}(\overline{p}_h(s'|s,a)-\breve{p}_h(s'|s,a))V^{\pi}_{h+1}(s';\breve{p},r)\Big|\Bigg]$$
$$=\sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[\![H]\!]} d^{\overline{p},\pi}_h(s,a)\Big|\sum_{s'\in\mathcal{S}}(\overline{p}_h(s'|s,a)-\breve{p}_h(s'|s,a))V^{\pi}_{h+1}(s';\breve{p},r)\Big|,$$

where at (1) we have unfolded the recursion.

The result follows by changing notation and by noticing that, by definition, $\overline{p}_h(s'|s,a) = \breve{p}_h(s'|s,a) = p'_h(s'|s,a)\ \forall s'\in\mathcal{S}$ in all $(s,a,h)\notin\mathcal{Z}$. ■

We are now ready to prove Theorem 7.

Proof of Theorem 7 Thanks to the definition of PAC algorithm in Definition 12, and thanks to Proposition 6, then it is clear that, if we show that CATY-OFF-IRL satisfies:

$$\mathop{\mathbb{P}}_{\mathcal{M},\pi^E,\pi^b}\Big(\sup_{r\in\mathcal{R}}\Big|\Delta_m(r)-\widehat{\Delta}_m(r)\Big|\leqslant\epsilon\ \wedge\ \sup_{r\in\mathcal{R}}\Big|\Delta_M(r)-\widehat{\Delta}_M(r)\Big|\leqslant\epsilon\Big)\geqslant 1-\delta,$$

then we have successfully proved that CATY-OFF-IRL is PAC.

Let $r\in\mathfrak{R}$ be arbitrary, and assume that event $\mathcal{E}$ holds if $|\mathcal{R}|=1$, otherwise assume that event $\mathcal{E}'$ holds. Then, we can write

$$\Big|\Delta_M(r)-\widehat{\Delta}_M(r)\Big|\leqslant\underbrace{\Big|\max_{p'\in[p]_{\equiv_{\mathcal{Z}}}}J^*(r;p')-\max_{\widehat{p}'\in[\widehat{p}]_{\equiv_{\widehat{\mathcal{Z}}}}}J^*(r;\widehat{p}')\Big|}_{=:\mathcal{I}}+\Big|\widehat{J}^E(r)-J^{\pi^E}(r;p)\Big|,$$
$$\Big|\Delta_m(r)-\widehat{\Delta}_m(r)\Big|\leqslant\underbrace{\Big|\min_{p'\in[p]_{\equiv_{\mathcal{Z}}}}J^*(r;p')-\min_{\widehat{p}'\in[\widehat{p}]_{\equiv_{\widehat{\mathcal{Z}}}}}J^*(r;\widehat{p}')\Big|}_{=:\mathcal{J}}+\Big|\widehat{J}^E(r)-J^{\pi^E}(r;p)\Big|,$$

where we have defined symbols $\mathcal{I},\mathcal{J}$. Next:

$$\begin{aligned}
\mathcal{I}&\overset{(1)}{\leqslant}\Big|\max_{p'\in[p]_{\equiv_{\mathcal{Z}}}}J^*(r;p')-\max_{\widehat{p}'\in[\widehat{p}]_{\equiv_{\mathcal{Z}}}}J^*(r;\widehat{p}')\Big|\\
&\overset{(2)}{=}\Big|\max_{p'\in\Delta^{\mathcal{S}}_{\mathcal{Z}^{\complement}}}J^*(r;p,p')-\max_{p'\in\Delta^{\mathcal{S}}_{\mathcal{Z}^{\complement}}}J^*(r;\widehat{p},p')\Big|\\
&\leqslant\max_{p'\in\Delta^{\mathcal{S}}_{\mathcal{Z}^{\complement}}}\Big|J^*(r;p,p')-J^*(r;\widehat{p},p')\Big|\\
&\leqslant\max_{p'\in\Delta^{\mathcal{S}}_{\mathcal{Z}^{\complement}}}\max_{\pi}\Big|J^{\pi}(r;p,p')-J^{\pi}(r;\widehat{p},p')\Big|\\
&\overset{(3)}{\leqslant}\max_{p'\in\Delta^{\mathcal{S}}_{\mathcal{Z}^{\complement}}}\max_{\pi}\sum_{(s,a,h)\in\mathcal{Z}}d^{p,p',\pi}_h(s,a)\Big|\sum_{s'\in\mathcal{S}}(p_h(s'|s,a)-\widehat{p}_h(s'|s,a))V^{\pi}_{h+1}(s';\widehat{p},p',r)\Big|,
\end{aligned}$$

where at (1) we use that, under $\mathcal{E}_3$ or $\mathcal{E}'_3$, we have $\widehat{\mathcal{Z}}=\mathcal{Z}$, at (2) we use the definitions of $[p]_{\equiv_{\mathcal{Z}}}$ and $[\widehat{p}]_{\equiv_{\mathcal{Z}}}$, and at (3) we apply Lemma 11. Note that we can show that also quantity $\mathcal{J}$ enjoys the same upper bound by using the property that $|\min_x f(x)-\min_x g(x)|\leqslant\max_x|f(x)-g(x)|$.

Now, when $|\mathcal{R}| = 1$, under the good event $\mathcal{E}$, we can upper bound the last term by:

$$\max_{p' \in \Delta^{\mathcal{S}}_{\mathcal{Z}^{\mathbb{C}}}} \max_{\pi} \sum_{(s,a,h) \in \mathcal{Z}} d_h^{p,p',\pi}(s,a) \cdot c \sqrt{\frac{H^3 \ln^2 \frac{12|\widehat{\mathcal{Z}}|}{\delta}}{\tau^b d_h^{p,\pi^b}(s,a)}} \leqslant cH \sqrt{\frac{H^3 \ln^2 \frac{12|\mathcal{Z}|}{\delta}}{\tau^b d_{\min}^{p,\pi^b}}} \leqslant \frac{\epsilon}{2},$$

from which we obtain the bound in the theorem for τ^b when $|\mathcal{R}| = 1$. For the bound for τ^E, we simply set the confidence bound of event $\mathcal{E}_1$ to be $\leqslant \epsilon/2$ and solve w.r.t. τ^E. The result follows through an application of Lemma 9.

For $|\mathcal{R}| \geqslant 1$, under good event $\mathcal{E}'$, by applying also the Pinsker's inequality, we can upper bound:

$$\max_{p' \in \Delta^{\mathcal{S}}_{\mathcal{Z}^{\mathbb{C}}}} \max_{\pi} \sum_{(s,a,h) \in \mathcal{Z}} d_h^{p,p',\pi}(s,a) \cdot cH \sqrt{\frac{\ln \frac{6|\widehat{\mathcal{Z}}|}{\delta} \beta(\tau^b, \delta/6)}{\tau^b d_h^{p,\pi^b}(s,a)}} \leqslant cH^2 \sqrt{\frac{\ln \frac{6|\mathcal{Z}|}{\delta} \beta(\tau^b, \delta/6)}{\tau^b d_{\min}^{p,\pi^b}}} \leqslant \frac{\epsilon}{2},$$

and solving w.r.t. τ^b using Lemma J.3 in Lazzati et al. (2024a) we obtain the bound in the theorem for τ^b when $|\mathcal{R}| \geqslant 1$. For the bound on τ^E, we simply impose that the confidence bound in event $\mathcal{E}'_1$ is smaller than $\epsilon/2$. Finally, we apply Lemma 10 to show that the guarantee holds with probability at least $1 - \delta$. This concludes the proof. ■

7. Discussion

In this section, we provide a discussion on the flexibility of the reward compatibility framework by presenting additional problem settings and extensions in which the adoption of reward compatibility for efficient learning is straightforward. Moreover, we discuss about some design choices made in this paper, and we collect additional insights about the practical "usage" of the rewards learned through IRL. We conclude with a discussion on the practical usefulness of our framework and bounds.

Reward Learning. In the context of Reward Learning (ReL), the learner receives a variety of expert feedbacks to learn the true reward function r^E (Jeon et al., 2020). From the "constraint" column of Table 2 in Jeon et al. (2020), we recognize that each feedback, similarly to the IRL feedback, provides a different kind of (inequality) constraint with which reducing the amount of rewards in $\mathfrak{R}$ that represent feasible candidates for r^E. It should be remarked that these constraints are "hard", in that a reward either satisfies the constraint or not (we might define a notion of feasible set as the set of rewards satisfying such constraints). To permit efficient learning when the transition model is unknown, our reward compatibility framework proposes to transform such "hard" constraints into "soft" constraints, by measuring the compatibility of the rewards with the constraints.

Other IRL settings. A popular alternative to the IRL setting with optimal expert (Ng and Russell, 2000), is that in which the expert is optimal in a certain *entropy-regularized MDP*

(Ziebart et al., 2010; Fu et al., 2017), that is,

$$\pi^E = \arg\max_{\pi\in\Pi} \mathop{\mathbb{E}}_{p,\pi}\left[\sum_{h\in[\![H]\!]} r_h(s_h,a_h) + \mathcal{H}(\pi_h(\cdot|s_h))\right] =: \arg\max_{\pi\in\Pi} \overline{J}^{\pi}(r;p),$$

where $\mathcal{H}$ denotes the entropy. In words, the objective of the expert consists in the maximization of the *entropy-regularised* expected utility $\overline{J}$. The advantage of this formulation is the existence of a unique optimal policy. To permit efficient learning when the dynamics is unknown, it is possible to extend our notion of reward compatibility to the maximum-entropy IRL framework as

$$\overline{\mathcal{C}}^{\mathrm{ENT}}_{p,\pi^E}(r) \coloneqq \max_{\pi} \overline{J}^{\pi}(r;p) - \overline{J}^{\pi^E}(r;p).$$

In this manner, we quantify the degree of compatibility of reward r with the given constraint. We observe that the same approach can be adopted also for other settings like, for instance, IRL with risk-sensitive agents (Lazzati and Metelli, 2025), once that the form of the objective of the expert has been defined. As a side comment, we would like to remark the difference between our notion of reward compatibility and existing IRL/IL objectives: reward compatibility represents an abstract measure of constraint satisfaction, i.e., it tells us how much a given reward complies with the (expert) information, while existing IRL/IL objectives (like the suboptimality gap that we considered throughout the paper) represent specific instantiations, which are useful depending on the problem setting at stake.

Robustness to misspecification. When the given constraint is not known exactly, our framework permits to be robust by accurately choosing the classification threshold Δ. For instance, when the expert's suboptimality is contained into $[L,U]$ for some $U \geqslant L \geqslant 0$, but we are uncertain about the specific values L, U, then we can use a larger classification threshold. Intuitively, the larger the classification threshold, the more uncertain we are on the feedback received.

Demonstrations in multiple environments. Consider the learning setting in which we are given many expert's demonstrations about the same reward r^E in a variety of environments (Amin and Singh, 2016; Cao et al., 2021). Clearly, multiple constraints reduce the partial identifiability of the problem and permit to retrieve a smaller feasible set. When the transition model is unknown, reward compatibility permits to cope with uncertainty in a straightforward way. For instance, let $\{\widehat{\mathcal{C}}_i(\cdot)\}_{i\in[\![N]\!]}$ be the estimated (non)compatibilities associated to demonstrations $\{\mathcal{D}_i^E\}_{i\in[\![N]\!]}$ in N environments. A meaningful objective consists in finding a reward r such that $\max_{i\in[\![N]\!]} \widehat{\mathcal{C}}_i(r) \leqslant \epsilon$, i.e., which is at most ϵ-(non)compatible (for some $\epsilon > 0$) with all the input demonstrations.

Multiplicative compatibility. Any reward $r \in \mathfrak{R}$ induces, in the considered environment $\mathcal{M}$ with dynamics p, an ordering in the space of policies Π, based on the performance $J^{\pi}(r;p)$ of each policy $\pi \in \Pi$. It is easy to notice that for any scaling and translation parameters $\alpha \in \mathbb{R}_{>0}, \beta \in \mathbb{R}$, the reward constructed as $r'(\cdot,\cdot) = \alpha r(\cdot,\cdot) + \beta$ induces the same ordering as r in the space of policies.[13] Nevertheless, the *additive* notion of (non)compatibility $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}$

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

in Definition 6 (setting with optimal expert), is such that, for any $r \in \mathfrak{R}$:

$$\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r+\beta) = \overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r) \qquad \forall \beta \in \mathbb{R},$$
$$\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(\alpha r) = \alpha \overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r) \neq \overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r) \qquad \forall \alpha \in \mathbb{R}_{>0}.$$

Simply put, the scale α of the reward matters, and rescaling the reward modifies the (non)compatibility. If we are interested in invariance to scale, we can define a *multiplicative* notion of compatibility,[14] denoted by $\mathcal{F}$ (defined only for non-negative rewards r and non-zero optimal performance $J^*(r;p)$) as

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

Clearly, the larger $\mathcal{F}_{\mathcal{M},\pi^E}(r)$, the closer is the performance of π^E to the optimal performance. By definition, we have

$$\mathcal{F}_{\mathcal{M},\pi^E}(\alpha r) = \mathcal{F}_{\mathcal{M},\pi^E}(r) \qquad \forall \alpha \in \mathbb{R}_{>0}$$
$$\mathcal{F}_{\mathcal{M},\pi^E}(r+\beta) \neq \mathcal{F}_{\mathcal{M},\pi^E}(r) \qquad \forall \beta \in \mathbb{R},$$

i.e., this definition does not care about the scaling α of the reward, but it is sensitive to the actual "location" β of the reward. Thus, intuitively, none of $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}, \overline{\mathcal{F}}_{\mathcal{M},\pi^E}$ can be seen as perfect. We prefer to use $\overline{\mathcal{C}}_{p,\pi^E}$ instead of $\mathcal{F}_{\mathcal{M},\pi^E}$ because (i) most of the RL literature prefers the additive notion of suboptimality instead of the multiplicative one, giving importance to the scale of the reward, and (ii) the additive notion of suboptimality is simpler to analyze from a theoretical viewpoint w.r.t. the multiplicative one.

When can a learned reward be "used" for (forward) RL? We are interested in applications of IRL like Apprenticeship Learning (AL). We say that a reward function r *can be "used" for (forward) RL* if the policy π obtained through the optimization of r performs acceptably under the true expert's reward r^E. What properties should the reward r, learned through IRL, satisfy in order to be "usable"? We now list and analyze some plausible requirements which are common in literature.

First, (i) we might ask that, being π^E optimal w.r.t. r^E, then any reward r such that $\pi^E \in \arg\max_\pi J^\pi(r)$ can be used, i.e., any reward in the feasible set $\mathcal{R}_{\mathcal{M},\pi^E}$ (Metelli et al., 2023). However, there are rewards $r \in \mathcal{R}_{\mathcal{M},\pi^E}$ that induce more than one optimal policy (e.g., both $\overline{\pi}, \pi^E$ as optimal), and optimal policies other than π^E (e.g., $\overline{\pi}$) are not guaranteed to perform well under r^E (actually, $\overline{\pi}$ can be an arbitrary policy in Π). Clearly, this is not satisfactory. Another possibility (ii) consists in rewards $r \in \mathcal{R}_{\mathcal{M},\pi^E}$ such that π^E is the *unique* optimal policy, similarly to what happens in entropy-regularized MDPs (Ziebart et al., 2010; Fu et al., 2017). However, in practice, due to computational limitations or uncertainty (e.g., estimated dynamics $\widehat{p}$), we can just afford to compute an ϵ-optimal policy for r in p. Since any policy can be ϵ-optimal under reward r, then we have no guarantee on its performance w.r.t. r^E. A third requirement (iii) asks for rewards r that make π^E at least ϵ-optimal, i.e., ϵ-compatible rewards based on Definition 6. However, since these

14. Multiplicative suboptimality has already been analysed in literature. E.g., see Theorem 7.2.7 in Puterman (1994), which is inspired by Ornstein (1969).

rewards represent a superset of the feasible set $\mathcal{R}_{\mathcal{M},\pi^E}$, then even this requirement is not satisfactory.

The three requirements described above do not provide guarantees that optimizing the considered reward r provides a policy with satisfactory performance w.r.t. the true r^E.

Remark 4 (Sufficient condition for reward usability)**.** *If we want to be sure that an ϵ-optimal policy for the learned reward r is at least $f(\epsilon)$-optimal for r^E (for some function f), then* it suffices that all the (at least) ϵ-optimal policies π under the learned r have visitation distribution close to that of π^E in 1-norm*:*

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

If we define distance d^{all} between rewards r, r' (see Section 3.1 of Zhao et al. (2024)) as

$$d^{\text{all}}(r,r') \coloneqq \sup_{\pi\in\Pi} |J^{\pi}(r;p) - J^{\pi}(r';p)|,$$

then we see that $d^{\text{all}}(r, r^E) \leqslant \epsilon$ for some small $\epsilon \geqslant 0$ represents a stronger condition for reward usability. Indeed, if $d^{\text{all}}(r, r^E)$ is small, then the performance of *any* policy as measured by r, not just optimal policy or ϵ-optimal policy, is similar to its performance as measured by r^E. Therefore, clearly, *rewards r with small distance to r^E w.r.t. d^{all}* can *be "used" for forward RL*. However, since expert demonstrations do not provide any information about the performance of policies other than π^E under r^E, we have the following result:

Proposition 12. *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, d_0, p)$ be a* known *MDP without reward, and let π^E be a* known *expert's policy. Let r^E the true* unknown *reward optimized by the expert to construct π^E. Then, there does not exist a learning algorithm that, for any $\epsilon, \delta \in (0,1)$, receives in input an arbitrary pair $(\mathcal{M}, \pi^E)$ and outputs a single reward r such that $d^{all}(r, r^E) \leqslant \epsilon$ w.p. $1-\delta$.*

Proof Sketch Simply, we can construct $\mathcal{M}$ to be an MDP without reward in which there is at least a policy $\pi \neq \pi^E$ such that $\mathcal{S}^{p,\pi} \cap \mathcal{S}^{p,\pi^E} = \varnothing$, i.e., the states visited by the two policies are different. Then, in the feasible set $\mathcal{R}_{\mathcal{M},\pi^E}$, there are both the rewards r', r'' that give $J^{\pi}(r';p) = 0$ and $J^{\pi}(r'';p) = H$. Since we do not know the performance of π under r^E, then both rewards are plausible. For this reason, whatever the output r of the algorithm, the distance $d^{\text{all}}(r, r^E)$ (note that d^{all} is a metric) cannot be smaller than $H/2$ in the worst case. By taking $\delta > 1/2$ and using triangle inequality, we can prove the result. ■

Nevertheless, Metelli et al. (2023); Lazzati et al. (2024a); Zhao et al. (2024) *seem* to provide sample efficient algorithms w.r.t. d^{all}.[15] By looking at Proposition 12, we realize that this is clearly a *contradiction*. What is the right interpretation? The trick is that the algorithms proposed in Metelli et al. (2023); Lazzati et al. (2024a); Zhao et al. (2024) are *not* able to output a single reward r which is close to r^E w.r.t. d^{all}, but, *for any possible reward*

15. Actually, Metelli et al. (2023); Lazzati et al. (2024a) use different notions of distance, like $d_\infty(r,r') \coloneqq \|r - r'\|_\infty$. However, we can write $\|r - r'\|_\infty \geqslant \|r - r'\|_1/(SAH)$, and by dual norms we have that $d^{\text{all}}(r,r') = \sup_{\pi\in\Pi} |\langle d^{p,\pi}, r - r'\rangle| \leqslant \sup_{\overline{d}:\|\overline{d}\|_\infty\leqslant 1} |\langle \overline{d}, r - r'\rangle| = \|r - r'\|_1$. Therefore, the guarantees of Metelli et al. (2023); Lazzati et al. (2024a) can be converted to d^{all} guarantees too.

$r^E = r^E(V, A)$ *parametrized by some value and advantage functions* V, A,[16] *they are able to output a reward* r *such that* $d^{all}(r, r^E(V, A))$ *is small.* In other words, it is like if these works *assume to know* the V, A parametrization of the true reward r^E. In this way, these works are able to output a reward r that can be used for "forward" RL, otherwise their algorithms cannot provide such guarantee.

Practical Insights. Our framework and algorithms can be used in practice to perform search in the reward space for the specific reward that we desire in a provably efficient manner. Any search algorithm (e.g., gradient based) can be used coupled with our algorithms. Indeed, a search algorithm starts from some initial reward r_0, and then proceeds through candidate iterates $r_1, ..., r_t, ...$ up to converge to some reward r_T. To make this search algorithm provably efficient, the simplest way is to guarantee that our estimate of (non)compatibility of each of these intermediate rewards r_t is accurate, and this is exactly the kind of guarantee provided by our framework/algorithms. Regarding the search method, it is important to note that our framework is not constrained to a specific method. Indeed, IRL is an ill-posed problem, and so the specific reward that we would like to compute from the set of compatible rewards (and so the method to find it) should be different depending on the application domain or additional prior information. Lastly, we mention that it is possible to search over the space of rewards in a realistic setting through gradient methods in a provably efficient way. For instance, if we care about just any arbitrary compatible reward, then note that, since the notion of (non)compatibility $\overline{\mathcal{C}}_{\mathcal{M},\pi^E}(r) = \max_{d^{p,\pi}} \langle d^{p,\pi}, r \rangle - J^{\pi^E}(r; p)$ is a convex function of r, then we can follow the gradient direction $d^{p,\pi^*} - d^{p,\pi^E}$ with gradient descent to find a compatible reward, and our algorithms give the sample complexity of the process (modulo some difference for rescaling ϵ to keep into account the approximation error). See Lazzati and Metelli (2025) for an example of this. As another example, in case we have additional (possibly convex) information on the reward function that we want to find (e.g., obtained through PbRL or some other human feedback), then we can still use a gradient method (but correcting the gradient direction for keeping into account the additional information), but our sample complexity bounds of this paper still hold. See Lazzati and Metelli (2026) for this.

8. Related Work

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

Feasible Set. Let us begin with works related to the feasible set. While the notion of feasible set has been introduced implicitly in Ng and Russell (2000), the first paper that analyses the sample complexity of estimating the feasible set in online IRL is Metelli et al. (2021). Authors in Metelli et al. (2021) adopt the simple generative model in tabular MDPs, and devise two sample efficient algorithms. Lindner et al. (2022) focuses on the same problem as Metelli et al. (2021), but adopts a forward model in tabular MDPs. By

16. While Zhao et al. (2024) makes this parametrization explicit, Metelli et al. (2023); Lazzati et al. (2024a) keep the parametrization implicit, but everything is analogous.

adopting RFE exploration algorithms, they devise sample efficient algorithms. However, as remarked in Zhao et al. (2024), the learning framework considered in Lindner et al. (2022) suffers from a major issue. Metelli et al. (2023) builds upon Metelli et al. (2021) to construct the first minimax lower bound for the problem of estimating the feasible set using a generative model. The lower bound is in the order of $\Omega\big(\frac{H^3SA}{\epsilon^2}(S + \log\frac{1}{\delta})\big)$, where S and A are the cardinality of the state and action spaces, H is the horizon, ϵ is the accuracy and δ the failure probability. In addition, Metelli et al. (2023) develops US-IRL, an efficient algorithm whose sample complexity matches the lower bound. Poiani et al. (2024) analyze a setting analogous to that of Metelli et al. (2023), in which there is availability of a single optimal expert and multiple suboptimal experts with known suboptimality. Lazzati et al. (2024a) analyse the problem of estimating the feasible set when no active exploration of the environment is allowed, but the learner is given a batch data set collected by some behavior policy π^b. In particular, Lazzati et al. (2024a) focus on two novel learning targets that are suited for the offline setting, i.e., a subset and a superset of the feasible set, and demonstrate that such sets are the tightest learnable subset and superset of the feasible set. They conclude by proposing a pessimistic algoroithm, PIRLO, to estimate them. Zhao et al. (2024) analyses the same offline setting as Lazzati et al. (2024a), but instead of focusing on the notion of feasible set directly, it considers the notion of reward mapping, which considers reward functions as parametrized by their value and advantage functions, and whose image coincides with the feasible set.

Other sample-efficient IRL settings. With regards to IRL works that do not consider the feasible set, we mention Lopes et al. (2009), which analyses an active learning framework for IRL. However, Lopes et al. (2009) assumes that the transition model is known, and its goal is to estimate the expert policy only. Works Komanduru and Honorio (2019) and Komanduru and Honorio (2021) provide, respectively, an upper bound and a lower bound to the sample complexity of IRL for β-strict separable problems in the tabular setting. However, both the setting considered and the bound obtained are fairly different from ours. Analogously, Dexter et al. (2021) provides a sample efficient IRL algorithm for β-strict separable problems with continuous state space. However, their setting is different from ours since they assume that the system can be modelled using a basis of orthonormal functions.

Identifiability and Reward Learning. As aforementioned, the IRL problem is ill-posed, thus, to retrieve a single reward, additional constraints shall be imposed. Amin and Singh (2016) analyse the setting in which demonstrations of an optimal policy for the same reward function are provided across environments with different transition models. In this way, authors can reduce the experimental unidentifiability, and recover the state-only reward function. Cao et al. (2021) and Kim et al. (2020) concern reward identifiability but in entropy-regularized MDPs (Ziebart et al., 2010; Fu et al., 2017). Such setting is in some sense easier than the common IRL setting, because entropy-regularization permits a unique optimal policy for any reward function. Cao et al. (2021) use expert demonstrations from multiple transition models and multiple discount factors to retrieve the reward function, while Kim et al. (2020) analyse properties of the dynamics of the MDP to increase the constraints. With regards to the more general field of Reward Learning (ReL), we mention Jeon et al. (2020), which introduce a framework that formalizes the constraints imposed by various kinds of human feedback (like demonstrations or preferences, Wirth et al., 2017).

Intuitively, multiple feedbacks about the same reward represent additional constraints beyond mere demonstrations. Skalse et al. (2023) characterize the partial identifiability of the reward function based on various reward learning data sources.

Online Apprenticeship Learning. The first works that provide a theoretical analysis of the AL setting when the transition model is unknown are Abbeel and Ng (2005) and Syed and Schapire (2007). Recently, Shani et al. (2022) formulate the online AL problem, which resembles the online IRL problem. The main difference is that in online AL the ultimate goal is to imitate the expert, while in IRL is to recover a reward function. Xu et al. (2023) improve the results in Shani et al. (2022) by combining an RFE algorithm with an efficient algorithm for the estimation of the visitation distribution of the deterministic expert's policy in tabular MDPs, presented in Rajaraman et al. (2020). We mention also Rajaraman et al. (2021); Swamy et al. (2022) for the sample complexity of estimating the expert's policy in problems with linear function approximation. In the context of Imitation Learning from Observation alone (ILfO) (Liu et al., 2018), Sun et al. (2019) propose a probably efficient algorithm for large-scale MDPs with unknown transition model. Liu et al. (2022) provide an efficient AL algorithm based on GAIL (Ho and Ermon, 2016) in Linear Kernel Episodic MDPs Zhou et al. (2021) with unknown transition model.

Others. We mention Klein et al. (2012) , which consider a classification approach for IRL. However, this is fairly different from our IRL problem formulation in Sections 4 and 6.

9. Conclusion

Motivated by major limitations of the feasible reward set as a unifying framework for sample-efficient IRL, in this paper, we have presented the powerful framework of *reward compatibility*, which permits efficient learning in many IRL problems. The major advantage of reward compatibility is its flexibility, since it can easily be adapted to a multitude of problem settings. In this paper, we have considered both an optimal and a suboptimal expert, and we have analysed online IRL in tabular and Linear MDPs, and offline IRL in tabular MDPs only, and we have provided sample efficient algorithms, `CATY-IRL` and `CATY-OFF-IRL`, for solving the newly-proposed IRL classification problem. We have also discussed on the flexibility of the framework introduced in a variety of *complementary* settings, and provided some insights about the usage of reward functions learned through IRL for (forward) RL.

Limitations. In this work, we have not analysed the offline IRL problem in Linear MDPs. In particular, as already mentioned, we note that the approach with the best and worst notions of (non)compatibility that we adopted cannot be extended straightforwardly to linear MDPs, because it is based on a notion of coverage best-suited for tabular MDPs. For this reason, an original definition of reward compatibility that permits to use existing notions of coverage of the state space for linear MDPs has to be devised, in order to analyse the offline IRL problem in linear MDPs through our framework. In addition, we acknowledge that the empirical validation of the proposed algorithms is beyond the scope of this work. Our focus has been on developing the theoretical framework and foundational aspects, leaving the empirical evaluation for future research.

Future Directions. Promising directions for future works concern the extension of the analysis of the *reward compatibility* framework beyond Linear MDPs to general function approximation and to the offline setting. Moreover, it would be fascinating to extend the notion of reward compatibility to other kinds of expert feedback (in the context of ReL), and to other IRL settings (e.g., Boltzmann rational experts), as discussed in Section 7. In such way, we believe that it will be possible to bring IRL closer to real-world applications.

Acknowledgments

This publication was funded with the contribution of Ministero dell'Università e della ricerca pursuant to D.D. n. 7206 of 17 April 2025 - BANDO FIS 2. Project FIS-2023-02598 (Starting Grant), title: "Unified Learning from Diverse Human Feedback" (HUmLrn). CUP: D53C25000710001.

References

P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning 21 (ICML)*, 2004.

P. Abbeel and A. Y. Ng. Exploration and apprenticeship learning in reinforcement learning. In *International Conference on Machine Learning 22 (ICML)*, 2005.

P. Abbeel, A. Coates, M. Quigley, and A. Ng. An application of reinforcement learning to aerobatic helicopter flight. In *Advances in Neural Information Processing Systems 19 (NeurIPS)*, 2006.

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

K. Amin and S. Singh. Towards resolving unidentifiability in inverse reinforcement learning, 2016.

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

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

M. G. Azar, R. Munos, and H. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013.

M. Barnes, M. Abueg, O. F. Lange, M. Deeds, J. Trader, D. Molitor, M. Wulfmeier, and S. O'Banion. Massively scalable inverse reinforcement learning in Google Maps, 2024.

H. Cao, S. Cohen, and L. Szpruch. Identifiability in inverse reinforcement learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 12362–12373, 2021.

C.-A. Cheng, A. Kolobov, and A. Agarwal. Policy improvement via imitation of multiple oracles, 2020.

G. Dexter, K. Bello, and J. Honorio. Inverse reinforcement learning in a continuous state space with formal guarantees. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 6972–6982, 2021.

C. Finn, S. Levine, and P. Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning 33 (ICML)*, volume 48, pages 49–58, 2016.

T. Freihaut and G. Ramponi. On multi-agent inverse reinforcement learning, 2024.

J. Fu, K. Luo, and S. Levine. Learning robust rewards with adversarial inverse reinforcement learning. In *International Conference on Learning Representations 5 (ICLR)*, 2017.

D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, 2016.

D. Hadfield-Menell, S. Milli, P. Abbeel, S. J. Russell, and A. Dragan. Inverse reward design. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, 2016.

F. Jarboui and V. Perchet. Offline inverse reinforcement learning, 2021.

H. J. Jeon, S. Milli, and A. Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 4415–4426, 2020.

N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning 34 (ICML)*, volume 70, pages 1704–1713, 2017.

C. Jin, A. Krishnamurthy, M. Simchowitz, and T. Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning 37 (ICML)*, volume 119, pages 4870–4879, 2020a.

C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory 33 (COLT)*, volume 125, pages 2137–2143, 2020b.

C. Jin, Q. Liu, and S. Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 13406–13418, 2021a.

Y. Jin, Z. Yang, and Z. Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning 38 (ICML)*, volume 139, pages 5084–5096, 2021b.

M. Jing, X. Ma, W. Huang, F. Sun, C. Yang, B. Fang, and H. Liu. Reinforcement learning from imperfect demonstrations under soft expert guidance, 2019.

M. S. Kakade. *On the Sample Complexity of Reinforcement Learning.* PhD thesis, 2003.

E. Kaufmann, P. Ménard, O. Darwiche Domingues, A. Jonsson, E. Leurent, and M. Valko. Adaptive reward-free exploration. In *International Conference on Algorithmic Learning Theory 32 (ALT)*, volume 132, pages 865–891, 2021.

K. Kim, Y. Gu, J. Song, S. Zhao, and S. Ermon. Domain adaptive imitation learning. In *International Conference on Machine Learning 37 (ICML)*, volume 119, pages 5286–5295, 2020.

E. Klein, M. Geist, B. Piot, and O. Pietquin. Inverse reinforcement learning through structured classification. In *Advances in Neural Information Processing Systems 25 (NeurIPS)*, 2012.

A. Komanduru and J. Honorio. On the correctness and sample complexity of inverse reinforcement learning. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

A. Komanduru and J. Honorio. A lower bound for the sample complexity of inverse reinforcement learning. In *International Conference on Machine Learning 38 (ICML)*, volume 139, pages 5676–5685, 2021.

A. Kurenkov, A. Mandlekar, R. Martin-Martin, S. Savarese, and A. Garg. AC-Teach: A Bayesian actor-critic method for policy learning with an ensemble of suboptimal teachers, 2019.

F. Lazzati and A. M. Metelli. Learning utilities from demonstrations in Markov decision processes. In *International Conference on Machine Learning 42 (ICML)*, volume 267, pages 32704–32770, 2025.

F. Lazzati and A. M. Metelli. Robustness in the face of partial identifiability in reward learning. In *International Conference on Learning Representations 14 (ICLR)*, volume 2026, pages 107353–107390, 2026.

F. Lazzati, M. Mutti, and A. M. Metelli. Offline inverse RL: New solution concepts and provably efficient algorithms. In *International Conference on Machine Learning 41 (ICML)*, 2024a.

F. Lazzati, M. Mutti, and A. M. Metelli. How does inverse RL scale to large state spaces? a provably efficient approach. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2024b.

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

X. Liu, T. Yoneda, C. Wang, M. R. Walter, and Y. Chen. Active policy improvement from multiple black-box oracles, 2023.

Y. Liu, A. Gupta, P. Abbeel, and S. Levine. Imitation from observation: Learning to imitate behaviors from raw video via context translation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1118–1125, 2018.

Z. Liu, Y. Zhang, Z. Fu, Z. Yang, and Z. Wang. Learning from demonstration: Provably efficient adversarial policy imitation with linear function approximation. In *International Conference on Machine Learning 39 (ICML)*, volume 162, pages 14094–14138, 2022.

M. Lopes, F. Melo, and L. Montesano. Active learning for reward estimation in inverse reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 31–46, 2009.

P. Menard, O. D. Domingues, A. Jonsson, E. Kaufmann, E. Leurent, and M. Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning 38 (ICML)*, volume 139, pages 7599–7608, 2021.

A. M. Metelli, G. Ramponi, A. Concetti, and M. Restelli. Provably efficient learning of transferable rewards. In *International Conference on Machine Learning 38 (ICML)*, volume 139, pages 7665–7676, 2021.

A. M. Metelli, F. Lazzati, and M. Restelli. Towards theoretical understanding of inverse reinforcement learning. In *International Conference on Machine Learning 40 (ICML)*, volume 202, pages 24555–24591, 2023.

B. Michini, M. Cutler, and J. P. How. Scalable reward learning from demonstration. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 303–308, 2013.

G. Neu and C. Szepesvári. Apprenticeship learning using inverse reinforcement learning and gradient methods. In *Conference on Uncertainty in Artificial Intelligence 23 (UAI 2007)*, pages 295–302, 2007.

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

D. Ornstein. On the existence of stationary optimal strategies. *Proceedings of the American Mathematical Society*, 20(2):563–569, 1969.

R. Poiani, G. Curti, A. M. Metelli, and M. Restelli. Inverse reinforcement learning with sub-optimal experts, 2024.

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

N. Rajaraman, L. Yang, J. Jiao, and K. Ramchandran. Toward the fundamental limits of imitation learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 2914–2924, 2020.

N. Rajaraman, Y. Han, L. Yang, J. Liu, J. Jiao, and K. Ramchandran. On the value of interaction and function approximation in imitation learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, volume 34, pages 1325–1336, 2021.

D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. In *International Joint Conference on Artifical Intelligence 20 (IJCAI)*, pages 2586–2591, 2007.

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

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

L. Shani, T. Zahavy, and S. Mannor. Online apprenticeship learning. In *AAAI Conference on Artificial Intelligence 36 (AAAI)*, pages 8240–8248, 2022.

J. M. V. Skalse, M. Farrugia-Roberts, S. Russell, A. Abate, and A. Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *International Conference on Machine Learning 40 (ICML)*, volume 202, pages 32033–32058, 2023.

W. Sun, A. Vemula, B. Boots, and D. Bagnell. Provably efficient imitation learning from observation alone. In *International Conference on Machine Learning 36 (ICML)*, volume 97, pages 6036–6045, 2019.

G. Swamy, N. Rajaraman, M. Peng, S. Choudhury, J. Bagnell, S. Z. Wu, J. Jiao, and K. Ramchandran. Minimax optimal online imitation learning via replay estimation. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, volume 35, pages 7077–7088, 2022.

U. Syed and R. E. Schapire. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems 20 (NeurIPS)*, 2007.

A. J. Wagenmaker, Y. Chen, M. Simchowitz, S. Du, and K. Jamieson. Reward-free RL is no harder than reward-aware RL in linear Markov decision processes. In *International Conference on Machine Learning 39 (ICML)*, volume 162, pages 22430–22456, 2022.

R. Wang, D. P. Foster, and S. M. Kakade. What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations 8 (ICLR)*, 2020a.

R. Wang, R. Salakhutdinov, and L. F. Yang. Reinforcement learning with general value function approximation: provably efficient approach via bounded eluder dimension. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 6123–6135, 2020b.

C. Wirth, R. Akrour, G. Neumann, and J. Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18:1–46, 2017.

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

T. Xu, Z. Li, Y. Yu, and Z. Luo. Provably efficient adversarial imitation learning with unknown transitions. In *Conference on Uncertainty in Artificial Intelligence 39 (UAI)*, volume 216, pages 2367–2378, 2023.

L. Yang and M. Wang. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning 36 (ICML)*, volume 97, pages 6995–7004, 2019.

B. Yue, J. Li, and G. Liu. Provably efficient exploration in inverse constrained reinforcement learning, 2024.

L. Zhao, M. Wang, and Y. Bai. Is inverse reinforcement learning harder than standard reinforcement learning? In *International Conference on Machine Learning 41 (ICML)*, 2024.

D. Zhou, Q. Gu, and C. Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes. In *Conference on Learning Theory 34 (COLT)*, pages 4532–4576, 2021.

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

B. D. Ziebart, J. A. Bagnell, and A. K. Dey. Modeling interaction via the principle of maximum causal entropy. In *International Conference on Machine Learning 27 (ICML)*, pages 1255–1262, 2010.