Generalizing Behavior via Inverse Reinforcement Learning with Closed-Form Reward Centroids

Filippo Lazzati **Alberto Maria Metelli**
Politecnico di Milano, Milan, Italy

Abstract

We study the problem of generalizing an expert agent's behavior, provided through demonstrations, to new environments and/or additional constraints. Inverse Reinforcement Learning (IRL) offers a promising solution by seeking to recover the expert's underlying *reward function*, which, if used for planning in the new settings, would reproduce the desired behavior. However, IRL is inherently *ill-posed*: multiple reward functions, forming the so-called *feasible set*, can explain the same observed behavior. Since these rewards may induce different policies in the new setting, in the absence of additional information, a decision criterion is needed to select which policy to deploy. In this paper, we propose a novel, principled criterion that selects the "average" policy among those induced by the rewards in a certain *bounded* subset of the feasible set. Remarkably, we show that this policy can be obtained by planning with the reward *centroid* of that subset, for which we derive a *closed-form expression*. We then present a provably efficient algorithm for estimating this centroid using an *offline* dataset of expert demonstrations only. Finally, we conduct *numerical simulations* that illustrate the relationship between the expert's behavior and the behavior produced by our method.

1 INTRODUCTION

Imitation Learning (IL) (Pomerleau, 1988; Abbeel and Ng, 2004) addresses the problem of *efficiently learning a desired behavior by imitating an expert's behavior* (Osa et al., 2018). By avoiding the difficult task of specifying a reward function, IL offers a direct and effective way to construct artificial agents with human-like capabilities using demonstrations of behavior alone (Ho and Ermon, 2016; Fu et al., 2017; Reddy et al., 2019; Garg et al., 2021). Consequently, it has been successfully applied to a variety of real-world applications and control problems, including autonomous driving (Kuderer et al., 2015), healthcare (Yu et al., 2019), motion planning (Finn et al., 2016), and path planning (Ratliff et al., 2006; Silver et al., 2010).

However, in many situations, the environment where the learner must operate has different dynamics or requires additional constraints compared to the environment where the expert's demonstrations were collected (Fu et al., 2017; Arora and Doshi, 2021; Schlaginhaufen and Kamgarpour, 2023). A common example are "sim-to-real" scenarios, where demonstrations are easily obtained in a simulator while the target environment is the (possibly more constrained) real world (OpenAI et al., 2019; Matas et al., 2018; Desai et al., 2020). In such cases, the expert's demonstrations provide valuable information about the desired behavior, but this information must be extrapolated, as merely transferring the expert's observed policy can lead to failure (Fu et al., 2017; Ni et al., 2021; Arora and Doshi, 2021).

Inverse Reinforcement Learning (IRL) (Russell, 1998; Ng and Russell, 2000) addresses this *generalization* problem by assuming the existence of an *expert's reward function* that, if used for planning in the new environment, would induce the desired behavior (Amin and Singh, 2016; Fu et al., 2017; Arora and Doshi, 2021; Rolland et al., 2022; Schlaginhaufen and Kamgarpour, 2024). The task is thus reduced to inferring the expert's reward function from demonstrations of behavior. This inference requires knowledge of the true relationship between the expert's policy and reward, referred to as the *model of behavior* (Skalse and Abate, 2023). Common models of behavior studied in the literature include assuming that the expert's policy is optimal for the expert's reward (Ng and Russell, 2000), soft-optimal (Ziebart, 2010; Fu et al., 2017), ϵ-optimal

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

(Poiani et al., 2024), the best policy satisfying certain constraints (Schlaginhaufen and Kamgarpour, 2023), or derived from a softmax over the optimal Q-function (Ramachandran and Amir, 2007).

Unfortunately, regardless of the chosen model of behavior, even if the expert's policy is known exactly, the expert's reward cannot be *uniquely* identified (Ng and Russell, 2000; Cao et al., 2021). Instead, there exists a set of equally plausible reward functions, known as the *feasible set*, that are consistent with the expert's behavior (Metelli et al., 2021, 2023). Crucially, these rewards *generalize differently*, meaning they induce different optimal policies in the target environment (Skalse et al., 2023), leading to ambiguity about which behavior should be deployed. To address this, many works in the literature aim to reduce the size of the feasible set using additional data or structural assumptions (Amin and Singh, 2016; Fu et al., 2017; Cao et al., 2021; Kim et al., 2021; Rolland et al., 2022; Schlaginhaufen and Kamgarpour, 2023, 2024).

In this work, we consider a more challenging and realistic setting where no additional data or structure is available, and we are requested to make the best possible decision given only the demonstrations. In this context, existing IRL approaches can be broadly grouped into methods that work well in the best-case, worst-case, or average-case. "Best-case" methods (Ziebart et al., 2008; Choi and Kim, 2011; Wulfmeier et al., 2016; Finn et al., 2016) arbitrarily select a reward from the feasible set and deploy its corresponding policy. "Worst-case" methods (Lazzati and Metelli, 2025) deploy the policy that minimizes the worst-case suboptimality over all rewards in the feasible set. "On-average" methods (Ramachandran and Amir, 2007; Michini and How, 2012; Bajgar et al., 2024) minimize expected suboptimality under a prior reward distribution.

Unsurprisingly, "best-case" methods tend to be *overly optimistic* and often fail to generalize, as observed in empirical studies such as (Finn et al., 2016; Fu et al., 2017; Viano et al., 2021). Conversely, "worst-case" methods can be *overly conservative*, and the resulting policy can be inefficient to compute (Lazzati and Metelli, 2025). Positioned between these extremes, "on-average" methods offer a more reasonable compromise. However, the common choices of priors considered in the literature either rely on *domain knowledge* or turn out to be *biased* in the set of policies that the expert can demonstrate (see Section 3.2), making them intuitively unsuitable for our specific setting, where no additional information is available.

In this paper, we introduce novel priors over rewards that are *unbiased* with respect to the policy that the expert will demonstrate. Building on them, we develop *provably efficient* "on-average" algorithms for generalizing the observed expert's behavior to new environments and constraints.

Contributions. We summarize the main contributions of this work below:

- We categorize existing IRL-based approaches for generalizing behavior into three groups, and we demonstrate that the most popular reward prior considered in the literature is *biased* (Section 3).
- We propose novel *unbiased* reward priors for three common models of behavior (Section 4.1), and we derive *closed-form* expressions for their centroids restricted to the feasible sets (Section 4.2).
- We develop provably efficient algorithms to estimate these centroids from *offline* data (Section 5).
- Finally, we conduct *numerical simulations* to illustrate the relationship between the expert's behavior and that produced by our method (Section 6).

Proofs are provided in Appendices C-F, and additional related work is discussed in Appendix A.

2 PRELIMINARIES

Notation. Given an integer $N \in \mathbb{N}$, we let $[\![N]\!] := \{1, 2, \dots, N\}$. Given two finite sets $\mathcal{X}, \mathcal{Y}$, we denote by $\Delta^{\mathcal{X}}$ and $\Delta^{\mathcal{X}}_{\mathcal{Y}}$, respectively, the set of probability measures on $\mathcal{X}$ and the set of functions from $\mathcal{Y}$ to $\Delta^{\mathcal{X}}$. For vectors $x \in \mathbb{R}^n$, we define $\|x\|_\infty := \max_i |x_i|$. Given a matrix W, we let $\det(W)$ be its determinant. We use notation $a \propto b$ to denote proportionality between two quantities a, b. Given a Riemannian manifold $\mathcal{X} \subseteq \mathbb{R}^n$ with dimensionality $k \leqslant n$ (Lee, 2019), we denote the integral of a function $f : \mathcal{X} \to \mathbb{R}$ over $\mathcal{X}$ as $\int_{\mathcal{X}} f(x)dx$, where dx represents the k-dimensional volume element of the manifold. We denote by $\mathrm{vol}(\mathcal{X}) := \int_{\mathcal{X}} dx$ the k-dimensional volume of $\mathcal{X}$. The *centroid* $C(\mathcal{X})$ is a generalization of the arithmetic mean: $C(\mathcal{X}) := \int_{\mathcal{X}} x dx / \mathrm{vol}(\mathcal{X}) \propto \int_{\mathcal{X}} x dx$. Appendix B provides additional notation for the supplementary material.

Markov Decision Processes (MDPs). An infinite-horizon discounted Markov Decision Process (MDP) (Puterman, 1994) is defined as a tuple $\mathcal{M}_r := (\mathcal{S}, \mathcal{A}, s_0, p, \gamma, r)$, where $\mathcal{S}$ is the finite state space ($S := |\mathcal{S}|$), $\mathcal{A}$ is the finite action space ($A := |\mathcal{A}|$), $s_0 \in \mathcal{S}$ is the initial state, $p \in \Delta^{\mathcal{S}}_{\mathcal{S}\times\mathcal{A}}$ is the transition model, $\gamma \in [0, 1)$ is the discount factor, and $r \in \mathbb{R}^{SA}$ is the reward function. We will use symbol $\mathcal{M} := (\mathcal{S}, \mathcal{A}, s_0, p, \gamma)$ to denote an MDP without reward (MDP\R), i.e., an MDP-like environment where there is no notion of reward (Abbeel and Ng, 2004). A policy $\pi \in \Pi := \Delta^{\mathcal{A}}_{\mathcal{S}}$ is a strategy that prescribes actions in states. We let $\mathbb{P}_{\mathcal{M},\pi}$ represent the probability

distribution induced by playing π in $\mathcal{M}_r$ or $\mathcal{M}$. We let $d_{\mathcal{M},\pi}(s,a) \coloneqq (1-\gamma)\sum_{t=0}^{+\infty}\gamma^t \mathbb{P}_{\mathcal{M},\pi}(s_t = s, a_t = a)$ be the state-action occupancy measure of π in $\mathcal{M}$, and $d_{\mathcal{M},\pi}(s) \coloneqq \sum_{a\in\mathcal{A}} d_{\mathcal{M},\pi}(s,a)$ be the state-only occupancy measure. We denote by $\mathcal{S}_{\mathcal{M},\pi}$ the set of states reachable with non-zero probability from s_0 in $\mathcal{M}$ when playing policy π. Formally: $\mathcal{S}_{\mathcal{M},\pi} \coloneqq \{s \in \mathcal{S} \,|\, d_{\mathcal{M},\pi}(s) > 0\}$. We define $W_{p,\pi} \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}}$ as the linear map that, to every pair of states $s, s' \in \mathcal{S}$, associates $W_{p,\pi}(s,s') = \mathbb{1}\{s' = s\} - \gamma \sum_{a\in\mathcal{A}} \pi(a|s)p(s'|s,a)$.

Value and advantage functions. We denote the value and Q-functions induced by a policy π in $\mathcal{M}_r$ as $V^\pi(s;p,r) \coloneqq \mathbb{E}_{\mathcal{M},\pi}\big[\sum_{t=0}^{+\infty}\gamma^t r(s_t,a_t)|s_0 = s\big]$ and $Q^\pi(s,a;p,r) \coloneqq \mathbb{E}_{\mathcal{M},\pi}\big[\sum_{t=0}^{+\infty}\gamma^t r(s_t,a_t)|s_0 = s, a_0 = a\big]$, where $\mathbb{E}_{\mathcal{M},\pi}$ denotes the expectation w.r.t. $\mathbb{P}_{\mathcal{M},\pi}$. Moreover, we let $V^\pi_\lambda(s;p,r) \coloneqq \mathbb{E}_{\mathcal{M},\pi}\big[\sum_{t=0}^{+\infty}\gamma^t\big(r(s_t,a_t) - \lambda\log\pi(a_t|s_t)\big)|s_0 = s\big]$ and $Q^\pi_\lambda(s,a;p,r) \coloneqq \mathbb{E}_{\mathcal{M},\pi}\big[\sum_{t=0}^{+\infty}\gamma^t\big(r(s_t,a_t) - \lambda\log\pi(a_t|s_t)\big)|s_0 = s, a_0 = a\big]$ denote the *soft* value and Q-functions associated to a policy π where $\lambda \geqslant 0$ is the regularization coefficient (Ziebart, 2010; Haarnoja et al., 2017). It is useful to introduce notation for the advantage and soft-advantage functions as $A^\pi(s,a;p,r) \coloneqq Q^\pi(s,a;p,r) - V^\pi(s;p,r)$ and $A^\pi_\lambda(s,a;p,r) \coloneqq Q^\pi_\lambda(s,a;p,r) - V^\pi_\lambda(s;p,r)$. The optimal counterparts of all these quantities will be denoted using a star $*$; namely, we let $V^*(s;p,r) \coloneqq \max_\pi V^\pi(s;p,r)$, $Q^*(s,a;p,r) \coloneqq \max_\pi Q^\pi(s,a;p,r)$, and $A^*(s,a;p,r) \coloneqq Q^*(s,a;p,r) - V^*(s;p,r)$ be the optimal value, Q and advantage functions, and $V^*_\lambda(s;p,r) \coloneqq \max_\pi V^\pi_\lambda(s;p,r)$, $Q^*_\lambda(s,a;p,r) \coloneqq \max_\pi Q^\pi_\lambda(s,a;p,r)$ and $A^*_\lambda(s,a;p,r) \coloneqq Q^*_\lambda(s,a;p,r) - V^*_\lambda(s;p,r)$ the soft-optimal value, Q and advantage functions. Finally, we define $\Pi^*(p,r) \coloneqq \{\pi \in \Pi \,|\, \forall s \in \mathcal{S} : V^\pi(s;p,r) = V^*(s;p,r)\}$.

Sets of rewards. Given an MDP\R $\mathcal{M}$, a set of states $\overline{\mathcal{S}} \subseteq \mathcal{S}$ and a policy π deterministic on $\overline{\mathcal{S}}$, we define the set of rewards that make π optimal in $\overline{\mathcal{S}}$ as:

$$\mathcal{R}^{\mathrm{OPT},\overline{\mathcal{S}}}_{\mathcal{M},\pi} \coloneqq \big\{r \in \mathbb{R}^{SA} \,\big|\, \forall s \in \overline{\mathcal{S}} : \pi(s) \in \arg\max_{a\in\mathcal{A}} Q^*(s,a;p,r)\big\}. \tag{1}$$

Similarly, given a policy π stochastic on $\overline{\mathcal{S}}$, we define the set of rewards that make π respectively soft-optimal (for some $\lambda > 0$) and Boltzmann w.r.t. Q^* (for some $\beta > 0$) in $\overline{\mathcal{S}}$ as:

$$\mathcal{R}^{\mathrm{MCE},\overline{\mathcal{S}}}_{\mathcal{M},\pi} \coloneqq \big\{r \in \mathbb{R}^{SA} \,\big|\, \forall s \in \overline{\mathcal{S}}, \forall a \in \mathcal{A} : \pi(a|s) \propto e^{\frac{1}{\lambda}Q^*_\lambda(s,a;p,r)}\big\}, \tag{2}$$

$$\mathcal{R}^{\mathrm{BIRL},\overline{\mathcal{S}}}_{\mathcal{M},\pi} \coloneqq \big\{r \in \mathbb{R}^{SA} \,\big|\, \forall s \in \overline{\mathcal{S}}, \forall a \in \mathcal{A} : \pi(a|s) \propto e^{\frac{1}{\beta}Q^*(s,a;p,r)}\big\}. \tag{3}$$

The meaning of the acronyms OPT, MCE, BIRL will be clear in the next paragraph.

Inverse Reinforcement Learning (IRL). In IRL (Russell, 1998; Ng and Russell, 2000), we are given demonstrations from an expert policy π_E in an MDP\R $\mathcal{M} \coloneqq (\mathcal{S}, \mathcal{A}, s_0, p, \gamma)$, and we aim to recover the expert's reward $r_E \in \mathbb{R}^{SA}$. To complete the formulation of an IRL problem, we need to specify a *model of behavior* (Skalse and Abate, 2023), namely, the relationship between π_E and r_E that an IRL algorithm must invert. In this paper, we focus on the popular models of behavior introduced by (Ng and Russell, 2000), (Ziebart, 2010) and (Ramachandran and Amir, 2007), that we abbreviate respectively with OPT, MCE and BIRL.

In OPT (OPTimal) (Ng and Russell, 2000), the expert's policy π_E is assumed to be optimal in $\mathcal{M}_{r_E}$:

$$\pi_E \in \arg\max_{\pi\in\Pi} V^\pi(s_0;p,r_E). \tag{4}$$

In MCE (Maximum Causal Entropy) (Ziebart, 2010), π_E is assumed to be soft-optimal in $\mathcal{M}_{r_E}$ for some $\lambda > 0$:

$$\pi_E = \arg\max_{\pi\in\Pi} V^\pi_\lambda(s_0;p,r_E). \tag{5}$$

In BIRL (Bayesian IRL) (Ramachandran and Amir, 2007), π_E is assumed to be softmax w.r.t. Q^* in $\mathcal{M}_{r_E}$ for some $\beta > 0$:

$$\pi_E(a|s) \propto e^{\frac{1}{\beta}Q^*(s,a;p,r_E)} \qquad \forall s \in \mathcal{S}_{\mathcal{M},\pi_E}, \forall a \in \mathcal{A}. \tag{6}$$

It is well-known that the IRL problem is *ill-posed* (Ng and Russell, 2000; Cao et al., 2021; Skalse et al., 2023), i.e., given any expert's policy π_E and a model of behavior $m \in \{\mathrm{OPT},\mathrm{MCE},\mathrm{BIRL}\}$, we are not able to uniquely retrieve the expert's reward r_E, because there is a set of rewards $\mathcal{R}^m_{\mathcal{M},\pi_E}$, known as the *feasible set* (Metelli et al., 2021, 2023), that satisfy the constraint defined by π_E and m (i.e., Eqs. 4, 5 or 6). It can be shown that the feasible sets $\mathcal{R}^{\mathrm{OPT}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\mathrm{MCE}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\mathrm{BIRL}}_{\mathcal{M},\pi_E}$ coincide, respectively, with $\mathcal{R}^{\mathrm{OPT},\mathcal{S}_{\mathcal{M},\pi_E}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\mathrm{MCE},\mathcal{S}_{\mathcal{M},\pi_E}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\mathrm{BIRL},\mathcal{S}_{\mathcal{M},\pi_E}}_{\mathcal{M},\pi_E}$ (see Appendix B.1). Finally, we remark that these sets are *unbounded* subsets of $\mathbb{R}^{SA}$ (Ng and Russell, 2000; Cao et al., 2021; Skalse et al., 2023).

3 PROBLEM SETTING AND EXISTING METHODS

In this section, we formalize the problem setting (Section 3.1) and review existing IRL-based approaches that address it (Section 3.2). Later, in Section 4, we introduce our proposed approach.

3.1 Problem Setting

We consider the setting where we are given a dataset $\mathcal{D}_E \sim \pi_E$ of trajectories collected by an expert's policy π_E in an MDP\R $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_0, p, \gamma)$, and we *know* the model of behavior $m \in \{\text{OPT,MCE,BIRL}\}$ that relates π_E to the *unknown* expert's reward r_E. Informally, our objective is to "generalize the observed expert's behavior" $\mathcal{D}_E$ to a new MDP\R $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, s_0', p', \gamma')$ under a cost constraint given by $c \in \mathbb{R}^{SA}$ and a budget $k \geqslant 0$. Formally, we aim to find a policy $\widehat{\pi}^*$ that satisfies the cost constraint $V^\pi(s_0'; p', c) \leqslant k$ and whose suboptimality in $\mathcal{M}'_{r_E}$ is as small as possible:[1]

$$\widehat{\pi}^* \in \arg\min_{\pi' \in \Pi_{c,k}} \Big(\max_{\pi \in \Pi_{c,k}} V^\pi(s_0'; p', r_E) - V^{\pi'}(s_0'; p', r_E) \Big), \quad (7)$$

where $\Pi_{c,k} := \{\pi \in \Pi \,|\, V^\pi(s_0'; p', c) \leqslant k\}$ is the set of feasible policies. For simplicity, we assume here that π_E is *known* on its support $\mathcal{S}_{\mathcal{M},\pi_E}$, corresponding to an infinite dataset $\mathcal{D}_E$ of expert's demonstrations. We will remove this assumption in Section 5. Finally, for convenience, we define $\pi_{\min} := \min_{(s,a) \in \mathcal{S} \times \mathcal{A}} \pi_E(a|s) > 0$ for MCE and BIRL.

3.2 Existing Methods

As explained in Section 2, no behavioral model $m \in \{\text{OPT,MCE,BIRL}\}$ allows to uniquely identify r_E from π_E alone. Therefore, we cannot directly optimize Eq. (7) as r_E is unknown. *In the absence of additional information*, the best we can do is replace the objective in Eq. (7) with a *surrogate* that does not depend on the unknown r_E, but instead leverages the fact that r_E belongs to the known feasible set $\mathcal{R}^m_{\mathcal{M},\pi_E}$. Below, we review common surrogate objectives considered in the literature.

"Best-case" methods (Ziebart et al., 2008; Choi and Kim, 2011; Wulfmeier et al., 2016; Finn et al., 2016). These approaches adopt an *optimistic* surrogate objective:

$$\widehat{\pi}_{\text{B}} \in \arg\min_{\pi' \in \Pi_{c,k}} \min_{r \in \mathcal{R}^m_{\mathcal{M},\pi_E}} \Big(\max_{\pi \in \Pi_{c,k}} V^\pi(s_0'; p', r) - V^{\pi'}(s_0'; p', r) \Big), \quad (8)$$

which can be optimized by first computing *any* reward r in the feasible set, and then deriving the policy $\widehat{\pi}_{\text{B}}$ it induces. In the *best-case* scenario where $r = r_E$, the resulting policy $\widehat{\pi}_{\text{B}}$ successfully generalizes the expert's behavior. However, as noted in empirical studies such as (Finn et al., 2016; Fu et al., 2017; Viano et al., 2021), this approach is often *overly optimistic*, and the resulting policy $\widehat{\pi}_{\text{B}}$ may fail to accurately reflect the expert's true preferences.

"Worst-case" methods (Lazzati and Metelli, 2025). Recently, (Lazzati and Metelli, 2025) proposed a *conservative* surrogate objective:

$$\widehat{\pi}_{\text{W}} \in \arg\min_{\pi' \in \Pi_{c,k}} \max_{r \in \mathcal{R}^m_{\mathcal{M},\pi_E} \cap \mathfrak{R}} \Big(\max_{\pi \in \Pi_{c,k}} V^\pi(s_0'; p', r) - V^{\pi'}(s_0'; p', r) \Big), \quad (9)$$

where $\mathfrak{R} := [-1, +1]^{SA}$ is the unit hypercube. Intuitively, $\widehat{\pi}_{\text{W}}$ minimizes the suboptimality under the most adversarial value that r_E can take on inside the feasible set $\mathcal{R}^m_{\mathcal{M},\pi_E}$, thereby representing the best policy in the *worst-case*. Note that restricting the otherwise *unbounded* set $\mathcal{R}^m_{\mathcal{M},\pi_E}$ to a bounded subset like $\mathfrak{R}$ is necessary to prevent the objective from becoming infinite for any feasible policy (see Appendix C.1). However, beyond potentially being *too conservative*, this approach has the problem that it is not clear what *bias* is introduced by using $\mathfrak{R}$ instead of another bounded subset of $\mathbb{R}^{SA}$, and also that Eq. (9) is generally *computationally hard*, as it involves maximizing a convex function, i.e., the pointwise maximum of linear functions (Boyd and Vandenberghe, 2004).

"On-average" methods (Ramachandran and Amir, 2007; Michini and How, 2012; Bajgar et al., 2024). These approaches select a function $w : \mathbb{R}^{SA} \to [0, 1]$ that assigns a weight to every reward, and then compute a policy $\widehat{\pi}_w$ minimizing the *average* suboptimality with respect to the rewards in the feasible set, *weighted by* w:[2]

$$\widehat{\pi}_w \in \arg\min_{\pi' \in \Pi_{c,k}} \int_{\mathcal{R}^m_{\mathcal{M},\pi_E}} w(r) \Big(\max_{\pi \in \Pi_{c,k}} V^\pi(s_0'; p', r) - V^{\pi'}(s_0'; p', r) \Big) dr, \quad (10)$$

This "on-average" formulation offers a reasonable compromise between the optimism of "best-case" methods and the conservatism of "worst-case" ones. Note that this approach is adopted by (Ramachandran and Amir, 2007; Michini and How, 2012; Bajgar et al., 2024) in a *Bayesian* fashion, by interpreting w as a *prior* distribution for r_E. Since our problem setting assumes no domain knowledge, w shall be as *uninformative* as possible. To this end, (Ramachandran and Amir, 2007) proposed the function $\overline{w}$ that assigns *equal weight* to all the rewards in the unit hypercube $\mathfrak{R} := [-1, +1]^{SA}$:[3]

$$\overline{w}(r) := \begin{cases} 1 & \text{if } r \in \mathfrak{R} \\ 0 & \text{otherwise} \end{cases}, \quad (11)$$

[1] Note that Eq. (7) can be equivalently rewritten as $\widehat{\pi}^* \in \arg\max_{\pi' \in \Pi_{c,k}} V^{\pi'}(s_0'; p', r_E)$, but the form in Eq. (7) is more convenient for the presentation of Section 3.2.

[2] Note that dr represents the volume element of the manifold $\mathcal{R}^m_{\mathcal{M},\pi_E}$ (see Section 2).

[3] Although (Ramachandran and Amir, 2007) focus on state-only rewards, our considerations still apply (see Appendix C.3).

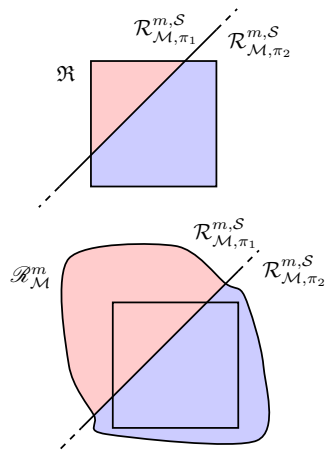

Figure 1: (Top) In general, the volumes of the red ($\mathfrak{R} \cap \mathcal{R}^{m,\mathcal{S}}_{\mathcal{M},\pi_1}$) and blue ($\mathfrak{R} \cap \mathcal{R}^{m,\mathcal{S}}_{\mathcal{M},\pi_2}$) regions are different. (Bottom) The set $\mathscr{R}^m_{\mathcal{M}}$ contains $\mathfrak{R}$ and expands $\mathfrak{R} \cap \mathcal{R}^{m,\mathcal{S}}_{\mathcal{M},\pi_1}$ and $\mathfrak{R} \cap \mathcal{R}^{m,\mathcal{S}}_{\mathcal{M},\pi_2}$ differently, so that the new red ($\mathscr{R}^m_{\mathcal{M}} \cap \mathcal{R}^{m,\mathcal{S}}_{\mathcal{M},\pi_1}$) and blue ($\mathscr{R}^m_{\mathcal{M}} \cap \mathcal{R}^{m,\mathcal{S}}_{\mathcal{M},\pi_2}$) regions have the same volumes.

as the constant prior $w(r) = 1 \ \forall r \in \mathbb{R}^{SA}$ is improper and cannot be used, since the feasible set $\mathcal{R}^m_{\mathcal{M},\pi_E}$ is unbounded and the integral in Eq. (10) would diverge. At first glance, $\overline{w}$ appears natural and uninformative. Interpreted in a Bayesian fashion, it models an expert whose reward r_E is uniformly distributed over $\mathfrak{R}$. Surprisingly, however, $\overline{w}$ turns out to be *biased* in the space of policies. To see this, let $m = \text{OPT}$ and define, for every deterministic policy π:

$$\overline{w}_\pi := \int_{\mathcal{R}^{\text{OPT},\mathcal{S}}_{\mathcal{M},\pi}} \overline{w}(r) dr = \text{vol}\big(\mathfrak{R} \cap \mathcal{R}^{\text{OPT},\mathcal{S}}_{\mathcal{M},\pi}\big). \tag{12}$$

$\overline{w}_\pi$ can be Bayesianly interpreted as the (unnormalized) probability that a reward drawn from $\overline{w}$ makes π optimal, i.e., it corresponds to *the probability that an OPT expert demonstrates policy π in $\mathcal{M}$* regardless of the initial state. As such, to be truly uninformative, we desire that all deterministic policies should be equally "likely" under $\overline{w}$. However, it turns out that this is not the case:

Proposition 3.1. *There exists an MDP\R $\mathcal{M}$ and two deterministic policies $\pi_1, \pi_2 \in \Pi$ such that:*

$$vol\big(\mathfrak{R} \cap \mathcal{R}^{OPT,\mathcal{S}}_{\mathcal{M},\pi_1}\big) \neq vol\big(\mathfrak{R} \cap \mathcal{R}^{OPT,\mathcal{S}}_{\mathcal{M},\pi_2}\big).$$

Proposition 3.2. *There exists an MDP\R $\mathcal{M}$ and two stochastic policies $\pi_1, \pi_2 \in \Pi$ such that, for all $m \in \{MCE, BIRL\}$:*

$$vol\big(\mathfrak{R} \cap \mathcal{R}^{m,\mathcal{S}}_{\mathcal{M},\pi_1}\big) \neq vol\big(\mathfrak{R} \cap \mathcal{R}^{m,\mathcal{S}}_{\mathcal{M},\pi_2}\big).$$

This situation is exemplified in Fig. 1 (top).[4] Roughly speaking, the issue is that, being biased, $\overline{\pi}$ effectively encodes some sort of prior knowledge, which we do *not* possess.

[4] In Appendix C.4, we provide related results considering also a uniform prior over transition models.

4 OUR APPROACH

In the previous section, we saw that "on-average" methods, by balancing optimism and conservatism through a prior over rewards w, should be the preferred approach for generalizing expert behavior. However, we also observed that the commonly used prior $\overline{w}$ introduces a bias in the space of policies, by assigning different weights to different policies. In this section, we propose novel *unbiased* priors over rewards $w^m_{\mathcal{M}}$ (Section 4.1) and leverage them to design efficient algorithms for generalizing expert behavior (Section 4.2).

4.1 Novel Priors over Rewards

Although the prior $\overline{w}$ is biased in the policy space, it has appealing properties: it is simple to represent, centered around the zero reward, and assigns equal weight to all rewards in the considered set ($\mathfrak{R}$). To retain these advantages while removing the bias, we propose replacing $\overline{w}$, the uniform prior over the hypercube $\mathfrak{R}$, with $w^m_{\mathcal{M}}$ (defined in Eq. 14 below), the uniform prior over the set $\mathscr{R}^m_{\mathcal{M}}$ (see Eq. 13), which resembles $\mathfrak{R}$ (Proposition 4.1) but avoids its bias (Propositions 4.2-4.4).

Specifically, for any $m \in \{\text{OPT,MCE,BIRL}\}$ and MDP\R $\mathcal{M}$, we define $\mathscr{R}^m_{\mathcal{M}}$ as the set of rewards whose (soft-)optimal value and advantage functions are bounded by positive constants $C^m_1, C^m_2 > 0$:

$$\begin{aligned}
\mathscr{R}^{\text{OPT}}_{\mathcal{M}} &:= \Big\{ r \in \mathbb{R}^{SA} \;\Big|\; \forall \pi \in \Pi^*(p,r), \forall s,a : \\
&\quad |V^\pi(s;p,r)| \leqslant C^{\text{OPT}}_1 k_\pi \wedge |A^\pi(s,a;p,r)| \leqslant C^{\text{OPT}}_2 \Big\}, \\
\mathscr{R}^{\text{MCE}}_{\mathcal{M}} &:= \Big\{ r \in \mathbb{R}^{SA} \;\Big|\; \forall s,a : \\
&\quad |V^*_\lambda(s;p,r)| \leqslant C^{\text{MCE}}_1 \wedge |A^*_\lambda(s,a;p,r)| \leqslant C^{\text{MCE}}_2 \Big\}, \\
\mathscr{R}^{\text{BIRL}}_{\mathcal{M}} &:= \Big\{ r \in \mathbb{R}^{SA} \;\Big|\; \forall s,a : \\
&\quad |V^*(s;p,r)| \leqslant C^{\text{BIRL}}_1 \wedge |A^*(s,a;p,r)| \leqslant C^{\text{BIRL}}_2 \Big\},
\end{aligned} \tag{13}$$

where $k_\pi := |\det(W_{p,\pi})|^{-1/S}$ for all π. Note that these sets replace the max-norm bound on the reward (as in $\mathfrak{R}$) with bounds on the max norm of the induced value and advantage functions. With appropriate choices of C^m_1, C^m_2, the sets $\mathscr{R}^m_{\mathcal{M}}$ remain close in size to the hypercube $\mathfrak{R}$, as desired:

Proposition 4.1. *Let $\mathcal{M}$ be any MDP\R. If we choose $C^{OPT}_1, C^{OPT}_2 \geqslant (1+\gamma)/(1-\gamma) =: \rho$, $C^{MCE}_1, C^{MCE}_2 \geqslant (2 + \lambda \log A)/(1-\gamma)$, and $C^{BIRL}_1, C^{BIRL}_2 \geqslant 1/(1-\gamma)$,*

then, for any $m \in \{OPT, MCE, BIRL\}$*:*

$$\mathfrak{R} \subseteq \mathscr{R}^m_{\mathcal{M}} \subseteq [-\rho C^m_1 - C^m_2, +\rho C^m_1]^{SA}.$$

In addition, the sets $\mathscr{R}^m_{\mathcal{M}}$ are not biased in the space of policies:[5]

Proposition 4.2. *For any MDP\R* $\mathcal{M}$ *and any pair of deterministic policies* $\pi_1, \pi_2 \in \Pi$*, it holds that:*

$$vol(\mathscr{R}^{OPT}_{\mathcal{M}} \cap \mathcal{R}^{OPT,\mathcal{S}}_{\mathcal{M},\pi_1}) = vol(\mathscr{R}^{OPT}_{\mathcal{M}} \cap \mathcal{R}^{OPT,\mathcal{S}}_{\mathcal{M},\pi_2}).$$

Proposition 4.3. *For any MDP\R* $\mathcal{M}$ *and any pair of stochastic policies* $\pi_1, \pi_2 \in \Pi$ *such that* $\pi_1(a|s) \geqslant \pi_{\min}$ *and* $\pi_2(a|s) \geqslant \pi_{\min}$ *in every* s, a*, if* $C^{MCE}_2 \geqslant \lambda \log \frac{1}{\pi_{\min}}$*, then:*

$$vol(\mathscr{R}^{MCE}_{\mathcal{M}} \cap \mathcal{R}^{MCE,\mathcal{S}}_{\mathcal{M},\pi_1}) = vol(\mathscr{R}^{MCE}_{\mathcal{M}} \cap \mathcal{R}^{MCE,\mathcal{S}}_{\mathcal{M},\pi_2}).$$

Proposition 4.4. *For any MDP\R* $\mathcal{M}$ *and any pair of stochastic policies* $\pi_1, \pi_2 \in \Pi$ *such that* $\pi_1(a|s) \geqslant \pi_{\min}$ *and* $\pi_2(a|s) \geqslant \pi_{\min}$ *in every* s, a*, if* $C^{BIRL}_2 \geqslant \beta \log \frac{1}{A\pi_{\min}}$*, then:*

$$vol(\mathscr{R}^{BIRL}_{\mathcal{M}} \cap \mathcal{R}^{BIRL,\mathcal{S}}_{\mathcal{M},\pi_1}) = vol(\mathscr{R}^{BIRL}_{\mathcal{M}} \cap \mathcal{R}^{BIRL,\mathcal{S}}_{\mathcal{M},\pi_2}).$$

Simply put, the sets $\mathscr{R}^m_{\mathcal{M}}$ expand the hypercube $\mathfrak{R}$ in directions that ensure all policies of interest for m are equally-represented in $\mathcal{M}$, as illustrated on the bottom of Fig. 1. Note that the additional in MCE and BIRL is mild, as C^m_2 can be chosen arbitrarily. We are now ready to define our new priors $w^m_{\mathcal{M}}$ as uniform distributions over $\mathscr{R}^m_{\mathcal{M}}$. Formally, for any $m \in \{\text{OPT,MCE,BIRL}\}$ and MDP\R $\mathcal{M}$, we define $w^m_{\mathcal{M}}$ as:

$$w^m_{\mathcal{M}}(r) := \begin{cases} 1 & \text{if } r \in \mathscr{R}^m_{\mathcal{M}} \\ 0 & \text{otherwise} \end{cases}. \tag{14}$$

At first glance, it might seem odd to use different priors for different $\mathcal{M}$ and m. However, as discussed in Section 3.2, we do *not* assume that $r_E \sim w^m_{\mathcal{M}}$. Instead, we use $w^m_{\mathcal{M}}$ as a mathematical tool that allows us to replace Eq. (7) with a surrogate objective that is less optimistic than Eq. (8), less conservative than Eq. (9), and still assigns equal weight to all relevant policies.

4.2 Algorithms for Generalizing Behavior via Closed-Form Reward Centroids

To generalize expert behavior using $w^m_{\mathcal{M}}$, we need to optimize Eq. (10) with $w = w^m_{\mathcal{M}}$. To this end:

Proposition 4.5. *For every* $w : \mathbb{R}^{SA} \to [0, 1]$*, let* $r_{w,\pi_E} = \int_{\mathcal{R}^m_{\mathcal{M},\pi_E}} w(r)dr$*. Then, Eq.* (10) *can be rewritten as:*

$$\widehat{\pi}_w \in \arg\max_{\pi' \in \Pi_{c,k}} V^{\pi'}(s'_0; p', r_{w,\pi_E}). \tag{15}$$

[5]In Appendix D.3, we show that these properties migh not hold in $\mathcal{M}' \neq \mathcal{M}$.

Let $\widehat{\pi}^m_{\mathcal{M}} := \widehat{\pi}_{w^m_{\mathcal{M}}}$ and $r^m_{\mathcal{M},\pi_E} := r_{w^m_{\mathcal{M}},\pi_E}$. Proposition 4.5 tells us that to compute the policy $\widehat{\pi}^m_{\mathcal{M}}$ that generalizes expert behavior using our new prior $w^m_{\mathcal{M}}$, we can first compute the (unnormalized) *centroid* $r^m_{\mathcal{M},\pi_E}$ of the feasible set:

$$r^m_{\mathcal{M},\pi_E} = \int_{\mathscr{R}^m_{\mathcal{M}} \cap \mathcal{R}^m_{\mathcal{M},\pi_E}} r dr,$$

and then perform (constrained) planning on it. Since there exists a rich body of literature on (constrained) planning algorithms (see (Puterman, 1994; Sutton and Barto, 2018)), the problem reduces to computing $r^m_{\mathcal{M},\pi_E}$. For doing so, one possibility is to use the Markov Chain Monte Carlo (MCMC)-based algorithms introduced in (Ramachandran and Amir, 2007; Michini and How, 2012; Bajgar et al., 2024) for approximating the mean reward r_{w,π_E} for almost any w. However, we can do significantly better. Interestingly, it turns out that, for any model of behavior m, MDP\R $\mathcal{M}$ and expert's policy π_E, the reward centroid $r^m_{\mathcal{M},\pi_E}$ admits a *closed-form expression*:

Theorem 4.6 (OPT)**.** *For any MDP\R* $\mathcal{M}$ *and deterministic policy* π_E*, the reward centroid* $r^{OPT}_{\mathcal{M},\pi_E}$ *after rescaling*[6] *is,* $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$*:*

$$r^{OPT}_{\mathcal{M},\pi_E}(s, a) = \begin{cases} 1 & \text{if } s \in \mathcal{S}_{\mathcal{M},\pi_E} \wedge \pi_E(s) = a \\ 0 & \text{if } s \in \mathcal{S}_{\mathcal{M},\pi_E} \wedge \pi_E(s) \neq a \\ \frac{1}{A} & \text{if } s \notin \mathcal{S}_{\mathcal{M},\pi_E} \end{cases}. \tag{16}$$

Theorem 4.7 (MCE)**.** *Let* $C^{MCE}_2 \geqslant \lambda \log \frac{1}{\pi_{\min}}$*. Then, for any MDP\R* $\mathcal{M}$ *in which* $\mathcal{S}_{\mathcal{M},\pi_E} = \mathcal{S}$*, the reward centroid* $r^{MCE}_{\mathcal{M},\pi_E}$ *after rescaling is,* $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$*:*

$$r^{MCE}_{\mathcal{M},\pi_E}(s, a) = \log \pi_E(a|s). \tag{17}$$

Theorem 4.8 (BIRL)**.** *Let* $C^{BIRL}_2 \geqslant \beta \log \frac{1}{A\pi_{\min}}$*. Then, for any MDP\R* $\mathcal{M}$ *in which* $\mathcal{S}_{\mathcal{M},\pi_E} = \mathcal{S}$*, the reward centroid* $r^{BIRL}_{\mathcal{M},\pi_E}$ *after rescaling is,* $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$*:*

$$r^{BIRL}_{\mathcal{M},\pi_E}(s, a) = \log \frac{\pi_E(a|s)}{\max_{a' \in \mathcal{A}} \pi_E(a'|s)}. \tag{18}$$

Some observations are in order. First, observe that the centroids depend only on π_E and the set of visited states $\mathcal{S}_{\mathcal{M},\pi_E}$. As a result, we will be able to efficiently estimate them from offline data $\mathcal{D}_E$ alone (see Section 5). Next, note that $r^{\text{OPT}}_{\mathcal{M},\pi_E}$, $r^{\text{MCE}}_{\mathcal{M},\pi_E}$ and $r^{\text{BIRL}}_{\mathcal{M},\pi_E}$ all assign higher rewards to actions that the expert executes with larger probability in the visited states $\mathcal{S}_{\mathcal{M},\pi_E}$. Intuitively, this means that they favor behaviors that aim to "replay" the actions demonstrated by the expert as much possible, as we will see in Section 6. Observe

[6]By "rescaling" a vector $x \in \mathbb{R}^n$ we mean applying a transformation $\alpha x + \beta$ with $\alpha > 0, \beta \in \mathbb{R}$. It is well-known that rescaling a reward function does not affect the rank of policies (Ng et al., 1999; Sutton and Barto, 2018).

Algorithm 1: OPT

Input: Data $\mathcal{D}_E$
$\widehat{\mathcal{S}} \leftarrow \{s | \exists i, h : s_h^i = s\}$
$\widehat{\mathcal{Z}} \leftarrow \{s, a | \exists i, h : s_h^i = s, a_h^i = a\}$
$\widehat{r}^{\mathrm{OPT}}(s,a) = \begin{cases} 1 & \text{if } (s,a) \in \widehat{\mathcal{Z}} \\ \frac{1}{A} & \text{if } s \notin \widehat{\mathcal{S}} \\ 0 & \text{otherwise} \end{cases}$
Return $\widehat{r}^{\mathrm{OPT}}$

also that we rescale the centroids just to obtain simpler equivalent forms. Finally, we require $\mathcal{S}_{\mathcal{M},\pi_E} = \mathcal{S}$ for MCE and BIRL to exclude unnecessarily complex cases where some states are unreachable by any policy. We conclude this section by showing that the centroids $r^m_{\mathcal{M},\pi_E}$ can be effectively used also for IL:[7]

Proposition 4.9. *Let $m \in \{OPT, MCE, BIRL\}$ be the model of behavior between π_E and r_E in an MDP$\backslash R\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_0, p, \gamma)$. Let $\pi^* \in \arg\max_\pi V^\pi(s_0; p, r^m_{\mathcal{M},\pi_E})$ be an optimal policy in $\mathcal{M}$ under $r^m_{\mathcal{M},\pi_E}$. Then, π^* is optimal also under r_E: $V^{\pi^*}(s_0; p, r_E) = V^*(s_0; p, r_E)$.*

Simply put, Proposition 4.9 shows that if the relation between π_E and r_E is accurately modeled by m, then planning with $r^m_{\mathcal{M},\pi_E}$ yields the same optimal policy as planning with r_E, a result that is not immediate especially for MCE and BIRL.

5 SAMPLE COMPLEXITY ANALYSIS

In this section, we drop the assumption that π_E is *known* in its support $\mathcal{S}_{\mathcal{M},\pi_E}$, and focus on estimating the centroids $r^{\mathrm{OPT}}_{\mathcal{M},\pi_E}$, $r^{\mathrm{MCE}}_{\mathcal{M},\pi_E}$ and $r^{\mathrm{BIRL}}_{\mathcal{M},\pi_E}$, from a dataset $\mathcal{D}_E = \{\tau^i\}_{i \in [\![N]\!]} = \{(s_1^i, a_1^i, s_2^i, a_2^i, \dots, s_H^i, a_H^i)\}_{i \in [\![N]\!]}$ of N trajectories, each of length H, collected by π_E in $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_0, p, \gamma)$. For the analysis, it is convenient to define $p^{\min,H}_{\mathcal{M},\pi_E}$ as the minimum probability with which π_E visits any state in its support in the first H steps:

$$p^{\min,H}_{\mathcal{M},\pi_E} := \min_{s \in \mathcal{S}_{\mathcal{M},\pi_E}} \mathbb{P}_{\mathcal{M},\pi_E}\big(\exists t \in [\![H]\!] : s_t = s\big). \quad (19)$$

Our algorithm for estimating $r^{\mathrm{OPT}}_{\mathcal{M},\pi_E}$ is reported in Algorithm 1. In short, we construct sets $\widehat{\mathcal{S}}$ and $\widehat{\mathcal{Z}}$ containing, respectively, all states and state-action pairs that have been visited by the expert in $\mathcal{D}_E$, and then we use them in place of $\mathcal{S}_{\mathcal{M},\pi_E}$ and π_E. This algorithm enjoys the following guarantees:

Theorem 5.1. *Let $\delta \in (0,1)$ and $H \geqslant S$. If π_E is deterministic, then, with probability at least $1-\delta$, it holds that $\widehat{r}^{OPT} = r^{OPT}_{\mathcal{M},\pi_E}$ provided that:*

$$NH \geqslant \frac{H \log \frac{|\mathcal{S}_{\mathcal{M},\pi_E}|}{\delta}}{p^{\min,H}_{\mathcal{M},\pi_E}}.$$

We remark that Theorem 5.1 provides the number of samples necessary to estimate the reward $r^{\mathrm{OPT}}_{\mathcal{M},\pi_E}$ *exactly*. Note that, since the available trajectories are H stages long, we need to assume that $H \geqslant S$ to ensure that $p^{\min,H}_{\mathcal{M},\pi_E} > 0$ (see Lemma F.2). Moreover, there is a trade-off in choosing H. Increasing it improves the probability $p^{\min,H}_{\mathcal{M},\pi_E}$, but also increases the numerator H, and viceversa. Regarding MCE and BIRL, we present our estimators in Algorithms 2 and 3. For every s, a, they both compute $N(s,a)$ as the number of trajectories $\tau^i \in \mathcal{D}_E$ in which state s is visited *at least once* and, at the *first visit*, the expert played action a.[8] Formally, let $t(s,i)$ denote the first stage $h \in [\![H]\!]$ at which $s_h^i = s$, or -1 otherwise. Then:

$$N(s,a) := \sum_{i \in [\![N]\!]} \mathbb{1}\{t(s,i) \neq -1 \wedge a^i_{t(s,i)} = a\}. \quad (20)$$

Next, both Algorithms 2 and 3 estimate the expert's policy π_E by clipping the empirical frequency $N(s,a)/N(s)$ using a threshold parameter $\pi'_{\min} > 0$ to avoid evaluating $\log 0$, and finally they estimate the rewards $r^{\mathrm{MCE}}_{\mathcal{M},\pi_E}$ and $r^{\mathrm{BIRL}}_{\mathcal{M},\pi_E}$ by taking the logarithm. We can show the following:

Algorithm 2: MCE

Input: Data $\mathcal{D}_E$, parameter $\pi'_{\min}$
$N(s,a) \leftarrow$ Eq. (20)
$N(s) \leftarrow \sum_{a \in \mathcal{A}} N(s,a)$
$\widehat{\pi}(a|s) \leftarrow \max\{\pi'_{\min}, \frac{N(s,a)}{\max\{1, N(s)\}}\}$
$\widehat{r}^{\mathrm{MCE}}(s,a) \leftarrow \log \widehat{\pi}(a|s)$
Return $\widehat{r}^{\mathrm{MCE}}$

Algorithm 3: BIRL

Input: Data $\mathcal{D}_E$, parameter $\pi'_{\min}$
$N(s,a) \leftarrow$ Eq. (20)
$N(s) \leftarrow \sum_{a \in \mathcal{A}} N(s,a)$
$\widehat{\pi}(a|s) \leftarrow \max\{\pi'_{\min}, \frac{N(s,a)}{\max\{1, N(s)\}}\}$
$\widehat{r}^{\mathrm{BIRL}}(s,a) \leftarrow \begin{cases} \log \frac{\widehat{\pi}(a|s)}{\max_{a'} \widehat{\pi}(a'|s)} & \text{if } N(s) > 0, \\ \log \pi'_{\min} & \text{if } N(s) = 0. \end{cases}$
Return $\widehat{r}^{\mathrm{BIRL}}$

Theorem 5.2. *Let $\epsilon, \delta \in (0,1)$ and $H \geqslant S$. Under the conditions of Theorem 4.7, if $\pi_{\min} \geqslant \pi'_{\min}$, then, with probability $1-\delta$, we have $\|\widehat{r}^{MCE} - r^{MCE}_{\mathcal{M},\pi_E}\|_\infty \leqslant \epsilon$*

[7] See Appendix E.3 for a result on $r^{\mathrm{OPT}}_{\mathcal{M},\pi_E}$ and $r^{\mathrm{BIRL}}_{\mathcal{M},\pi_E}$ in $\mathcal{M}' \neq \mathcal{M}$ in absence of constraints.

[8] This estimator simplifies the theoretical analysis. In practice, we count all occurrences of s, a.

provided that:

$$NH \geqslant \frac{H \log^2 \frac{4SA}{\delta}}{p^{\min,H}_{\mathcal{M},\pi_E}} \cdot \frac{16}{\epsilon^2 \pi'_{\min}}.$$

Theorem 5.3. *Let $\epsilon, \delta \in (0,1)$ and $H \geqslant S$. Under the conditions of Theorem 4.8, if $\pi_{\min} \geqslant \pi'_{\min}$, then, with probability $1-\delta$, we have $\|\widehat{r}^{BIRL} - r^{BIRL}_{\mathcal{M},\pi_E}\|_\infty \leqslant \epsilon$ provided that:*

$$NH \geqslant \frac{H \log^2 \frac{8SA}{\delta}}{p^{\min,H}_{\mathcal{M},\pi_E}} \cdot \frac{33}{\epsilon^2 \pi'_{\min}}.$$

Theorems 5.2 and 5.3 establish sample complexity rates comparable to that of OPT in terms of H, $p^{\min,H}_{\mathcal{M},\pi_E}$ and $\log \frac{1}{\delta}$. In addition, they depend linearly on $\frac{1}{\pi_{\min}}$ and $\frac{1}{\epsilon^2}$ since they need to estimate π_E and not only $\mathcal{S}_{\mathcal{M},\pi_E}$. The next proposition bounds the error incurred when optimizing the estimate $\widehat{r}^m$ instead of $r^m_{\mathcal{M},\pi_E}$:

Proposition 5.4. *For any MDP\R $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, s'_0, p', \gamma')$, cost $c \in \mathbb{R}^{SA}$ and budget $k \geqslant 0$, it holds that: $|\max_{\pi \in \Pi_{c,k}} V^\pi(s'_0; p', \widehat{r}^m) - \max_{\pi \in \Pi_{c,k}} V^\pi(s'_0; p', r^m_{\mathcal{M},\pi_E})| \leqslant \|\widehat{r}^m - r^m_{\mathcal{M},\pi_E}\|_\infty/(1-\gamma)$.*

6 NUMERICAL SIMULATIONS

This section provides high-level insights into how the proposed IRL algorithms generalize expert behavior.[9] To this end, we conduct simulations in small-scale MDP\Rs with deterministic transitions and assume full knowledge of π_E over its support, allowing us to analyze the *exact* reward centroids (more simulations and details in Appendix G).[10]

Experimental setting. We consider two grid-world environments: $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_0, p, \gamma)$ and $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, s'_0, p', \gamma')$, with a 2D state space $\mathcal{S} = [\![10]\!] \times [\![10]\!]$ and five actions $\mathcal{A} = \{\leftarrow, \rightarrow, \uparrow, \downarrow, \cdot\}$. In $\mathcal{M}$, each directional action moves the agent deterministically to the adjacent cell in the corresponding direction, while $\cdot$ keeps the agent in the current cell. In $\mathcal{M}'$, the transition model p' reverses the direction of the arrow actions (e.g., $\leftarrow$ moves right), while $\cdot$ behaves as in p. Initial states s_0, s'_0 and discount factors γ, γ' vary across simulations. In some simulations, $\mathcal{M}'$ includes "hard" constraints with cost $c \in \mathbb{R}^{SA}$ and budget $k \geqslant 0$ that restrict access to certain states. We visualize the state space $\mathcal{S}$ as a 10×10 grid, marking s_0 or s'_0 in red and constrained states (due to c, k) in brown (see Fig. 2-3). For deterministic policies, we display the action taken in each state (see Fig. 2 (c)); for stochastic ones, each action is shown with size proportional to its selection probability (see Fig. 2 (b)). Policies defined only on subsets of $\mathcal{S}$ (e.g., π_E on $\mathcal{S}_{\mathcal{M},\pi_E}$) leave the remaining cells blank (see Fig. 2 (a)). To indicate that a policy π is executed in $\mathcal{M}$ or $\mathcal{M}'$, we color each cell by the state-only occupancy measure $d_{\mathcal{M},\pi}$ or $d_{\mathcal{M}',\pi}$ (darker blue means higher occupancy).

Simulations and results. For OPT, we used the deterministic policy π_E shown in Fig. 2 (a), which moves right for a few steps before stopping. We found that the policy $\pi^{\text{OPT}}_{\mathcal{M}}$ obtained through our method (i.e., by planning in $\mathcal{M}'$ with $r^{\text{OPT}}_{\mathcal{M},\pi_E}$, see Eq. 15), is quite similar to the policy $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$, defined as the policy whose state-action occupancy measure in $\mathcal{M}'$ is closest to that of the expert $d_{\mathcal{M},\pi_E}$ in $\mathcal{M}$ (details in Appendix G.1). Intuitively, both $\pi^{\text{OPT}}_{\mathcal{M}}$ and $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ aim to "re-play" in $\mathcal{M}'$ the actions demonstrated by the expert in $\mathcal{M}$ (as shown by Fig. 2 (b)-(c) and (d)-(e)), with the difference that $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ weights states in $\mathcal{S}_{\mathcal{M},\pi_E}$ by the expert's discounted visitation frequency $d_{\mathcal{M},\pi_E}$, while $\pi^{\text{OPT}}_{\mathcal{M}}$ treats all visited states equally (see Appendix G.3). For MCE and BIRL, the results are quite similar. Here, we considered as π_E the stochastic policy represented in Fig. 3 (a). As shown by Fig. 3 (b)-(c) and (d)-(e) (and by Fig. 9 (b)-(c) and (d)-(e) in Appendix G.2), both $\pi^{\text{MCE}}_{\mathcal{M}}$ and $\pi^{\text{BIRL}}_{\mathcal{M}}$ resemble $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ in trying to re-play the actions *mostly* played by the expert in its support, with the difference that, as for OPT, they treat the states in $\mathcal{S}_{\mathcal{M},\pi_E}$ uniformly.

7 CONCLUSION

In this paper, we proposed a novel "on-average" approach for generalizing expert behavior and derived three efficient IRL algorithms for estimating the corresponding reward centroids from offline data. We also provided numerical simulations to illustrate the resulting policies' behavior.

Limitations and future directions. Our approach is currently limited to tabular settings. A key direction for future work is to identify structural assumptions that enable scaling to large or continuous spaces. Broader empirical studies could also clarify in which application domains our method generalizes well and when additional structure is needed.

[9] The code for running our simulations can be found at `https://github.com/filippolazzati/IRL-centroids`.

[10] Our algorithms address the vanilla IRL setting, where only the model of behavior of the expert is known. As such, comparing them with approaches that exploit domain knowledge would be unfair. Similarly, testing their average performance across domains would actually test the models (OPT, MCE, BIRL) rather than our algorithms. For these reasons, we present here simulations that provide high-level intuition about their output. We mention also that comparisons with (Ramachandran and Amir, 2007; Michini and How, 2012; Bajgar et al., 2024) under our prior are unfair, as they try to approximate our closed-form centroids.

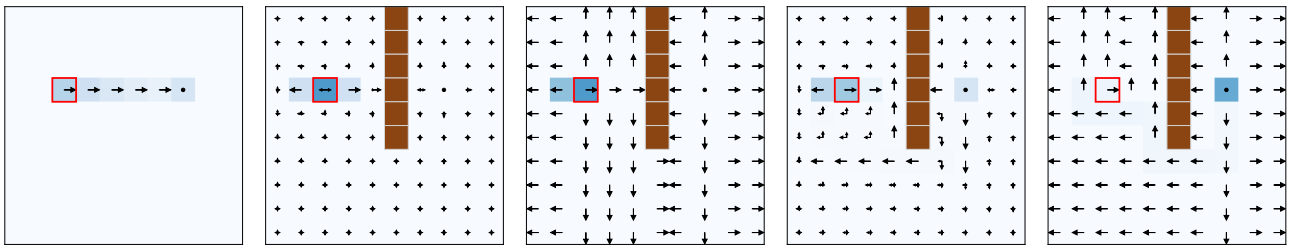

Figure 2: From left to right: (a) deterministic π_E in $\mathcal{M}$ with $\gamma = 0.7$, (b)-(c) $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ and $\pi^{\text{OPT}}_{\mathcal{M}}$ in $\mathcal{M}'$ with $\gamma' = 0.7$, (d)-(e) $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ and $\pi^{\text{OPT}}_{\mathcal{M}}$ in $\mathcal{M}'$ with $\gamma' = 0.95$.

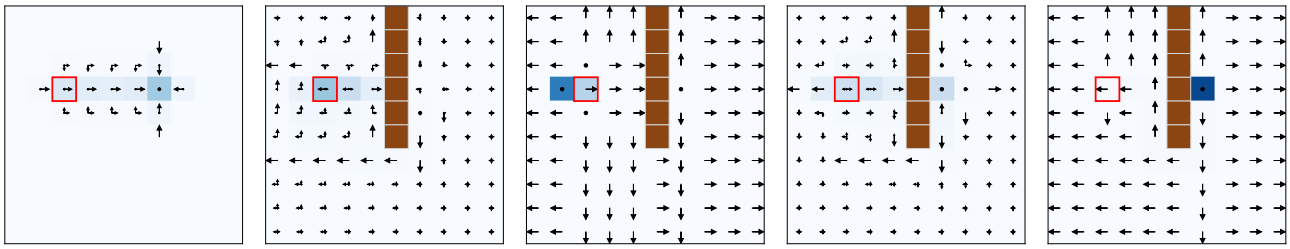

Figure 3: From left to right: (a) stochastic π_E in $\mathcal{M}$ with $\gamma = 0.9$, (b)-(c) $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ and $\pi^{\text{BIRL}}_{\mathcal{M}}$ in $\mathcal{M}'$ with $\gamma' = 0.7$, (d)-(e) $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ and $\pi^{\text{BIRL}}_{\mathcal{M}}$ in $\mathcal{M}'$ with $\gamma' = 0.99$.

Acknowledgements

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

Dean A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *Advances in Neural Information Processing Systems 1 (NeurIPS)*, 1988.

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

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

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

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

Siddharth Reddy, Anca D. Dragan, and Sergey Levine. Sqil: Imitation learning via reinforcement learning with sparse rewards, 2019. URL `https://arxiv.org/abs/1905.11108`.

Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, and Stefano Ermon. Iq-learn: Inverse soft-q learning for imitation. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 4028–4039, 2021.

Markus Kuderer, Shilpa Gulati, and Wolfram Burgard. Learning driving styles for autonomous vehicles from demonstration. In *IEEE International Conference on Robotics and Automation 32 (ICRA)*, pages 2641–2646, 2015.

Chao Yu, Jiming Liu, and Hongyi Zhao. Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. *BMC Medical Informatics and Decision Making*, 19, 2019.

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

Nathan D. Ratliff, J. Andrew Bagnell, and Martin A. Zinkevich. Maximum margin planning. In *International Conference on Machine Learning 23 (ICML)*, pages 729–736, 2006.

David Silver, J. Andrew Bagnell, and Anthony Stentz. Learning from demonstration for autonomous navigation in complex unstructured terrain. *The Inter-*

national Journal of Robotics Research, 29(12):1565–1592, 2010.

Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021.

Andreas Schlaginhaufen and Maryam Kamgarpour. Identifiability and generalizability in constrained inverse reinforcement learning. In *International Conference on Machine Learning 40 (ICML)*, volume 202, pages 30224–30251, 2023.

OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving rubik's cube with a robot hand, 2019.

Jan Matas, Stephen James, and Andrew J. Davison. Sim-to-real reinforcement learning for deformable object manipulation. In *Conference on Robot Learning 2 (CoRL)*, volume 87, pages 734–743, 2018.

Siddharth Desai, Ishan Durugkar, Haresh Karnan, Garrett Warnell, Josiah Hanna, and Peter Stone. An imitation from observation approach to transfer learning with dynamics mismatch. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 3917–3929, 2020.

Tianwei Ni, Harshit Sikchi, Yufei Wang, Tejus Gupta, Lisa Lee, and Ben Eysenbach. f-irl: Inverse reinforcement learning via state marginal matching. In *Conference on Robot Learning (CoRL)*, volume 155, pages 529–551, 2021.

Stuart Russell. Learning agents for uncertain environments (extended abstract). In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory 11 (COLT)*, pages 101–103, 1998.

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

Kareem Amin and Satinder Singh. Towards resolving unidentifiability in inverse reinforcement learning, 2016.

Paul Rolland, Luca Viano, Norman Schürhoff, Boris Nikolov, and Volkan Cevher. Identifiability and generalizability from multiple experts in inverse reinforcement learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 550–564, 2022.

Andreas Schlaginhaufen and Maryam Kamgarpour. Towards the transferability of rewards recovered via regularized inverse reinforcement learning, 2024.

Joar Skalse and Alessandro Abate. Misspecification in inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence 37 (AAAI)*, 2023.

Brian D. Ziebart. Modeling purposeful adaptive behavior with the principle of maximum causal entropy, 2010.

Riccardo Poiani, Gabriele Curti, Alberto Maria Metelli, and Marcello Restelli. Inverse reinforcement learning with sub-optimal experts, 2024.

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

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

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

Alberto Maria Metelli, Filippo Lazzati, and Marcello Restelli. Towards theoretical understanding of inverse reinforcement learning. In *International Conference on Machine Learning 40 (ICML)*, pages 24555–24591, 2023.

Joar Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *International Conference on Machine Learning 40 (ICML)*, volume 202, pages 32033–32058, 2023.

Kuno Kim, Shivam Garg, Kirankumar Shiragur, and Stefano Ermon. Reward identification in inverse reinforcement learning. In *International Conference on Machine Learning 38 (ICML)*, pages 5496–5505, 2021.

Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence 23 (AAAI)*, volume 3, pages 1433–1438, 2008.

Jaedeug Choi and Kee-eung Kim. Map inference for bayesian inverse reinforcement learning. In *Advances in Neural Information Processing Systems 24 (NeurIPS)*, 2011.

Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. Maximum entropy deep inverse reinforcement learning, 2016.

Filippo Lazzati and Alberto Maria Metelli. On the partial identifiability in reward learning: Choosing

the best reward, 2025. URL https://arxiv.org/abs/2501.06376.

Bernard Michini and Jonathan P. How. Improving the efficiency of bayesian inverse reinforcement learning. In *IEEE International Conference on Robotics and Automation 29 (ICRA)*, pages 3651–3656, 2012.

Ondrej Bajgar, Alessandro Abate, Konstantinos Gatsis, and Michael A. Osborne. Walking the values in bayesian inverse reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence 40 (UAI)*, 2024.

Luca Viano, Yu-Ting Huang, Parameswaran Kamalaruban, Adrian Weller, and Volkan Cevher. Robust inverse reinforcement learning under transition dynamics mismatch. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 25917–25931, 2021.

J.M. Lee. *Introduction to Riemannian Manifolds*. Springer International Publishing, 2019.

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

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning 34 (ICML)*, volume 70, pages 1352–1361, 2017.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, 2018.

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *International Conference on Machine Learning 16 (ICML)*, pages 278–287, 1999.

Ruohan Wang, Carlo Ciliberto, Pierluigi Vito Amadori, and Yiannis Demiris. Random expert distillation: Imitation learning via expert policy support estimation. In *International Conference on Machine Learning 36 (ICML)*, volume 97, pages 6536–6544, 2019.

Kianté Brantley, Wen Sun, and Mikael Henaff. Disagreement-regularized imitation learning. In *International Conference on Learning Representations*, 2020. URL https://api.semanticscholar.org/CorpusID:211131570.

Fumihiro Sasaki, Tetsuya Yohira, and Atsuo Kawaguchi. Sample efficient imitation learning for continuous control. In *International Conference on Learning Representations 7 (ICLR)*, 2019.

Filippo Lazzati, Mirco Mutti, and Alberto Maria Metelli. Offline inverse rl: New solution concepts and provably efficient algorithms. In *International Conference on Machine Learning 41 (ICML)*, 2024.

Dennis S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, 2011.

Jim Hefferon. *Linear Algebra*. Department of Mathematics & Applied Mathematics/Virginia Commonwealth University, 2009.

D.L. Cohn. *Measure Theory: Second Edition*. Springer New York, 2013.

William Munger Boothby. *An introduction to differentiable manifolds and Riemannian geometry*. Academic Press, Orlando, FL, 1986.

Michael Stoll. Integration and manifolds, Fall 2007.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 27395–27407, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] See Section 3 for the setting and assumptions, and Sections 4-5 for our algorithms.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] Theoretical analysis in Section 5, empirical in Section 6.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] See attached folder.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes] See the appendix.
   (c) Clear explanations of any assumptions. [Yes] Each theoretical result is preceded or followed by an intuitive explanation.

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

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

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

A Additional Related Work

In this appendix, we discuss additional related work that was omitted from the main paper due to space constraints.

In Section 3.2, we categorized existing IRL methods into three groups, based on the strategy used to address the partial identifiability of the expert's reward function. Here, we introduce a fourth category of methods, which we refer to as "heuristic" methods (Ng and Russell, 2000; Ratliff et al., 2006). These are closely related to "best-case" methods. Specifically, heuristic methods select a reward function from the feasible set using an application-specific heuristic, which is generally not justified beyond the given context. For example, (Ng and Russell, 2000; Ratliff et al., 2006) employ the heuristic of *margin maximization* to select a reward function from the feasible set, a choice mainly justified within the IL setting.

The reward function obtained through OPT (see Eq. 16) can be computed using knowledge of the expert's support $\mathcal{S}_{\mathcal{M},\pi_E}$, making it similar to expert support estimation methods in IL found in the literature (Wang et al., 2019; Brantley et al., 2020). These works present IL algorithms based on reward functions derived directly from $\mathcal{S}_{\mathcal{M},\pi_E}$. However, we emphasize that these approaches are specifically tailored to IL and typically involve implicit structural assumptions to scale to large or continuous state-action spaces (see also (Reddy et al., 2019; Sasaki et al., 2019)).

Finally, we note that the reward derived for MCE in Eq.(17) was previously discussed in Section 4 of (Fu et al., 2017) as a promising approach for IL. However, it was shown to potentially yield errors when applied to novel environments. This observation led (Fu et al., 2017) to introduce additional assumptions aimed at reducing the size of the feasible set to obtain a more transferable reward. In contrast, our work provides a principled justification for this reward based on the minimization of average error in the absence of additional assumptions.

B Additional Notation

In this section, we provide additional notation that will be useful for the presentation of the contents in the appendix. In addition, in Appendix B.1, we show the equivalence between $\mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\text{MCE}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\text{BIRL}}_{\mathcal{M},\pi_E}$ and $\mathcal{R}^{\text{OPT},\mathcal{S}_{\mathcal{M},\pi_E}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\text{MCE},\mathcal{S}_{\mathcal{M},\pi_E}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\text{BIRL},\mathcal{S}_{\mathcal{M},\pi_E}}_{\mathcal{M},\pi_E}$, while in Appendix B.2 we provide useful results that permit to rewrite the feasible sets $\mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\text{MCE}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\text{BIRL}}_{\mathcal{M},\pi_E}$ more explicitly.

Given a policy π and a subset of states $\overline{\mathcal{S}} \subseteq \mathcal{S}$, we define as $[\pi]_{\overline{\mathcal{S}}} := \{\pi' \in \Delta^{\mathcal{A}}_{\mathcal{S}} \mid \forall s \in \overline{\mathcal{S}} : \pi'(\cdot|s) = \pi(\cdot|s) \wedge \forall s \notin \overline{\mathcal{S}}, \exists a \in \mathcal{A} : \pi'(a|s) = 1\}$ the set of policies that in $\overline{\mathcal{S}}$ coincide with π, and outside $\overline{\mathcal{S}}$ are deterministic.

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_0, p, \gamma)$ be an MDP\R, and let $V \in \mathbb{R}^S$ be an arbitrary vector and $A \in \mathbb{R}^{S\times(A-1)}$ an arbitrary matrix. Then, for any deterministic policy π, we define the linear operator $T_{p,\pi} : \mathbb{R}^S \times \mathbb{R}^{S\times(A-1)} \to \mathbb{R}^{S\times A}$ as the matrix that takes in input a pair (V, A) and return a reward $r = T_{p,\pi}(V, A)$ in the following way. For all $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$r(s,a) = V(s) - \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)V(s') + \begin{cases} 0 & \text{if } a = \pi(s), \\ A(s,a) & \text{otherwise.} \end{cases} \tag{21}$$

A property that will be useful in various proofs is that: $|\det(T_{p,\pi})| = |\det(W_{p,\pi})|$ (see Proposition D.4).

Similarly to $T_{p,\pi}$, for any SA-dimensional vector $\eta \in \mathbb{R}^{SA}$, we introduce the linear map $U_{p,\eta} : \mathbb{R}^S \to \mathbb{R}^{SA}$ that takes in input a vector $V \in \mathbb{R}^S$ and returns a reward $r = U_{p,\eta}(V)$ such that, for all $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$r(s,a) = V(s) - \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)V(s') + \eta(s,a). \tag{22}$$

For any MDP\R $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_0, p, \gamma)$, it is useful to define a partition of the set of all possible rewards $\mathbb{R}^{SA} = \mathscr{R}^1_p \cup \mathscr{R}^{>1}_p$ into the set of rewards that induce in $\mathcal{M}$ a unique optimal policy from every state $\mathscr{R}^1_p$, and those that do not $\mathscr{R}^{>1}_p$. Formally:

$$\mathscr{R}^1_p := \Big\{ r \in \mathbb{R}^{SA} \;\Big|\; |\Pi^*(r,p)| = 1 \Big\}, \tag{23}$$

$$\mathscr{R}^{>1}_p := \Big\{ r \in \mathbb{R}^{SA} \;\Big|\; |\Pi^*(r,p)| > 1 \Big\}. \tag{24}$$

B.1 Equivalence among Sets of Rewards

We now show that the feasible sets $\mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\text{MCE}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\text{BIRL}}_{\mathcal{M},\pi_E}$ coincide, respectively, with the sets $\mathcal{R}^{\text{OPT},\mathcal{S}_{\mathcal{M},\pi_E}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\text{MCE},\mathcal{S}_{\mathcal{M},\pi_E}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\text{BIRL},\mathcal{S}_{\mathcal{M},\pi_E}}_{\mathcal{M},\pi_E}$. First, we write $\mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\text{MCE}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\text{BIRL}}_{\mathcal{M},\pi_E}$ formally:

$$\mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E} := \big\{r \in \mathbb{R}^{SA} \,\big|\, \pi_E \in \arg\max_{\pi} V^{\pi}(s_0; p, r)\big\}, \tag{25}$$

$$\mathcal{R}^{\text{MCE}}_{\mathcal{M},\pi_E} := \big\{r \in \mathbb{R}^{SA} \,\big|\, \pi_E = \arg\max_{\pi} V^{\pi}_{\lambda}(s_0; p, r)\big\}, \tag{26}$$

$$\mathcal{R}^{\text{BIRL}}_{\mathcal{M},\pi_E} := \big\{r \in \mathbb{R}^{SA} \,\big|\, \pi_E(a|s) \propto e^{\frac{1}{\beta}Q^*(s,a;p,r)} \ \forall s \in \mathcal{S}^{\mathcal{M},\pi_E}, \forall a \in \mathcal{A}\big\}, \tag{27}$$

from which the equality $\mathcal{R}^{\text{BIRL}}_{\mathcal{M},\pi_E} = \mathcal{R}^{\text{BIRL},\mathcal{S}_{\mathcal{M},\pi_E}}_{\mathcal{M},\pi_E}$ is immediate. Regarding OPT, the equality $\mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E} = \mathcal{R}^{\text{OPT},\mathcal{S}_{\mathcal{M},\pi_E}}_{\mathcal{M},\pi_E}$ follows directly from Lemma E.1 of (Lazzati et al., 2024). Finally, w.r.t. MCE, simply observe that the constraint $\pi_E = \arg\max_{\pi} V^{\pi}_{\lambda}(s_0; p, r)$ holds only for the states $s \in \mathcal{S}_{\mathcal{M},\pi_E}$, i.e., those reachable from s_0 playing π_E in $\mathcal{M}$. Thus, the reward functions $r \in \mathcal{R}^{\text{MCE}}_{\mathcal{M},\pi_E}$ can take on arbitrary values in $s' \notin \mathcal{S}_{\mathcal{M},\pi_E}$. Then, the equality $\mathcal{R}^{\text{MCE}}_{\mathcal{M},\pi_E} = \mathcal{R}^{\text{MCE},\mathcal{S}_{\mathcal{M},\pi_E}}_{\mathcal{M},\pi_E}$ follows by using Eq. (5) of (Cao et al., 2021) (or, equivalently, Eq. (6) of (Haarnoja et al., 2017)).

B.2 Rewriting the Feasible Sets Explicitly

The propositions below permit to rewrite the feasible sets $\mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\text{MCE}}_{\mathcal{M},\pi_E}$, $\mathcal{R}^{\text{BIRL}}_{\mathcal{M},\pi_E}$ more explicitly when $\mathcal{S} = \mathcal{S}_{\mathcal{M},\pi_E}$. Observe that, by making explicit the number of degrees of freedom in these sets, they also allow us to realize that $\mathcal{R}^{\text{OPT},\mathcal{S}}_{\mathcal{M},\pi_E}$ is SA-dimensional, while both $\mathcal{R}^{\text{MCE},\mathcal{S}}_{\mathcal{M},\pi_E}$ and $\mathcal{R}^{\text{BIRL},\mathcal{S}}_{\mathcal{M},\pi_E}$ are S-dimensional.

Proposition B.1. *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_0, p, \gamma)$ be an MDP\R, and let π be any policy. Then, for all and only the rewards r in the set $\mathcal{R}^{OPT,\mathcal{S}}_{\mathcal{M},\pi}$ there exist two vectors $V \in \mathbb{R}^S$ and $A \in (-\infty, 0]^{S\times A}$ such that, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$:*

$$r(s,a) = V(s) - \gamma \sum_{s'} p(s'|s,a)V(s') + \begin{cases} 0 & \text{if } \pi(a|s) > 0, \\ A(s,a) & \text{otherwise.} \end{cases}$$

Moreover, V and A coincide with, respectively, $V^*(\cdot; p, r)$ and $A^*(\cdot, \cdot; p, r)$.

Proof. See Lemma 3.2 of (Metelli et al., 2021). □

Proposition B.2. *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_0, p, \gamma)$ be an MDP\R, and let π be any stochastic policy. Then, for all and only the rewards r in the set $\mathcal{R}^{MCE,\mathcal{S}}_{\mathcal{M},\pi}$ there exists a vector $V \in \mathbb{R}^S$ such that, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$:*

$$r(s,a) = V(s) - \gamma \sum_{s'} p(s'|s,a)V(s') + \lambda \log \pi(a|s).$$

Moreover, $V(s)$ coincides with $V^*_{\lambda}(s; p, r)$, and $\lambda \log \pi(a|s)$ coincides with $A^*_{\lambda}(s, a; p, r)$.

Proof. See Theorem 1 of (Cao et al., 2021). □

Proposition B.3. *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_0, p, \gamma)$ be an MDP\R, and let π be any stochastic policy. Then, for all and only the rewards r in the set $\mathcal{R}^{BIRL,\mathcal{S}}_{\mathcal{M},\pi}$ there exists a vector $V \in \mathbb{R}^S$ such that, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$:*

$$r(s,a) = V(s) - \gamma \sum_{s'} p(s'|s,a)V(s') + \beta \log \frac{\pi(a|s)}{\max_{a' \in \mathcal{A}} \pi(a'|s)}.$$

Moreover, $V(s)$ coincides with $V^*(s; p, r)$, and $\beta \log(\pi(a|s)/\max_{a' \in \mathcal{A}} \pi(a'|s))$ coincides with $A^*(s, a; p, r)$.

Proof. We begin by showing that every reward in the set can be written in this way. Starting from the definition of $\mathcal{R}^{\text{BIRL},\mathcal{S}}_{\mathcal{M},\pi}$, for any $r \in \mathcal{R}^{\text{BIRL},\mathcal{S}}_{\mathcal{M},\pi}$, for every $(s,a) \in \mathcal{S} \times \mathcal{A}$, it holds that:

$$\pi(a|s) = \frac{e^{\frac{1}{\beta}Q^*(s,a;p,r)}}{\sum_{a'} e^{\frac{1}{\beta}Q^*(s,a';p,r)}}$$

$$= \frac{e^{\frac{1}{\beta}(Q^*(s,a;p,r)-V^*(s;p,r))}}{\sum_{a'} e^{\frac{1}{\beta}(Q^*(s,a';p,r)-V^*(s;p,r))}}$$
$$= \frac{e^{\frac{1}{\beta}A^*(s,a;p,r)}}{\sum_{a'} e^{\frac{1}{\beta}A^*(s,a';p,r)}}.$$

Then, by rearranging and taking the logarithm on both parts, we get:

$$\frac{1}{\beta}A^*(s,a;p,r) = \log \pi(a|s) + \log \sum_{a' \in \mathcal{A}} e^{\frac{1}{\beta}A^*(s,a';p,r)}.$$

Now, by applying the following formula:

$$\beta \log \sum_{a' \in \mathcal{A}} e^{\frac{1}{\beta}A^*(s,a';p,r)} = -\beta \log \max_{a' \in \mathcal{A}} \pi(a'|s),$$

which has been derived in the proof of Lemma C.2 of (Skalse et al., 2023), we obtain:

$$A^*(s,a;p,r) = \beta \log \frac{\pi(a|s)}{\max_{a'} \pi(a'|s)}.$$

By using the definition of optimal advantage function: $A^*(s,a;p,r) = Q^*(s,a;p,r) - V^*(s;p,r)$, and also the Bellman optimality equation (Puterman, 1994): $Q^*(s,a;p,r) = r(s,a) + \gamma \sum_{s'} p(s'|s,a)V^*(s';p,r)$, rearranging, and using symbol V for $V^*(\cdot;p,r)$, we get the result:

$$r(s,a) = V(s) - \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a)V(s') + \beta \log \frac{\pi(a|s)}{\max_{a'} \pi(a'|s)}.$$

To show that if a reward function r can be written in the form described in the statement of the proposition (for some $V \in \mathbb{R}^S$) then it belongs to $\mathcal{R}^{\text{BIRL},\mathcal{S}}_{\mathcal{M},\pi}$, observe that, in every $s \in \mathcal{S}$, there is at least one action a for which (observe that $\log(\pi(a|s)/\max_{a'} \pi(a'|s)) = 0$):

$$r(s,a) = V(s) - \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a)V(s').$$

Therefore, through a reasoning analogous to that of Lemma 3.2 of (Metelli et al., 2021), we realize that V coincides with $V^*(\cdot;p,r)$, and, in addition:

$$A^*(s,a;p,r) = \beta \log \frac{\pi(a|s)}{\max_{a'} \pi(a'|s)}.$$

Thus, the proof can be concluded by reversing the passages in the first part of this proof, to show that this implies that:

$$\pi(a|s) = \frac{e^{\frac{1}{\beta}Q^*(s,a;p,r)}}{\sum_{a'} e^{\frac{1}{\beta}Q^*(s,a';p,r)}},$$

and so $r \in \mathcal{R}^{\text{BIRL},\mathcal{S}}_{\mathcal{M},\pi}$. □

C Additional Results and Proofs for Section 3.2

In this appendix, we collect various results and proofs for Section 3.2. Specifically, in Appendix C.1, we show the importance of restricting the feasible set to a bounded set in the "worst-case" approaches. Next, we present the proofs of Propositions E.2-3.2 (Appendix C.2), and we explain how to extend the result for OPT to state-only rewards (Appendix C.3). Finally, in Appendix C.4, we extend these results to the setting where we are given a uniform distribution over transition models.

C.1 On the Importance of $\mathfrak{R}$ in the "Worst-case" Approach

In this appendix, we show that, if we do not restrict the feasible set $\mathcal{R}^m_{\mathcal{M},\pi_E}$ in Eq. (9) to a bounded set, e.g., $\mathcal{R}^m_{\mathcal{M},\pi_E} \cap \mathfrak{R}$, then there are problem instances where any feasible policy $\pi' \in \Pi_{c,k}$ achieves the objective value $\max_{r \in \mathcal{R}^m_{\mathcal{M},\pi_E} \cap \mathfrak{R}} \big(\max_{\pi \in \Pi_{c,k}} V^\pi(s'_0;p',r) - V^{\pi'}(s'_0;p',r)\big) = +\infty$. As a consequence, any policy π' solves the

Figure 4: Problem instance considered in Appendix C.1.

optimization problem in Eq. (9).

We focus on the problem instance exemplified in Fig. 4. Assume that the expert demonstrates policy π_E with $\pi_E(a_1|s_0) = 1$ in the MDP without reward $\mathcal{M}$ depicted on the left of Fig. 4, where s_0 is the initial state and γ is any scalar in$(0, 1)$, and we want to "generalize π_E" to the MDP without reward $\mathcal{M}'$ on the right of Fig. 4, where s_0 is again the initial state and γ' is any value in $[0, 1)$. Assume that there are no constraints (e.g., $c(s, a) = 0$ and $k = 1$) and that the model of behavior is $m = \text{OPT}$. Then, the robust approach of (Lazzati and Metelli, 2025) without restricting the feasible set $\mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E}$ to the hypercube $\mathfrak{R}$ involves the optimization of:

$$\widehat{\pi}_{\text{W}} \in \arg\min_{\pi' \in \Pi} \max_{r \in \mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E}} \Big(V^*(s'_0; p', r) - V^{\pi'}(s'_0; p', r) \Big).$$

Now, we show that, for any policy $\pi \in \Pi = \Delta^3$, there exists a reward function $\overline{r} \in \mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E}$ that makes the objective diverge:

$$\max_{\pi \in \Pi} V^{\pi}(s'_0; p', \overline{r}) - V^{\pi'}(s'_0; p', \overline{r}) \to +\infty.$$

We begin with the deterministic policy π such that $\pi(a_1|s_0) = 1$. Observe that the reward r_1 such that:

$$r_1(s, a) = \begin{cases} 0 & \text{if } (s, a) = (s_0, a_1) \vee (s_0, a_2) \vee (s_0, a_3), \\ +k' & \text{otherwise,} \end{cases}$$

for any $k' \geqslant 0$ belongs to the feasible set $\mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E}$ (because it makes all policies to be optimal), and provides a suboptimality value for π in $\mathcal{M}'$ of:

$$V^*(s'_0; p', r_1) - V^{\pi}(s'_0; p', r_1) = 0 + \gamma \frac{k'}{1-\gamma} - \frac{0}{1-\gamma} = \frac{\gamma}{1-\gamma} k' \to +\infty,$$

for $k' \to \infty$. If π plays deterministically a_2, i.e., $\pi(a_2|s_0) = 1$, then we find reward r_2 such that:

$$r_2(s, a) = \begin{cases} +k' & \text{if } (s, a) = (s_0, a_1) \vee (s_1, a), \\ 0 & \text{otherwise,} \end{cases}$$

for any $k' \geqslant 0$ belongs to the feasible set $\mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E}$, and provides a suboptimality value for π in $\mathcal{M}'$ of:

$$V^*(s'_0; p', r_2) - V^{\pi}(s'_0; p', r_2) = \frac{k'}{1-\gamma} - \frac{0}{1-\gamma} = \frac{k'}{1-\gamma} \to +\infty,$$

for $k' \to \infty$. A similar derivation can be carried out also for the policy that plays a_3 deterministically. Note that, using r_2, we can show the same result also for any stochastic policy. Let π be any stochastic policy in Δ^3. Then, its suboptimality w.r.t. r_2 is:

$$V^*(s'_0; p', r_2) - V^{\pi}(s'_0; p', r_2) = \frac{k'}{1-\gamma} - \pi(a_1|s_0) \frac{k'}{1-\gamma} = \frac{1 - \pi(a_1|s_0)}{1-\gamma} k' \to +\infty,$$

for $k' \to \infty$, as $\pi(a_1|s_0) < 1$ since π is stochastic.

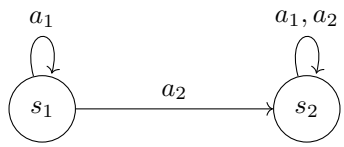

Figure 5: MDP\R used in the proof of Proposition 3.1.

C.2 The Hypercube is *Biased* in the Space of Policies

Proposition 3.1. *There exists an MDP\R $\mathcal{M}$ and two deterministic policies $\pi_1, \pi_2 \in \Pi$ such that:*

$$vol\big(\mathfrak{R} \cap \mathcal{R}^{OPT,\mathcal{S}}_{\mathcal{M},\pi_1}\big) \neq vol\big(\mathfrak{R} \cap \mathcal{R}^{OPT,\mathcal{S}}_{\mathcal{M},\pi_2}\big).$$

Proof. Consider the MDP\R $\mathcal{M}$ depicted in Fig. 5, and let π_1 be the policy that plays action a_1 in both states s_1, s_2. Note that there are four deterministic policies in this MDP\R, thus, we can prove the result by showing that:

$$\frac{\mathrm{vol}\big(\mathfrak{R} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M},\pi_1}\big)}{\mathrm{vol}\big(\mathfrak{R}\big)} \neq \frac{1}{4}.$$

Since $\mathfrak{R} = [-1,+1]^{SA} = [-1,+1]^4$, then $\mathrm{vol}\big(\mathfrak{R}\big) = 2^4 = 16$, thus the result follows by showing that:

$$\mathrm{vol}\big(\mathfrak{R} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M},\pi_1}\big) \neq 4.$$

Let us now compute the integration region $\mathfrak{R} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M},\pi_1}$. For any $r \in \mathbb{R}^{SA}$, let us abbreviate $r_{ij} := r(s_i, a_j)$. We can compute the value and Q-functions of policy π_1 under any r as:

$$\begin{aligned}
V^{\pi_1}(s_1; p, r) &= \frac{r_{11}}{1-\gamma}, \\
V^{\pi_1}(s_2; p, r) &= \frac{r_{21}}{1-\gamma}, \\
Q^{\pi_1}(s_1, a_2; p, r) &= r_{12} + \gamma V^{\pi_1}(s_2; p, r), \\
Q^{\pi_1}(s_2, a_2; p, r) &= r_{22} + \gamma V^{\pi_1}(s_2; p, r).
\end{aligned}$$

Using the policy improvement theorem (Puterman, 1994), we have that a reward r belongs to $\mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M},\pi_1}$ if and only if:

$$\begin{cases} Q^{\pi_1}(s_1, a_2; p, r) \leqslant V^{\pi_1}(s_1; p, r), \\ Q^{\pi_1}(s_2, a_2; p, r) \leqslant V^{\pi_1}(s_2; p, r), \end{cases} \iff \begin{cases} r_{12} + \gamma \frac{r_{21}}{1-\gamma} \leqslant \frac{r_{11}}{1-\gamma}, \\ r_{22} + \gamma \frac{r_{21}}{1-\gamma} \leqslant \frac{r_{21}}{1-\gamma}, \end{cases}$$

which can be rearranged to:

$$\begin{cases} r_{11} \geqslant (1-\gamma) r_{12} + \gamma r_{21}, \\ r_{21} \geqslant r_{22}. \end{cases}$$

To simplify the calculations, let $\gamma \to 1$, so that the integration region reduces to:

$$\mathfrak{R} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M},\pi_1} = \Big\{ r \in [-1,+1]^4 \,|\, r_{11} \geqslant r_{21} \geqslant r_{22} \Big\}.$$

Thus, we can write:

$$\begin{aligned}
\int_{\mathfrak{R} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M},\pi_1}} dr_{11} dr_{12} dr_{21} dr_{22} &= \int_{-1}^{+1} dr_{12} \int_{-1}^{+1} \int_{r_{22}}^{+1} \int_{r_{21}}^{+1} dr_{11} dr_{21} dr_{22} \\
&= 2 \cdot \int_{-1}^{+1} \int_{r_{22}}^{+1} \int_{r_{21}}^{+1} dr_{11} dr_{21} dr_{22} \\
&= 2 \cdot \int_{-1}^{+1} \int_{r_{22}}^{+1} (1 - r_{21}) dr_{21} dr_{22} \\
&= 2 \cdot \int_{-1}^{+1} \frac{1}{2} (1 - r_{22})^2 dr_{22}
\end{aligned}$$

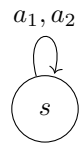

Figure 6: MDP\R used in the proof of Proposition 3.2.

$$= 2 \cdot \frac{4}{3} = \frac{8}{3}.$$

Since $\frac{8}{3} \neq 4$, this concludes the proof. □

Proposition 3.2. *There exists an MDP\R $\mathcal{M}$ and two stochastic policies $\pi_1, \pi_2 \in \Pi$ such that, for all $m \in$ {MCE,BIRL}:*

$$vol\big(\mathfrak{R} \cap \mathcal{R}^{m,\mathcal{S}}_{\mathcal{M},\pi_1}\big) \neq vol\big(\mathfrak{R} \cap \mathcal{R}^{m,\mathcal{S}}_{\mathcal{M},\pi_2}\big).$$

Proof. We begin with MCE. Take $\lambda = 1$ for simplicity. Consider the MDP\R with one state and two actions reported in Fig. 6. By applying Proposition B.2, we have that, for any policy π and for any reward $r \in \mathcal{R}^{\text{MCE},\mathcal{S}}_{\mathcal{M},\pi}$, there exists a scalar V such that:

$$\begin{aligned} r(s, a_1) &= (1-\gamma)V + \log \pi(a_1|s), \\ r(s, a_2) &= (1-\gamma)V + \log \pi(a_2|s). \end{aligned}$$

By rearranging the second equation, we have:

$$V = \frac{r(s, a_2) - \log \pi(a_2|s)}{1-\gamma},$$

and replacing this value of V into the first one, we get:

$$r(s, a_1) = r(s, a_2) + \log \frac{\pi(a_1|s)}{1-\pi(a_1|s)}.$$

Therefore, by denoting $x := r(s, a_2)$, we realize that the integration region $\mathfrak{R} \cap \mathcal{R}^{\text{MCE},\mathcal{S}}_{\mathcal{M},\pi}$ is:

$$\mathfrak{R} \cap \mathcal{R}^{\text{MCE},\mathcal{S}}_{\mathcal{M},\pi} = \Big\{ x \in [-1, +1] \mid -1 - \log \frac{\pi(a_1|s)}{1-\pi(a_1|s)} \leqslant x \leqslant +1 - \log \frac{\pi(a_1|s)}{1-\pi(a_1|s)} \Big\}.$$

Clearly, this is a 1-dimensional region embedded in $\mathbb{R}^2$, whose 1-dimensional volume is:

$$\begin{aligned} \text{vol}(\mathfrak{R} \cap \mathcal{R}^{\text{MCE},\mathcal{S}}_{\mathcal{M},\pi}) &= \int_{\max\{-1, -1-\log \frac{\pi(a_1|s)}{1-\pi(a_1|s)}\}}^{\min\{+1, +1-\log \frac{\pi(a_1|s)}{1-\pi(a_1|s)}\}} dx \\ &= \min\Big\{+1, +1 - \log \frac{\pi(a_1|s)}{1-\pi(a_1|s)}\Big\} - \max\Big\{-1, -1 - \log \frac{\pi(a_1|s)}{1-\pi(a_1|s)}\Big\}. \end{aligned}$$

If we take policies π_1 and π_2 such that $\pi_1(a_1|s) = 1/2$ and $\pi_2(a_1|s) = 1/3$, we obtain the volumes:

$$\begin{aligned} &\text{vol}(\mathfrak{R} \cap \mathcal{R}^{\text{MCE},\mathcal{S}}_{\mathcal{M},\pi_1}) = 2, \\ &\text{vol}(\mathfrak{R} \cap \mathcal{R}^{\text{MCE},\mathcal{S}}_{\mathcal{M},\pi_2}) = 1 - (-1 + \log 2) = 2 - \log 2, \end{aligned}$$

which are different.

Regarding BIRL, the proof is analogous to that for MCE. Specifically, set $\beta = 1$ and consider the same MDP\R in Fig. 6. Thanks to Proposition B.3, we see that, for any policy π and for any reward $r \in \mathcal{R}^{\text{BIRL},\mathcal{S}}_{\mathcal{M},\pi}$, there exists a scalar V such that (take $\pi(a_2|s) \geqslant \pi(a_1|s)$ w.l.o.g.):

$$\begin{aligned} r(s, a_1) &= (1-\gamma)V + \log \frac{\pi(a_1|s)}{1-\pi(a_1|s)}, \\ r(s, a_2) &= (1-\gamma)V. \end{aligned}$$

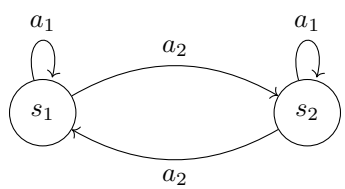

Figure 7: MDP\R used in the proof of Proposition C.1.

By isolating V in the second expression, and replacing in the first one, we get:

$$r(s, a_1) = r(s, a_2) + \log \frac{\pi(a_1|s)}{1 - \pi(a_1|s)},$$

as in MCE. Therefore, for this simple example, the integration regions $\mathfrak{R} \cap \mathcal{R}^{\text{MCE},\mathcal{S}}_{\mathcal{M},\pi} = \mathfrak{R} \cap \mathcal{R}^{\text{BIRL},\mathcal{S}}_{\mathcal{M},\pi}$ coincide. Thus, the same choices of π_1, π_2 as in MCE allow to prove the result. □

C.3 Uniform Prior for State-only Rewards

We present here an analogous result to Proposition 3.1 for state-only rewards.

Proposition C.1. *Let $R_{\max} > 0$ be arbitrary, and consider state-only rewards. There exists an MDP\R $\mathcal{M}$ and two deterministic policies π_1, π_2 such that:*

$$vol\big([-1, +1]^S \cap \mathcal{R}^{OPT,\mathcal{S}}_{\mathcal{M},\pi_1}\big) \neq vol\big([-1, +1]^S \cap \mathcal{R}^{OPT,\mathcal{S}}_{\mathcal{M},\pi_2}\big).$$

Proof. The proof is analogous to that of Proposition 3.1, but we make use of the MDP\R in Fig. 7. Let π_1 be the policy that plays action a_2 in both states s_1, s_2. Note that there are four deterministic policies in this MDP\R, thus, we can prove the result by showing that:

$$\frac{\text{vol}\big([-1, +1]^S \cap \mathcal{R}^{\text{OPT},\mathcal{S}}_{\mathcal{M},\pi_1}\big)}{\text{vol}\big([-1, +1]^S\big)} \neq \frac{1}{4}.$$

Since $[-1, +1]^S = [-1, +1]^2$, then $\text{vol}\big([-1, +1]^S\big) = 2^2 = 4$, thus the result follows by showing that:

$$\text{vol}\big([-1, +1]^S \cap \mathcal{R}^{\text{OPT},\mathcal{S}}_{\mathcal{M},\pi_1}\big) \neq 1.$$

Let us now compute the integration region $[-1, +1]^S \cap \mathcal{R}^{\text{OPT},\mathcal{S}}_{\mathcal{M},\pi_1}$. For any $r \in \mathbb{R}^{SA}$, let us abbreviate $r_i := r(s_i)$ (the reward is state-only). We can compute the value and Q-functions of policy π_1 under any state-only reward r as:

$$\begin{aligned}
V^{\pi_1}(s_1; p, r) &= r_1 + \gamma V^{\pi_1}(s_2; p, r),\\
V^{\pi_1}(s_2; p, r) &= r_2 + \gamma V^{\pi_1}(s_1; p, r),\\
Q^{\pi_1}(s_1, a_1; p, r) &= r_1 + \gamma V^{\pi_1}(s_1; p, r),\\
Q^{\pi_1}(s_2, a_1; p, r) &= r_2 + \gamma V^{\pi_1}(s_2; p, r),
\end{aligned}$$

thus, using the policy improvement theorem (Puterman, 1994), r belongs to $\mathcal{R}^{\text{OPT},\mathcal{S}}_{\mathcal{M},\pi_1}$ if and only if:

$$\begin{cases} Q^{\pi_1}(s_1, a_1; p, r) \leqslant V^{\pi_1}(s_1; p, r), \\ Q^{\pi_1}(s_2, a_1; p, r) \leqslant V^{\pi_1}(s_2; p, r), \end{cases} \iff \begin{cases} r_1 + \gamma V^{\pi_1}(s_1; p, r) \leqslant r_1 + \gamma V^{\pi_1}(s_2; p, r), \\ r_2 + \gamma V^{\pi_1}(s_2; p, r) \leqslant r_2 + \gamma V^{\pi_1}(s_1; p, r), \end{cases}$$

which can be simplified to:

$$\begin{cases} V^{\pi_1}(s_1; p, r) \leqslant V^{\pi_1}(s_2; p, r), \\ V^{\pi_1}(s_2; p, r) \leqslant V^{\pi_1}(s_1; p, r). \end{cases} \iff V^{\pi_1}(s_1; p, r) = V^{\pi_1}(s_2; p, r).$$

Since the value functions can be computed as:

$$V^{\pi_1}(s_1; p, r) = \frac{r_1 + \gamma r_2}{1 - \gamma^2},$$

$$V^{\pi_1}(s_2; p, r) = \frac{r_2 + \gamma r_1}{1 - \gamma^2},$$

the constraint imposes that:

$$r_1 + \gamma r_2 = r_2 + \gamma r_1 \qquad \Longleftrightarrow \qquad r_1 = r_2.$$

Therefore, the integration region:

$$[-1, +1]^S \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M},\pi_1} = \Big\{ r \in [-1, +1]^S \,|\, r_1 = r_2 \Big\},$$

has zero volume, which concludes the proof. □

C.4 A Uniform Prior over Transition Models Too

In this appendix we consider the setting in which the MDP\R $\mathcal{M}$ is not fixed, but its transition model p is sampled from a uniform distribution over $\Delta^{\mathcal{S}}_{\mathcal{S}\times\mathcal{A}}$. Interestingly, it turns out that the uniform prior over the hypercube (see Eq. 11) assigns equal weight to all the deterministic policies on the *average* of the transition models. However, for MCE and BIRL, this does not hold.

Proposition C.2. *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_0, p, \gamma)$ be an MDP\R whose transition model p is a random variable uniformly distributed over $\Delta^{\mathcal{S}}_{\mathcal{S}\times\mathcal{A}}$. Then, for any pair of deterministic policies π_1, π_2, it holds that:*

$$\mathbb{E}_p[vol\big(\mathfrak{R} \cap \mathcal{R}^{OPT,\mathcal{S}}_{\mathcal{M},\pi_1}\big)] = \mathbb{E}_p[vol\big(\mathfrak{R} \cap \mathcal{R}^{OPT,\mathcal{S}}_{\mathcal{M},\pi_2}\big)],$$

where $\mathbb{E}_p$ denotes the expectation w.r.t. p.

Proof. The result can be proved by showing that, for any possible pair of policies π_1, π_2 and any transition model p', there always exists another transition model p'' for which:

$$\mathrm{vol}\big(\mathfrak{R} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{p',\pi_1}\big) = \mathrm{vol}\big(\mathfrak{R} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{p'',\pi_2}\big),$$

and:

$$\mathrm{vol}\big(\mathfrak{R} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{p'',\pi_1}\big) = \mathrm{vol}\big(\mathfrak{R} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{p',\pi_2}\big).$$

where we used p', p'' in place of $(\mathcal{S}, \mathcal{A}, s_0, p', \gamma)$ and $(\mathcal{S}, \mathcal{A}, s_0, p'', \gamma)$. To this aim, fix two policies π_1, π_2 and a transition model p', and construct the transition model p'' such that, for every $s, s' \in \mathcal{S}$:

$$\begin{aligned}
&p''(s'|s, \pi_1(s)) = p'(s'|s, \pi_2(s)),\\
&p''(s'|s, \pi_2(s)) = p'(s'|s, \pi_1(s)),\\
&p''(s'|s, a) = p'(s'|s, a) \qquad \text{if } a \neq \pi_1(s) \wedge \pi_2(s).
\end{aligned}$$

Clearly, p'' guarantees that the role of π_2 is the same as that of π_1 in p', and that the role of π_1 is the same as that of π_2 in p'. Therefore:

$$\mathrm{vol}\big(\mathfrak{R} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{p',\pi_1}\big) = \mathrm{vol}\big(\mathfrak{R} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{p'',\pi_2}\big),$$

and:

$$\mathrm{vol}\big(\mathfrak{R} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{p'',\pi_1}\big) = \mathrm{vol}\big(\mathfrak{R} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{p',\pi_2}\big).$$

This concludes the proof. □

Proposition C.3. *There exist a state space $\mathcal{S}$, action space $\mathcal{A}$, discount factor γ, initial state s_0, coefficients $\lambda, \beta \geqslant 0$, and a pair of stochastic policies π_1, π_2 such that, if we construct $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_0, p, \gamma)$ as an MDP\R whose transition model p is a random variable uniformly distributed over $\Delta^{\mathcal{S}}_{\mathcal{S}\times\mathcal{A}}$, then:*

$$\begin{aligned}
&\mathbb{E}_p[vol\big(\mathfrak{R} \cap \mathcal{R}^{MCE,\mathcal{S}}_{\mathcal{M},\pi_1}\big)] \neq \mathbb{E}_p[vol\big(\mathfrak{R} \cap \mathcal{R}^{MCE,\mathcal{S}}_{\mathcal{M},\pi_2}\big)] \quad \wedge\\
&\mathbb{E}_p[vol\big(\mathfrak{R} \cap \mathcal{R}^{BIRL,\mathcal{S}}_{\mathcal{M},\pi_1}\big)] \neq \mathbb{E}_p[vol\big(\mathfrak{R} \cap \mathcal{R}^{BIRL,\mathcal{S}}_{\mathcal{M},\pi_2}\big)],
\end{aligned}$$

where $\mathbb{E}_p$ denotes the expectation w.r.t. p.

Proof. Simply, observe that, for $S = 1$, the expectation over transition models reduces to evaluate the single transition model that brings every action back to the current state. For this problem instance, we have already

shown that there exists a pair of policies π_1, π_2 for which the volumes are different (see the proof of Proposition 3.2). □

D Additional Results and Proofs for Section 4.1

In this appendix, we provide the missing proofs for Proposition 4.1 (Appendix D.1) and for Propositions 4.2-4.4 (Appendix D.2). In addition, we show that priors $w^n_{\mathcal{M}}$ might be unbiased in new environments $\mathcal{M}''$ (Appendix D.3).

D.1 Relation with the Hypercube

To prove Proposition 4.1, we use the following three lemmas.

Proposition D.1. *Let $\mathcal{M}$ be any MDP\R. For any $C_1^{OPT}, C_2^{OPT} > 0$, it holds that:*

$$\mathscr{R}^{OPT}_{\mathcal{M}} \supseteq \Big[-(1-\gamma)\min\Big\{\frac{C_1^{OPT}}{1+\gamma}, C_2^{OPT}\Big\}, +(1-\gamma)\min\Big\{\frac{C_1^{OPT}}{1+\gamma}, C_2^{OPT}\Big\}\Big]^{SA},$$

and that:

$$\mathscr{R}^{OPT}_{\mathcal{M}} \subseteq \Big[-\frac{1+\gamma}{1-\gamma}C_1^{OPT} - C_2^{OPT}, +\frac{1+\gamma}{1-\gamma}C_1^{OPT}\Big]^{SA}.$$

Proof. For any (stochastic) policy $\pi \in \Pi$, define the set:

$$\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M},\pi} := \Big\{r \in \mathbb{R}^{SA} \;\Big|\; \pi \in \Pi^*(p,r) \wedge \forall s,a: \; |V^\pi(s;p,r)| \leqslant C_1^{\mathrm{OPT}} k_\pi \wedge |A^\pi(s,a;p,r)| \leqslant C_2^{\mathrm{OPT}}\Big\},$$

as the set of all the rewards that make policy π optimal at every possible state $s \in \mathcal{S}$ and whose value and advantage functions are bounded respectively by $C_1^{\mathrm{OPT}} k_\pi$ and C_2^{OPT}.

We show that, for any MDP\R $\mathcal{M}$ and policy $\pi \in \Pi$, it holds that:

$$\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M},\pi} \supseteq [-(1-\gamma)\min\{C_1^{\mathrm{OPT}} k_\pi, C_2^{\mathrm{OPT}}\}, +(1-\gamma)\min\{C_1^{\mathrm{OPT}} k_\pi, C_2^{\mathrm{OPT}}\}]^{SA}. \tag{28}$$

To this aim, observe that, for any $r \in [-r_{\max}, +r_{\max}]^{SA}$ (with $r_{\max} > 0$), the value and advantage functions $V^\pi(s;p,r)$ and $A^\pi(s,a;p,r)$ induced by policy π in $\mathcal{M}_r$ are bounded by $\frac{r_{\max}}{1-\gamma}$ (Puterman, 1994). Therefore, if we set $r_{\max} = (1-\gamma)\min\{C_1^{\mathrm{OPT}} k_\pi, C_2^{\mathrm{OPT}}\}$, we have that $|V^\pi(s;p,r)| \leqslant C_1^{\mathrm{OPT}} k_\pi$ and $|A^\pi(s,a;p,r)| \leqslant C_2^{\mathrm{OPT}}$, and therefore Eq. (28) holds.

Now, let us show that, for any MDP\R $\mathcal{M}$ and policy $\pi \in \Pi$, it holds that:

$$\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M},\pi} \subseteq [-(1+\gamma)C_1^{\mathrm{OPT}} k_\pi - C_2^{\mathrm{OPT}}, +(1+\gamma)C_1^{\mathrm{OPT}} k_\pi]^{SA}. \tag{29}$$

This can be done by first observing that any $r \in \mathscr{R}^{\mathrm{OPT}}_{\mathcal{M},\pi}$, since $\pi \in \Pi^*(p,r)$ by definition, can be rewritten through Proposition B.1 as:

$$r(s,a) = V(s) - \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a)V(s') - \begin{cases} 0 & \text{if } \pi(a|s) > 0, \\ A(s,a) & \text{otherwise,} \end{cases}$$

where $V \in \mathbb{R}^S$ and $A \in (-\infty, 0]^{S \times A}$ coincide with $V^*(\cdot;p,r)$ and $A^*(\cdot,\cdot;p,r)$ (see Proposition B.1). However, by definition of $\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M},\pi}$ we know that $|V^\pi(s;p,r)| \leqslant C_1^{\mathrm{OPT}} k_\pi$ and $|A^\pi(s,a;p,r)| \leqslant C_2^{\mathrm{OPT}}$, therefore, by applying triangle's inequality, we have that:

$$\begin{aligned} \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |r(s,a)| &\leqslant |V(s)| + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a)|V(s')| + \begin{cases} 0 & \text{if } \pi(a|s) > 0, \\ |A(s,a)| & \text{otherwise,} \end{cases} \\ &\leqslant (1+\gamma)C_1^{\mathrm{OPT}} k_\pi + C_2^{\mathrm{OPT}}. \end{aligned}$$

Thus, Eq. (29) holds.

To conclude the proof, given the relationship between $\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}}$ and all the sets $\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M},\pi}$, it suffices to show that, for

any MDP\R $\mathcal{M}$ and policy $\pi \in \Pi$, it holds that:

$$k_\pi \in \Big[\frac{1}{1+\gamma}, \frac{1}{1-\gamma}\Big].$$

To do so, we write:

$$\begin{aligned}(k_\pi)^{-S} &\overset{(1)}{=} |\det(W_{p,\pi})| \\ &\overset{(2)}{=} |\det(I - \gamma P^\pi)| \\ &\overset{(3)}{=} \Big|\prod_{i=1}^{S}(1-\gamma\lambda_i)\Big|,\end{aligned}$$

where at (1) we use that the determinant of the inverse of a matrix coincides with the inverse of the determinant of the matrix (Bernstein, 2011), at (2) we use the definition of $W_{p,\pi}$, where $I \in \mathbb{R}^{S\times S}$ is the identity matrix, and $P^\pi \in \mathbb{R}^{S\times S}$ associates, to every pair of states $s, s' \in \mathcal{S}$, the value $\sum_{a\in\mathcal{A}} \pi(a|s)p(s'|s,a)$, at (3) we use the fact that the characteristic polynomial satisfies $\det(xI - P^\pi) = \prod_{i=1}^{S}(x-\lambda_i)$ for any $x \in \mathbb{R}$, where λ_i are the eigenvalues of P^π, and then we replaced $x = 1/\gamma$.

By noting that P^π is a right stochastic matrix, we have that its eigenvalues satisfy $|\lambda_i| \leqslant 1$ for all $i \in [\![S]\!]$, and so:

$$\Big(\frac{1}{k_\pi}\Big)^S \in \Big[(1-\gamma)^S, (1+\gamma)^S\Big] \iff k_\pi \in \Big[\frac{1}{1+\gamma}, \frac{1}{1-\gamma}\Big].$$

This concludes the proof. □

Proposition D.2. *Let $\mathcal{M}$ be any MDP\R. For any $C_1^{MCE}, C_2^{MCE} > 0$, it holds that:*

$$\mathscr{R}_{\mathcal{M}}^{MCE} \subseteq [-(1+\gamma)C_1^{MCE} - C_2^{MCE}, +(1+\gamma)C_1^{MCE}]^{SA}.$$

Moreover, define $k_1 := (1-\gamma)C_1^{MCE} - \lambda \log A$ and $k_2 := (1-\gamma)C_2^{MCE}/2 - \lambda/2 \log A$. Then, if $\min\{k_1, k_2\} > 0$, we also have that:

$$\mathscr{R}_{\mathcal{M}}^{MCE} \supseteq [-\min\{k_1,k_2\}, +\min\{k_1,k_2\}]^{SA}.$$

Proof. For the first part of the statement, for any $r \in \mathscr{R}_{\mathcal{M}}^{\mathrm{MCE}}$ we can apply Proposition B.2 to rewrite it as:

$$\begin{aligned}r(s,a) &= V(s) - \gamma \sum_{s'} p(s'|s,a)V(s') + \lambda \log \pi(a|s) \\ &= V_\lambda^*(s;p,r) - \gamma \sum_{s'} p(s'|s,a)V_\lambda^*(s';p,r) + A_\lambda^*(s,a;p,r).\end{aligned}$$

The result follows by applying triangle's inequality and noting that $V_\lambda^*(s;p,r)$ and $A_\lambda^*(s,a;p,r)$ are bounded since $r \in \mathscr{R}_{\mathcal{M}}^{\mathrm{MCE}}$.

For the second part of the proposition, let $r \in [-r_{\max}, +r_{\max}]^{SA}$ (for some $r_{\max} > 0$). Then, we can show that, for every s, a:

$$\begin{aligned}|V_\lambda^*(s;p,r)| &\leqslant \frac{r_{\max} + \lambda \log A}{1-\gamma}, \\ |A_\lambda^*(s,a;p,r)| &\leqslant \frac{2r_{\max} + \lambda \log A}{1-\gamma}.\end{aligned}$$

To do so, for any $s \in \mathcal{S}$, we can write (see also (Cao et al., 2021) and (Haarnoja et al., 2017)):

$$\begin{aligned}V_\lambda^*(s;p,r) &= \lambda \log \sum_{a\in\mathcal{A}} e^{\frac{1}{\lambda}Q_\lambda^*(s,a;p,r)} \\ &\overset{(1)}{\leqslant} \max_{a\in\mathcal{A}} Q_\lambda^*(s,a;p,r) + \lambda \log A \\ &= \max_{a\in\mathcal{A}} r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)V_\lambda^*(s';p,r) + \lambda \log A\end{aligned}$$

$$
\begin{aligned}
&\overset{(2)}{\leqslant} r_{\max} + \gamma \max_{s' \in \mathcal{S}} V^*_\lambda(s'; p, r) + \lambda \log A \\
&\overset{(3)}{\leqslant} r_{\max} + \lambda \log A + \gamma (r_{\max} + \lambda \log A + \gamma \max_{s' \in \mathcal{S}} V^*_\lambda(s'; p, r)) \\
&\overset{(4)}{\leqslant} \frac{r_{\max} + \lambda \log A}{1-\gamma},
\end{aligned}
$$

where at (1) we used the log-sum-exp inequality: $\log \sum_{i \in [\![n]\!]} e^{x_i} \leqslant \max_i x_i + \log n$, at (2) we use that r is in the hypercube by hypothesis, at (3) we unfold the recursion, obtaining the expression in (4) since it is a geometric series.

Regarding the advantage function, we begin by lower bounding the soft Q^*_λ:

$$
\begin{aligned}
Q^*_\lambda(s, a; p, r) &= r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) V^*_\lambda(s'; p, r) \\
&\overset{(5)}{\geqslant} -r_{\max} + \gamma \min_{s' \in \mathcal{S}} V^*_\lambda(s'; p, r) \\
&= -r_{\max} + \gamma \min_{s' \in \mathcal{S}} \lambda \log \sum_{a' \in \mathcal{A}} e^{\frac{1}{\lambda} Q^*_\lambda(s', a'; p, r)} \\
&\overset{(6)}{\geqslant} -r_{\max} + \gamma \min_{s' \in \mathcal{S}} \max_{a' \in \mathcal{A}} Q^*_\lambda(s', a'; p, r) \\
&\geqslant -r_{\max} + \gamma (-r_{\max} + \gamma \min_{s' \in \mathcal{S}} \max_{a' \in \mathcal{A}} Q^*_\lambda(s', a'; p, r)) \\
&\overset{(7)}{=} -\frac{r_{\max}}{1-\gamma},
\end{aligned}
$$

where at (5) we use that r is in the hypercube and bound the expectation with the minimum, at (6) we used the log-sum-exp inequality: $\log \sum_i e^{x_i} \geqslant \max_i x_i$, and then we have unfolded the recursion, obtaining the expression in (7).

Then, we can bound the soft advantage combining these bounds:

$$
\begin{aligned}
A^*_\lambda(s, a; p, r) &\coloneqq Q^*_\lambda(s, a; p, r) - V^*_\lambda(s; p, r) \\
&\geqslant -\frac{r_{\max}}{1-\gamma} - \frac{r_{\max} + \lambda \log A}{1-\gamma} \\
&= -\frac{2 r_{\max} + \lambda \log A}{1-\gamma}.
\end{aligned}
$$

Therefore, a sufficient condition for the reward r to belong to set $\mathscr{R}^{\mathrm{MCE}}_{\mathcal{M}}$ is that:

$$
\begin{cases}
\frac{r_{\max} + \lambda \log A}{1-\gamma} \leqslant C^{\mathrm{MCE}}_1, \\
\frac{2 r_{\max} + \lambda \log A}{1-\gamma} \leqslant C^{\mathrm{MCE}}_2,
\end{cases}
\iff
\begin{cases}
r_{\max} \leqslant (1-\gamma) C^{\mathrm{MCE}}_1 - \lambda \log A, \\
r_{\max} \leqslant (1-\gamma) C^{\mathrm{MCE}}_2 / 2 - \lambda / 2 \log A.
\end{cases}
$$

By taking the minimum and imposing that $r_{\max} \geqslant 0$, we get the result. □

Proposition D.3. *Let $\mathcal{M}$ be any MDP\R. For any $C^{BIRL}_1, C^{BIRL}_2 > 0$, it holds that:*

$$
\mathscr{R}^{BIRL}_{\mathcal{M}} \subseteq [-(1+\gamma) C^{BIRL}_1 - C^{BIRL}_2, +(1+\gamma) C^{BIRL}_1]^{SA},
$$

and also that:

$$
\mathscr{R}^{BIRL}_{\mathcal{M}} \supseteq [-(1-\gamma) \min\{C^{BIRL}_1, C^{BIRL}_2\}, +(1-\gamma) \min\{C^{BIRL}_1, C^{BIRL}_2\}]^{SA}.
$$

Proof. For the first part, observe that any $r \in \mathscr{R}^{\mathrm{BIRL}}_{\mathcal{M}}$ can be rewritten using Proposition B.3 as:

$$
\begin{aligned}
r(s, a) &= V(s) - \gamma \sum_{s'} p(s'|s, a) V(s') + \beta \log \frac{\pi(a|s)}{\max_{a' \in \mathcal{A}} \pi(a'|s)} \\
&= V^*(s; p, r) - \gamma \sum_{s'} p(s'|s, a) V^*(s'; p, r) + A^*(s, a; p, r).
\end{aligned}
$$

Then, the result follows by applying triangle's inequality and the bounds on $V^*(s;p,r)$ and $A^*(s,a;p,r)$ in the definition of $\mathscr{R}_{\mathcal{M}}^{\text{BIRL}}$.

For the second part, observe that, as in the proof of Proposition D.1, if $r \in [-r_{\max}, +r_{\max}]$ (with $r_{\max} > 0$), then its optimal value and advantage functions are bounded by $\frac{r_{\max}}{1-\gamma}$. By asking that this quantity is smaller than both C_1^{BIRL} and C_2^{BIRL}, we get the result. □

We are now ready to prove Proposition 4.1.

Proposition 4.1. *Let $\mathcal{M}$ be any MDP\R. If we choose $C_1^{OPT}, C_2^{OPT} \geqslant (1+\gamma)/(1-\gamma) =: \rho$, $C_1^{MCE}, C_2^{MCE} \geqslant (2+\lambda \log A)/(1-\gamma)$, and $C_1^{BIRL}, C_2^{BIRL} \geqslant 1/(1-\gamma)$, then, for any $m \in \{OPT, MCE, BIRL\}$:*

$$\mathfrak{R} \subseteq \mathscr{R}_{\mathcal{M}}^{m} \subseteq [-\rho C_1^m - C_2^m, +\rho C_1^m]^{SA}.$$

Proof. Apply Proposition D.1,D.2 and D.3 by choosing the values that guarantee that the sets in the lower inclusions contain the unit hypercube $\mathfrak{R}$. □

D.2 The New Priors are Unbiased

Before proving the results, we need the following proposition.

Proposition D.4. *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_0, p, \gamma)$ be an MDP\R and let π be any deterministic policy. Then, it holds that:*

$$|det(T_{p,\pi})| = |det(W_{p,\pi})|.$$

Proof. Assume that π prescribes action a_1 in every possible state. Then, $T_{p,\pi}$ can be written as:

$$T_{p,\pi} = \begin{bmatrix} 1-\gamma p(s_1|s_1,a_1) & -\gamma p(s_2|s_1,a_1) & \dots & -\gamma p(s_S|s_1,a_1) & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 1-\gamma p(s_1|s_1,a_2) & -\gamma p(s_2|s_1,a_2) & \dots & -\gamma p(s_S|s_1,a_2) & 1 & \dots & 0 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1-\gamma p(s_1|s_1,a_A) & -\gamma p(s_2|s_1,a_A) & \dots & -\gamma p(s_S|s_1,a_A) & 0 & \dots & 1 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ -\gamma p(s_1|s_2,a_1) & 1-\gamma p(s_2|s_2,a_1) & \dots & -\gamma p(s_S|s_2,a_1) & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ -\gamma p(s_1|s_2,a_2) & 1-\gamma p(s_2|s_2,a_2) & \dots & -\gamma p(s_S|s_2,a_2) & 0 & \dots & 0 & 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ -\gamma p(s_1|s_2,a_A) & 1-\gamma p(s_2|s_2,a_A) & \dots & -\gamma p(s_S|s_2,a_A) & 0 & \dots & 0 & 0 & \dots & 1 & \dots & 0 & \dots & 0 \\ -\gamma p(s_1|s_S,a_1) & -\gamma p(s_2|s_S,a_1) & \dots & 1-\gamma p(s_S|s_S,a_1) & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ -\gamma p(s_1|s_S,a_2) & -\gamma p(s_2|s_S,a_2) & \dots & 1-\gamma p(s_S|s_S,a_2) & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ -\gamma p(s_1|s_S,a_A) & -\gamma p(s_2|s_S,a_A) & \dots & 1-\gamma p(s_S|s_S,a_A) & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{bmatrix},$$

where on the right part of the matrix we do not have any 1 in all and only the rows corresponding to the action a_1, played by π. To compute $\det(T_{p,\pi})$, let us apply the Laplace's formula (see Theorem 1.5 of (Hefferon, 2009)) by iteratively expanding by cofactors from the last column SA up to column $S+1$ included. Clearly, since these columns are all zeros with exactly one 1, then after these iterations we obtain that, for some $x \in \{0,1\}$:

$$\det(T_{p,\pi}) = (-1)^x \det \begin{bmatrix} 1-\gamma p(s_1|s_1,a_1) & -\gamma p(s_2|s_1,a_1) & \dots & -\gamma p(s_S|s_1,a_1) \\ -\gamma p(s_1|s_2,a_1) & 1-\gamma p(s_2|s_2,a_1) & \dots & -\gamma p(s_S|s_2,a_1) \\ \dots & \dots & \dots & \dots \\ -\gamma p(s_1|s_S,a_1) & -\gamma p(s_2|s_S,a_1) & \dots & 1-\gamma p(s_S|s_S,a_1) \end{bmatrix},$$

since all the rows corresponding to the actions non prescribed by π are neglected. However, this matrix is exactly $W_{p,\pi}$, therefore, the result follows when $\pi(s) = a_1$ in every state.

Generalizing to arbitrary deterministic policies π is immediate once realizing that $W_{p,\pi}$ contains all and only the rows that, in $T_{p,\pi}$, are followed by only zeros. □

Proposition 4.2. *For any MDP\R $\mathcal{M}$ and any pair of deterministic policies $\pi_1, \pi_2 \in \Pi$, it holds that:*

$$vol(\mathscr{R}_{\mathcal{M}}^{OPT} \cap \mathcal{R}_{\mathcal{M},\pi_1}^{OPT,\mathcal{S}}) = vol(\mathscr{R}_{\mathcal{M}}^{OPT} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{OPT,\mathcal{S}}).$$

Proof. For any deterministic policy π in any MDP\R $\mathcal{M}$, we can write:

$$\text{vol}(\mathscr{R}_{\mathcal{M}}^{\text{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi}^{\text{OPT},\mathcal{S}}) \overset{(1)}{=} \text{vol}(\mathscr{R}_{\mathcal{M}}^{\text{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi}^{\text{OPT},\mathcal{S}} \cap \mathscr{R}_p^1) + \text{vol}(\mathscr{R}_{\mathcal{M}}^{\text{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi}^{\text{OPT},\mathcal{S}} \cap \mathscr{R}_p^{>1})$$

$$
\begin{aligned}
&\overset{(2)}{=} \mathrm{vol}(\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1) \\
&\overset{(3)}{=} |\det(T_{p,\pi})| \mathrm{vol}(\mathcal{V}_\pi) \\
&\overset{(4)}{=} |\det(W_{p,\pi})| \mathrm{vol}(\mathcal{V}_\pi) \\
&\overset{(5)}{=} k_\pi^{-S} \mathrm{vol}(\mathcal{V}_\pi),
\end{aligned}
$$

where at (1) we use the additivity of the integral (see Eq. (24) for a definition of $\mathscr{R}_p^1$ and $\mathscr{R}_p^{>1}$), at (2) we realize that the set $\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^{>1}$, containing all the rewards that induce at least two optimal policies from every state in $\mathcal{M}$ and with bounded value and advantage functions, has zero SA-dimensional volume since the optimality of another policy π' beyond π represents an additional constraint (the considered volume is SA-dimensional because of Proposition B.1, where it is shown that the OPT feasible set has SA degrees of freedom). At (3) we use the change of variables formula in Proposition 6.1.3 of (Cohn, 2013), after having observed that matrix[11] $T_{p,\pi}$ (along with Proposition B.1) permits to reparametrize set $\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1$ using set $\mathcal{V}_\pi$ as:

$$
\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1 = \Big\{ r \in \mathbb{R}^{SA} \,|\, \exists (V,A) \in \mathcal{V}_\pi : \, r = T_{p,\pi}(V,A) \Big\},
$$

where

$$
\mathcal{V}_\pi := \Big\{ V \in [-C_1^{\mathrm{OPT}} k_\pi, +C_1^{\mathrm{OPT}} k_\pi]^S, A \in [-C_2^{\mathrm{OPT}}, 0)^{S \times (A-1)} \Big\}.
$$

At (4) we use Proposition D.4 and at (5) we use the definition of k_π.

We can compute the volume of $\mathcal{V}_\pi$ as:

$$
\begin{aligned}
\mathrm{vol}(\mathcal{V}_\pi) &:= \int_{\mathcal{V}^\pi} dV dA \\
&= \int_{[-C_1^{\mathrm{OPT}} k_\pi, +C_1^{\mathrm{OPT}} k_\pi]^S} dV \cdot \int_{[-C_2^{\mathrm{OPT}}, 0)^{S \times (A-1)}} dA \\
&= \prod_{s \in \mathcal{S}} \int_{[-C_1^{\mathrm{OPT}} k_\pi, +C_1^{\mathrm{OPT}} k_\pi]} dV(s) \cdot \prod_{s \in \mathcal{S}, a \in \mathcal{A} \setminus \{\pi(s)\}} \int_{[-C_2^{\mathrm{OPT}}, 0)^{S \times (A-1)}} dA(s,a) \\
&= 2^S k_\pi^S (C_1^{\mathrm{OPT}})^S (C_2^{\mathrm{OPT}})^{S(A-1)}.
\end{aligned}
$$

Thus, replacing into the previous expression, we obtain:

$$
\mathrm{vol}(\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi}^{\mathrm{OPT},\mathcal{S}}) = 2^S (C_1^{\mathrm{OPT}})^S (C_2^{\mathrm{OPT}})^{S(A-1)}.
$$

Since this quantity is constant for all the deterministic policies π, this concludes the proof. □

Proposition 4.3. *For any MDP\R $\mathcal{M}$ and any pair of stochastic policies $\pi_1, \pi_2 \in \Pi$ such that $\pi_1(a|s) \geqslant \pi_{\min}$ and $\pi_2(a|s) \geqslant \pi_{\min}$ in every s, a, if $C_2^{MCE} \geqslant \lambda \log \frac{1}{\pi_{\min}}$, then:*

$$
vol(\mathscr{R}_{\mathcal{M}}^{MCE} \cap \mathcal{R}_{\mathcal{M},\pi_1}^{MCE,\mathcal{S}}) = vol(\mathscr{R}_{\mathcal{M}}^{MCE} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{MCE,\mathcal{S}}).
$$

Proof. For any MDP\R $\mathcal{M}$, observe that, for any stochastic policy π, if it does not hold that:

$$
\min_{(s,a) \in \mathcal{S} \times \mathcal{A}} A_\lambda^\pi(s,a;p,r) \geqslant -C_2^{\mathrm{MCE}},
$$

then the intersection $\mathscr{R}_{\mathcal{M}}^{\mathrm{MCE}} \cap \mathcal{R}_{\mathcal{M},\pi}^{\mathrm{MCE},\mathcal{S}}$ is empty. Since, by hypothesis, $\min_{(s,a) \in \mathcal{S} \times \mathcal{A}} \pi(a|s) \geqslant \pi_{\min}$, and since we know that (Cao et al., 2021; Haarnoja et al., 2017):

$$
A_\lambda^*(s,a;p,r) = \lambda \log \pi(a|s),
$$

then the above condition can be imposed to hold as long as:

$$
C_2^{\mathrm{MCE}} \geqslant \lambda \log \frac{1}{\pi_{\min}}.
$$

[11] The definition of $T_{p,\pi}$ is provided in Eq. (21).

For this reason, let π be a stochastic policy satisfying that $C_2^{\mathrm{MCE}} \geqslant \lambda \log \frac{1}{\pi_{\min}}$. Define $\eta(s,a) = \lambda \log \pi(a|s)$ for all s, a for conveniency. We can write:

$$\begin{aligned}\mathrm{vol}(\mathscr{R}_{\mathcal{M}}^{\mathrm{MCE}} \cap \mathcal{R}_{\mathcal{M},\pi}^{\mathrm{MCE},\mathcal{S}}) &\overset{(1)}{=} \sqrt{|\det(J_{p,\eta}^{\intercal} J_{p,\eta})|}\mathrm{vol}_S(\mathcal{V}_\pi') \\ &\overset{(2)}{=} \sqrt{|\det(J_{p,\eta}^{\intercal} J_{p,\eta})|} 2^S (C_1^{\mathrm{MCE}})^S,\end{aligned}$$

where at (1) we use the change of variables formula in (Boothby, 1986; Stoll, 2007), after having observed that the *rectangular*[12] matrix $U_{p,\eta}$, defined in Eq. (22), along with Proposition B.2, permits to reparametrize the set $\mathscr{R}_{\mathcal{M}}^{\mathrm{MCE}} \cap \mathcal{R}_{\mathcal{M},\pi}^{\mathrm{MCE},\mathcal{S}}$ using set $\mathcal{V}_\pi'$ as:

$$\mathscr{R}_{\mathcal{M}}^{\mathrm{MCE}} \cap \mathcal{R}_{\mathcal{M},\pi}^{\mathrm{MCE},\mathcal{S}} = \Big\{ r \in \mathbb{R}^{SA} \mid \exists V \in \mathcal{V}_\pi' : r = U_{p,\eta}(V) \Big\},$$

where

$$\mathcal{V}_\pi' := [-C_1^{\mathrm{MCE}}, +C_1^{\mathrm{MCE}}]^S.$$

Specifically, we use symbol $J_{p,\eta}(V) \in \mathbb{R}^{SA \times S}$ to denote the Jacobian of the operator $U_{p,\eta}$ evaluated in any $V \in \mathbb{R}^S$; however, since $J_{p,\eta}(V)$ is a constant function of V (as shown below), then we use $J_{p,\eta}$. At (2) we compute the volume of set $\mathcal{V}_\pi'$, which is an hypercube.

Observe that $2^S (C_1^{\mathrm{MCE}})^S$ does not depend on the policy π. The proof is concluded after having observed that the Jacobian $J_{p,\eta}$ is a constant function. Indeed, for any $V \in \mathbb{R}^S$, we have:

$$\begin{aligned}J_{p,\eta}(V) &:= \begin{bmatrix} \frac{\partial [U_{p,\eta}(V)]_{s_1,a_1}}{\partial V(s_1)} & \frac{\partial [U_{p,\eta}(V)]_{s_1,a_1}}{\partial V(s_2)} & \dots & \frac{\partial [U_{p,\eta}(V)]_{s_1,a_1}}{\partial V(s_S)} \\ \frac{\partial [U_{p,\eta}(V)]_{s_1,a_2}}{\partial V(s_1)} & \frac{\partial [U_{p,\eta}(V)]_{s_1,a_2}}{\partial V(s_2)} & \dots & \frac{\partial [U_{p,\eta}(V)]_{s_1,a_2}}{\partial V(s_S)} \\ \dots & \dots & \dots & \dots \\ \frac{\partial [U_{p,\eta}(V)]_{s_S,a_A}}{\partial V(s_1)} & \frac{\partial [U_{p,\eta}(V)]_{s_S,a_A}}{\partial V(s_2)} & \dots & \frac{\partial [U_{p,\eta}(V)]_{s_S,a_A}}{\partial V(s_S)} \end{bmatrix} \\ &= \begin{bmatrix} 1-\gamma p(s_1|s_1,a_1) & -\gamma p(s_2|s_1,a_1) & \dots & -\gamma p(s_S|s_1,a_1) \\ 1-\gamma p(s_1|s_1,a_2) & -\gamma p(s_2|s_1,a_2) & \dots & -\gamma p(s_S|s_1,a_2) \\ \dots & \dots & \dots & \dots \\ -\gamma p(s_1|s_S,a_A) & -\gamma p(s_2|s_S,a_A) & \dots & 1-\gamma p(s_S|s_S,a_A) \end{bmatrix}.\end{aligned}$$

□

Proposition 4.4. *For any MDP\R $\mathcal{M}$ and any pair of stochastic policies $\pi_1, \pi_2 \in \Pi$ such that $\pi_1(a|s) \geqslant \pi_{\min}$ and $\pi_2(a|s) \geqslant \pi_{\min}$ in every s, a, if $C_2^{BIRL} \geqslant \beta \log \frac{1}{A\pi_{\min}}$, then:*

$$vol(\mathscr{R}_{\mathcal{M}}^{BIRL} \cap \mathcal{R}_{\mathcal{M},\pi_1}^{BIRL,\mathcal{S}}) = vol(\mathscr{R}_{\mathcal{M}}^{BIRL} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{BIRL,\mathcal{S}}).$$

Proof. For any MDP\R $\mathcal{M}$, observe that, for any stochastic policy π, if it does not hold that:

$$\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} A^\pi(s,a;p,r) \geqslant -C_2^{\mathrm{BIRL}},$$

then the intersection $\mathscr{R}_{\mathcal{M}}^{\mathrm{BIRL}} \cap \mathcal{R}_{\mathcal{M},\pi}^{\mathrm{BIRL},\mathcal{S}}$ is empty. Since, by hypothesis, $\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \pi(a|s) \geqslant \pi_{\min}$, and since we know that (see (Skalse et al., 2023) and Proposition B.3):

$$A^*(s,a;p,r) = \beta \log \frac{\pi(a|s)}{\max_{a'\in\mathcal{A}} \pi(a'|s)},$$

then the above condition can be imposed to hold as long as:

$$C_2^{\mathrm{BIRL}} \geqslant \beta \log \frac{1}{A\pi_{\min}},$$

where we also used that $\log \max_a \pi(a|s) \geqslant \log \frac{1}{A}$. Having made this condition to hold, the proof is then analogous to that of Proposition 4.3, with the only difference that we define $\eta(s,a) = \beta \log(\pi(a|s)/\max_{a'\in\mathcal{A}} \pi(a'|s))$ at all s, a. □

[12] The set $\mathscr{R}_{\mathcal{M}}^{\mathrm{MCE}} \cap \mathcal{R}_{\mathcal{M},\pi}^{\mathrm{MCE},\mathcal{S}}$ is S-dimensional but is embedded into $\mathbb{R}^{SA}$, so we cannot apply Proposition 6.1.3 of (Cohn, 2013).

Figure 8: MDP\R considered in the proof of Proposition D.5. In the left there is $\mathcal{M}$, while on the right there is $\mathcal{M}'$.

D.3 The New Priors can be Biased in new Environments

We now show that the new priors $w^m_{\mathcal{M}}$ can be biased in environments different from $\mathcal{M}$.

Proposition D.5. *There exists a pair of MDP\R $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_0, p, \gamma)$ and $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, s_0, p, \gamma')$ with two states s_1, s_2 and two actions a_1, a_2 in s_1 and one action a_1 in s_2, and a pair of deterministic policies π_1, π_2 such that, if we set $C_1^{OPT} = 1$ and if we let $\gamma \to 1$, then for any choice of $C_2^{OPT} > 0$, we have:*

$$vol(\mathscr{R}^{OPT}_{\mathcal{M}} \cap \mathcal{R}^{OPT,\mathcal{S}}_{\mathcal{M}',\pi_1}) \neq vol(\mathscr{R}^{OPT}_{\mathcal{M}} \cap \mathcal{R}^{OPT,\mathcal{S}}_{\mathcal{M}',\pi_2}).$$

Proof. For the proof, we consider the $\mathcal{M}, \mathcal{M}'$ in Fig. 8, and we define $x := r(s_1, a_1), y := r(s_1, a_2), z := r(s_1, a_1)$ for simplicity. Let π_1, π_2 be the deterministic policies that play actions a_1, a_2 in state s_1. Now, let us make explicit the sets $\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}}$, $\mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M}',\pi_1}$ and $\mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M}',\pi_2}$. It is immediate to observe that, in $\mathcal{M}'$, for $\gamma \to 1$, policy π_1 is optimal if $x \geqslant z$, and viceversa π_2 is optimal if $x \leqslant z$:

$$\mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M}',\pi_1} = \Big\{x, y, z \in \mathbb{R} \,|\, x \geqslant z\Big\},$$
$$\mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M}',\pi_2} = \Big\{x, y, z \in \mathbb{R} \,|\, x \leqslant z\Big\}.$$

Regarding $\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}}$, we need to compute k_π for any possible (stochastic) policy $\pi \in \Pi$. We have:

$$(k_\pi)^{-S} = |\det(W_{p,\pi})| = |\det \begin{bmatrix} 1-\gamma & 0 \\ 0 & 1-\gamma \end{bmatrix}| = (1-\gamma)^2.$$

Since we have two states, $S = 2$ and so:

$$k_\pi = \frac{1}{1-\gamma},$$

for every possible policy π.[13] To represent set $\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}}$ explicitly as a function of x, y, z, observe that it can be rewritten as:

$$\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}} = (\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M},\pi_1} \cap \mathscr{R}^1_p) \cup (\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M},\pi_2} \cap \mathscr{R}^1_p) \cup (\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}} \cap \mathscr{R}^{>1}_p),$$

where set $\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}} \cap \mathscr{R}^{>1}_p$ has zero 3-dimensional volume and so is uninteresting. By using Proposition B.1 and also that $(1-\gamma)k_\pi = 1$ for any π and γ, it can be shown that:

$$\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M},\pi_1} \cap \mathscr{R}^1_p = \Big\{x, y, z \in \mathbb{R} \,|\, x, z \in [-C_1^{\mathrm{OPT}}, +C_1^{\mathrm{OPT}}] \wedge y \in [x - C_2^{\mathrm{OPT}}, x]\Big\},$$
$$\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M},\pi_2} \cap \mathscr{R}^1_p = \Big\{x, y, z \in \mathbb{R} \,|\, y, z \in [-C_1^{\mathrm{OPT}}, +C_1^{\mathrm{OPT}}] \wedge x \in [y - C_2^{\mathrm{OPT}}, y]\Big\}.$$

Since $\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}} \cap \mathscr{R}^{>1}_p$ has zero 3-dimensional volume, we can write:

$$\begin{aligned}
\mathrm{vol}(\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M}',\pi_1}) &= \mathrm{vol}(\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M},\pi_1} \cap \mathscr{R}^1_p \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M}',\pi_1}) \\
&\quad + \mathrm{vol}(\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M},\pi_2} \cap \mathscr{R}^1_p \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M}',\pi_1}) \\
&= \mathrm{vol}(\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M},\pi_1} \cap \mathscr{R}^1_p \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M}',\pi_2}) \\
&\quad + \mathrm{vol}(\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M},\pi_2} \cap \mathscr{R}^1_p \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M}',\pi_1}),
\end{aligned}$$

[13] To be precise, we have never shown in the paper that this choice of set gives the same volume $\mathrm{vol}(\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M},\pi_1}) \neq \mathrm{vol}(\mathscr{R}^{\mathrm{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\mathrm{OPT},\mathcal{S}}_{\mathcal{M},\pi_2})$ in any $\mathcal{M}$ for any possible pair of policies π_1, π_2 even though some states might have a number of actions different than that of other states. Nevertheless, this can be proved rather easily analogously to Proposition 4.2. Of course, an analogous of Proposition B.1 will be needed.

where in the second passage we have observed that $\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_1}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1$ allows x, z to vary freely in the same interval, therefore, using the formulas computed earlier for $\mathcal{R}_{\mathcal{M}',\pi_1}^{\mathrm{OPT},\mathcal{S}}$ and $\mathcal{R}_{\mathcal{M}',\pi_2}^{\mathrm{OPT},\mathcal{S}}$, we realize that by symmetry $\mathrm{vol}(\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_1}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1 \cap \mathcal{R}_{\mathcal{M}',\pi_1}^{\mathrm{OPT},\mathcal{S}}) = \mathrm{vol}(\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_1}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1 \cap \mathcal{R}_{\mathcal{M}',\pi_2}^{\mathrm{OPT},\mathcal{S}})$.

For this reason, the claim in the proposition can be proved by showing that:

$$\mathrm{vol}(\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1 \cap \mathcal{R}_{\mathcal{M}',\pi_1}^{\mathrm{OPT},\mathcal{S}}) \neq \mathrm{vol}(\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1 \cap \mathcal{R}_{\mathcal{M}',\pi_2}^{\mathrm{OPT},\mathcal{S}}).$$

To make things even easier, we will prove it by first computing $\mathrm{vol}(\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1)$, then computing $\mathrm{vol}(\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1 \cap \mathcal{R}_{\mathcal{M}',\pi_1}^{\mathrm{OPT},\mathcal{S}})$, and finally showing that their ratio is different from 1/2:

$$\frac{\mathrm{vol}(\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1 \cap \mathcal{R}_{\mathcal{M}',\pi_1}^{\mathrm{OPT},\mathcal{S}})}{\mathrm{vol}(\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1)} \neq \frac{1}{2}.$$

We begin with the volume of $\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1$. We can write:

$$\begin{aligned}
\mathrm{vol}(\mathcal{R}_{p,\pi_2}) &= \int_{-C_1^{\mathrm{OPT}}}^{+C_1^{\mathrm{OPT}}} \int_{-C_1^{\mathrm{OPT}}}^{+C_1^{\mathrm{OPT}}} \int_{y-C_2^{\mathrm{OPT}}}^{y} dxdydz \\
&= \int_{-C_1^{\mathrm{OPT}}}^{+C_1^{\mathrm{OPT}}} \int_{-C_1^{\mathrm{OPT}}}^{+C_1^{\mathrm{OPT}}} C_2^{\mathrm{OPT}} dydz \\
&= 4C_1^{\mathrm{OPT}} C_2^{\mathrm{OPT}} = 4C_2^{\mathrm{OPT}},
\end{aligned}$$

by recalling that $C_1^{\mathrm{OPT}} = 1$ by hypothesis.

Now, concerning the volume of $\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1 \cap \mathcal{R}_{\mathcal{M}',\pi_1}^{\mathrm{OPT},\mathcal{S}}$, note that we can write this region more explicitly (modulo some zero-volume sets) as:

$$\begin{aligned}
\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1 \cap \mathcal{R}_{\mathcal{M}',\pi_1}^{\mathrm{OPT},\mathcal{S}} = \Big\{ x, y, z \in \mathbb{R} \mid & -1 \leqslant y \leqslant +1 \ \wedge \\
& -1 \leqslant z \leqslant +1 \ \wedge \\
& y - C_2^{\mathrm{OPT}} \leqslant x \leqslant y \ \wedge \\
& x \geqslant z \Big\}.
\end{aligned}$$

We partition it into 4 regions $\mathcal{R} := \mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1 \cap \mathcal{R}_{\mathcal{M}',\pi_1}^{\mathrm{OPT},\mathcal{S}} = I_1 \cup I_2 \cup I_3 \cup I_4$, so that the resulting volume will be the sum of the volumes of these regions:

$$\begin{aligned}
I_1 &:= \mathcal{R} \cap \{z \geqslant y - C_2^{\mathrm{OPT}} \ \wedge \ z \geqslant 1 - C_2^{\mathrm{OPT}}\}, \\
I_2 &:= \mathcal{R} \cap \{z \geqslant y - C_2^{\mathrm{OPT}} \ \wedge \ z \leqslant 1 - C_2^{\mathrm{OPT}}\}, \\
I_3 &:= \mathcal{R} \cap \{z \leqslant y - C_2^{\mathrm{OPT}} \ \wedge \ z \leqslant -1 - C_2^{\mathrm{OPT}}\}, \\
I_4 &:= \mathcal{R} \cap \{z \leqslant y - C_2^{\mathrm{OPT}} \ \wedge \ z \geqslant -1 - C_2^{\mathrm{OPT}}\}.
\end{aligned}$$

We remark that, for I_1, I_2, it holds that $z \leqslant y$ (otherwise they are empty).

Now, if we denote $m := \max\{-1, 1 - C_2^{\mathrm{OPT}}\}$, the volume of I_1 is (we must enforce $z \leqslant y$ otherwise this becomes the signed volume):

$$\begin{aligned}
I_1 &= \int_{m}^{+1} \int_{z}^{+1} \int_{z}^{y} dxdydz \\
&= \int_{m}^{+1} \int_{z}^{+1} (y - z) dydz
\end{aligned}$$

$$= \int_{m}^{+1} (\frac{z^2}{2} - z + \frac{1}{2}) dydz$$
$$= \frac{1}{6} - \frac{m^3}{6} + \frac{m^2}{2} - \frac{m}{2}.$$

Depending on C_2^{OPT}, we get different values:

$$I_1 = \begin{cases} \frac{(C_2^{\mathrm{OPT}})^3}{6} & \text{if } C_2^{\mathrm{OPT}} \leqslant 2, \\ \frac{4}{3} & \text{otherwise,} \end{cases}$$

where we used that, for $C_2^{\mathrm{OPT}} \geqslant 2$, we have $m = -1$, and replaced, and similarly for the other term:

$$\frac{1}{6} - \frac{(1-C)^3}{6} + \frac{(1-C)^2}{2} - \frac{(1-C)}{2} = \frac{(C_2^{\mathrm{OPT}})^3}{6}.$$

W.r.t. I_2, we can write:

$$I_2 = \int_{-1}^{1-C_2^{\mathrm{OPT}}} \int_{z}^{z+C_2^{\mathrm{OPT}}} \int_{z}^{y} dxdydz$$
$$= \int_{-1}^{1-C_2^{\mathrm{OPT}}} \int_{z}^{z+C_2^{\mathrm{OPT}}} (y-z) dydz$$
$$= \int_{-1}^{1-C_2^{\mathrm{OPT}}} (\frac{y^2}{2}\Big|_{z}^{z+C_2^{\mathrm{OPT}}} - zC_2^{\mathrm{OPT}}) dz$$
$$= \frac{(C_2^{\mathrm{OPT}})^2}{2} \int_{-1}^{1-C_2^{\mathrm{OPT}}} dz$$
$$= \frac{(C_2^{\mathrm{OPT}})^2}{2} (2 - C_2^{\mathrm{OPT}}),$$

where we are assuming that $C_2^{\mathrm{OPT}} \leqslant 2$. Otherwise, $I_2 = 0$.

Regarding I_3, simply note that it is empty, so the volume is 0, because we are asking that $z \leqslant -1 - C_2^{\mathrm{OPT}} \leqslant -1$.

Thus, w.r.t. I_4, we can write (the upper bound $1 - C_2^{\mathrm{OPT}}$ on z arises because we ask $z + C_2^{\mathrm{OPT}} \leqslant 1$):

$$I_4 = \int_{-1}^{1-C_2^{\mathrm{OPT}}} \int_{z+C_2^{\mathrm{OPT}}}^{1} \int_{y-C_2^{\mathrm{OPT}}}^{y} dxdydz$$
$$= \int_{-1}^{1-C_2^{\mathrm{OPT}}} \int_{z+C_2^{\mathrm{OPT}}}^{1} C dydz$$
$$= \int_{-1}^{1-C_2^{\mathrm{OPT}}} (1 - z - C_2^{\mathrm{OPT}}) C dz$$
$$= \frac{C_2^{\mathrm{OPT}}}{2} (C_2^{\mathrm{OPT}} - 2)^2,$$

which holds assuming $C_2^{\mathrm{OPT}} \leqslant 2$, otherwise it is 0.

Finally, we get that the volume of $\mathcal{R} = \mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1 \cap \mathcal{R}_{\mathcal{M}',\pi_1}^{\mathrm{OPT},\mathcal{S}}$ is (we sum all contributions as a function

of C_2^{OPT}):

$$\begin{aligned}
\mathrm{vol}(\mathcal{R}) &= I_1 + I_2 + I_3 + I_4 \\
&= \begin{cases} \frac{(C_2^{\mathrm{OPT}})^3}{6} + \frac{(C_2^{\mathrm{OPT}})^2}{2}(2 - C_2^{\mathrm{OPT}}) + 0 + \frac{C_2^{\mathrm{OPT}}}{2}(C_2^{\mathrm{OPT}} - 2)^2 & \text{if } C_2^{\mathrm{OPT}} \leqslant 2, \\ \frac{4}{3} + 0 + 0 + 0 & \text{otherwise,} \end{cases} \\
&= \begin{cases} \frac{(C_2^{\mathrm{OPT}})^3}{6} - (C_2^{\mathrm{OPT}})^2 + 2C_2^{\mathrm{OPT}} & \text{if } C_2^{\mathrm{OPT}} \leqslant 2, \\ \frac{4}{3} & \text{otherwise.} \end{cases}
\end{aligned}$$

By taking the ratio, we get:

$$\frac{\mathrm{vol}(\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1 \cap \mathcal{R}_{\mathcal{M}',\pi_1}^{\mathrm{OPT},\mathcal{S}})}{\mathrm{vol}(\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \cap \mathcal{R}_{\mathcal{M},\pi_2}^{\mathrm{OPT},\mathcal{S}} \cap \mathscr{R}_p^1)} = \begin{cases} \frac{1}{3C_2^{\mathrm{OPT}}} & \text{if } C_2^{\mathrm{OPT}} \geqslant 2, \\ \frac{(C_2^{\mathrm{OPT}})^2}{24} - \frac{C_2^{\mathrm{OPT}}}{4} + \frac{1}{2} & \text{if } 0 < C_2^{\mathrm{OPT}} < 2, \end{cases}$$

which is different from $\frac{1}{2}$ for every $C_2^{\mathrm{OPT}} > 0$.

This concludes the proof. □

E Additional Results and Proofs for Section 4.2

In Appendix E.1, we prove Proposition 4.5. In Appendix E.2, we provide the missing proofs for Theorems 4.6, 4.7 and 4.8, while in Appendix E.3 we prove Proposition 4.9 and Proposition E.1. Finally, Appendix E.4 contains additional centroid calculations for OPT.

E.1 Derivation of Eq. (15)

Proposition 4.5. *For every $w : \mathbb{R}^{SA} \to [0,1]$, let $r_{w,\pi_E} = \int_{\mathcal{R}_{\mathcal{M},\pi_E}^m} w(r)dr$. Then, Eq. (10) can be rewritten as:*

$$\widehat{\pi}_w \in \arg\max_{\pi' \in \Pi_{c,k}} V^{\pi'}(s_0'; p', r_{w,\pi_E}). \tag{15}$$

Proof. The derivation is analogous to the proof of Theorem 3 of (Ramachandran and Amir, 2007). Starting from Eq. (10), we can write:

$$\begin{aligned}
\widehat{\pi}_w &\in \arg\min_{\pi' \in \Pi_{c,k}} \int_{\mathcal{R}_{\mathcal{M},\pi_E}^m} w(r)\Big(\max_{\pi \in \Pi_{c,k}} V^{\pi}(s_0'; p', r) - V^{\pi'}(s_0'; p', r)\Big)dr \\
&\overset{(1)}{=} \arg\min_{\pi' \in \Pi_{c,k}} \left(\int_{\mathcal{R}_{\mathcal{M},\pi_E}^m} w(r) \max_{\pi \in \Pi_{c,k}} V^{\pi}(s_0'; p', r)dr - \int_{\mathcal{R}_{\mathcal{M},\pi_E}^m} w(r)V^{\pi'}(s_0'; p', r)dr\right) \\
&\overset{(2)}{=} \arg\min_{\pi' \in \Pi_{c,k}} -\int_{\mathcal{R}_{\mathcal{M},\pi_E}^m} w(r)V^{\pi'}(s_0'; p', r)dr \\
&\overset{(3)}{=} \arg\max_{\pi' \in \Pi_{c,k}} \int_{\mathcal{R}_{\mathcal{M},\pi_E}^m} w(r)V^{\pi'}(s_0'; p', r)dr \\
&\overset{(4)}{=} \arg\max_{\pi' \in \Pi_{c,k}} V^{\pi'}(s_0'; p', r_{w,\pi_E}),
\end{aligned}$$

where at (1) we use the linearity of the integral operator, at (2) we realize that the first term does not depend on π', and, as such, it can be removed, at (3) we remove the minus by replacing argmin with argmax, finally, at (4) we realize that $V^{\pi'}(s_0'; p', r)$ is a linear function of r (it can be seen as the dot product between the occupancy measure of π' and r (Puterman, 1994)), thus, by defining $r_{w,\pi_E} := \int_{\mathcal{R}_{\mathcal{M},\pi_E}^m} w(r)dr$, and recalling that $\mathcal{S}$ and $\mathcal{A}$ are finite, we get the result. □

E.2 The Centroids of the Feasible Sets w.r.t. our New Priors

Theorem 4.6 (OPT). *For any MDP\R $\mathcal{M}$ and deterministic policy π_E, the reward centroid $r^{OPT}_{\mathcal{M},\pi_E}$ after rescaling*[14] *is, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$:*

$$r^{OPT}_{\mathcal{M},\pi_E}(s,a) = \begin{cases} 1 & \text{if } s \in \mathcal{S}_{\mathcal{M},\pi_E} \ \wedge \ \pi_E(s) = a \\ 0 & \text{if } s \in \mathcal{S}_{\mathcal{M},\pi_E} \ \wedge \ \pi_E(s) \neq a \\ \frac{1}{A} & \text{if } s \notin \mathcal{S}_{\mathcal{M},\pi_E} \end{cases}. \tag{16}$$

Proof. We can write:

$$\begin{aligned}
r^{\text{OPT}}_{\mathcal{M},\pi_E} &\overset{(1)}{:=} \int_{\mathscr{R}^{\text{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E}} r dr \\
&\overset{(2)}{=} \int_{\mathscr{R}^{\text{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E} \cap \mathscr{R}^1_p} r dr + \int_{\mathscr{R}^{\text{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E} \cap \mathscr{R}^{>1}_p} r dr \\
&\overset{(3)}{=} \int_{\mathscr{R}^{\text{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E} \cap \mathscr{R}^1_p} r dr \\
&\overset{(4)}{=} \sum_{\pi \in [\pi_E]_{\mathcal{S}_{\mathcal{M},\pi_E}}} \int_{\mathscr{R}^{\text{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\text{OPT},\mathcal{S}}_{\mathcal{M},\pi} \cap \mathscr{R}^1_p} r dr \\
&\overset{(5)}{=} \sum_{\pi \in [\pi_E]_{\mathcal{S}_{\mathcal{M},\pi_E}}} \int_{\mathcal{V}_\pi} |\det(T_{p,\pi})| T_{p,\pi}(V,A) dV dA \\
&\overset{(6)}{=} \sum_{\pi \in [\pi_E]_{\mathcal{S}_{\mathcal{M},\pi_E}}} |\det(W_{p,\pi})| \int_{\mathcal{V}_\pi} T_{p,\pi}(V,A) dV dA \\
&\overset{(7)}{=} \sum_{\pi \in [\pi_E]_{\mathcal{S}_{\mathcal{M},\pi_E}}} |\det(W_{p,\pi})| T_{p,\pi}(\overline{V}_\pi, \overline{A}_\pi),
\end{aligned}$$

where at (1) we use the definition, at (2) we use the additivity property of the integral for partitioning in rewards that induce a single optimal policy $\mathscr{R}^1_p$ and those that do not (see Eq. 23), at (3) we recognize that set $\mathscr{R}^{\text{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E} \cap \mathscr{R}^{>1}_p$ has zero volume as $\mathscr{R}^{>1}_p$ imposes an additional constraint on the rewards in this set (using Proposition B.1 it can be seen that this set is described by at most $SA - 1 < SA$ parameters), at (4) we apply Corollary D.1 in (Lazzati et al., 2024) to observe that[15] $\mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E} = \bigcup_{\pi \in [\pi_E]_{\mathcal{S}_{\mathcal{M},\pi_E}}} \mathcal{R}^{\text{OPT},\mathcal{S}}_{\mathcal{M},\pi}$, and then we use the distributive property of intersection over union and the additivity of the integral,[16] at (5), similarly to the proof of Proposition 4.2, we use the change of variables formula in Theorem 6.1.7 of (Cohn, 2013), after having observed that matrix[17] $T_{p,\pi}$ (along with Proposition B.1) permits to reparametrize set $\mathscr{R}^{\text{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\text{OPT},\mathcal{S}}_{\mathcal{M},\pi} \cap \mathscr{R}^1_p$ using set $\mathcal{V}_\pi$ as:

$$\mathscr{R}^{\text{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\text{OPT},\mathcal{S}}_{\mathcal{M},\pi} \cap \mathscr{R}^1_p = \Big\{ r \in \mathbb{R}^{SA} \,|\, \exists (V,A) \in \mathcal{V}_\pi : \, r = T_{p,\pi}(V,A) \Big\},$$

where

$$\mathcal{V}_\pi := \Big\{ V \in [-C^{\text{OPT}}_1 k_\pi, +C^{\text{OPT}}_1 k_\pi]^S, A \in [-C^{\text{OPT}}_2, 0)^{S \times (A-1)} \Big\}.$$

At (6) we apply Proposition D.4, and note that $|\det(W_{p,\pi})|$ does not depend on V, A, at (7) we applied the linearity of operator $T_{p,\pi}$, and defined $\overline{V}_\pi := \int_{\mathcal{V}_\pi} V dV dA$ and $\overline{A}_\pi := \int_{\mathcal{V}_\pi} A dV dA$.

[14] By "rescaling" a vector $x \in \mathbb{R}^n$ we mean applying a transformation $\alpha x + \beta$ with $\alpha > 0, \beta \in \mathbb{R}$. It is well-known that rescaling a reward function does not affect the rank of policies (Ng et al., 1999; Sutton and Barto, 2018).

[15] See Appendix B for a definition of $[\cdot]_{\overline{\mathcal{S}}}$ for any set $\overline{\mathcal{S}} \subseteq \mathcal{S}$.

[16] To be precise, the sets $\mathscr{R}^{\text{OPT}}_{\mathcal{M}} \cap \mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E} \cap \mathscr{R}^1_p$ are open, so this passage cannot be done as-is, but additional formalism should be adopted. Nevertheless, we avoid introducing it here to keep the exposition from becoming unnecessarily cumbersome.

[17] The definition of $T_{p,\pi}$ is provided in Eq. (21).

Now, we turn to the computation of $\overline{V}_\pi$ and $\overline{A}_\pi$. For any state $\overline{s} \in \mathcal{S}$, we have:

$$\begin{aligned}
\overline{V}_\pi(\overline{s}) &:= \int_{\mathcal{V}^\pi} V(\overline{s}) dV dA \\
&= \int_{[-C_1^{\mathrm{OPT}} k_\pi, +C_1^{\mathrm{OPT}} k_\pi]^S} V(\overline{s}) dV \cdot \int_{[-C_2^{\mathrm{OPT}}, 0)^{S \times (A-1)}} dA \\
&= \prod_{s \in \mathcal{S} \setminus \{\overline{s}\}} \int_{[-C_1^{\mathrm{OPT}} k_\pi, +C_1^{\mathrm{OPT}} k_\pi]} dV(s) \cdot \int_{[-C_1^{\mathrm{OPT}} k_\pi, +C_1^{\mathrm{OPT}} k_\pi]} V(\overline{s}) dV(\overline{s}) \\
&\qquad \cdot \prod_{s \in \mathcal{S}, a \in \mathcal{A} \setminus \{\pi(s)\}} \int_{[-C_2^{\mathrm{OPT}}, 0)} dA(s,a) \\
&= (2C_1^{\mathrm{OPT}} k_\pi)^{S-1} \cdot (C_2^{\mathrm{OPT}})^{S(A-1)} \cdot \left[\frac{V(\overline{s})^2}{2} \right]_{-C_1^{\mathrm{OPT}} k_\pi}^{+C_1^{\mathrm{OPT}} k_\pi} \\
&= 0.
\end{aligned}$$

In a similar manner, for any $\overline{s} \in \mathcal{S}$, $\overline{a} \in \mathcal{A} \setminus \{\pi(\overline{s})\}$, we have:

$$\begin{aligned}
\overline{A}_\pi(\overline{s}, \overline{a}) &:= \int_{\mathcal{V}^\pi} A(\overline{s}, \overline{a}) dV dA \\
&= \int_{[-C_1^{\mathrm{OPT}} k_\pi, +C_1^{\mathrm{OPT}} k_\pi]^S} dV \cdot \int_{[-C_2^{\mathrm{OPT}}, 0)^{S \times (A-1)}} A(\overline{s}, \overline{a}) dA \\
&= \prod_{s \in \mathcal{S}} \int_{[-C_1^{\mathrm{OPT}} k_\pi, +C_1^{\mathrm{OPT}} k_\pi]} dV(s) \cdot \prod_{s \in \mathcal{S}, a \in \mathcal{A} \setminus \{\pi(s)\} \wedge (s,a) \neq (\overline{s}, \overline{a})} \int_{[-C_2^{\mathrm{OPT}}, 0)} dA(s,a) \\
&\qquad \cdot \int_{[-C_2^{\mathrm{OPT}}, 0)} A(\overline{s}, \overline{a}) dA(\overline{s}, \overline{a}) \\
&= (2C_1^{\mathrm{OPT}} k_\pi)^{S} \cdot (C_2^{\mathrm{OPT}})^{S(A-1)-1} \cdot \left[\frac{A(\overline{s}, \overline{a})^2}{2} \right]_{-C_2^{\mathrm{OPT}}}^{0} \\
&= -2^{S-1} \cdot (C_1^{\mathrm{OPT}} k_\pi)^{S} \cdot (C_2^{\mathrm{OPT}})^{S(A-1)+1}.
\end{aligned}$$

Replacing the values of $\overline{V}_\pi$ and $\overline{A}_\pi$ into the previous expression, we get that:

$$\begin{aligned}
r^{\mathrm{OPT}}_{\mathcal{M}, \pi_E}(s,a) &\overset{(8)}{=} \begin{cases} 0 & \text{if } s \in \mathcal{S}_{\mathcal{M}, \pi_E}, a = \pi_E(s), \\ -\sum_{\pi \in [\pi_E]_{\mathcal{S}_{\mathcal{M}, \pi_E}}} |\det(W_{p,\pi})| x_\pi & \text{if } s \in \mathcal{S}_{\mathcal{M}, \pi_E}, a \neq \pi_E(s), \\ -\sum_{\pi \in [\pi_E]_{\mathcal{S}_{\mathcal{M}, \pi_E}} : \pi(s) \neq a} |\det(W_{p,\pi})| x_\pi & \text{otherwise} \end{cases} \\
&\overset{(9)}{\propto} \begin{cases} 0 & \text{if } s \in \mathcal{S}_{\mathcal{M}, \pi_E}, a = \pi_E(s), \\ -\sum_{\pi \in [\pi_E]_{\mathcal{S}_{\mathcal{M}, \pi_E}}} 1 & \text{if } s \in \mathcal{S}_{\mathcal{M}, \pi_E}, a \neq \pi_E(s), \\ -\sum_{\pi \in [\pi_E]_{\mathcal{S}_{\mathcal{M}, \pi_E}} : \pi(s) \neq a} 1 & \text{otherwise} \end{cases} \\
&\overset{(10)}{=} \begin{cases} 0 & \text{if } s \in \mathcal{S}_{\mathcal{M}, \pi_E}, a = \pi_E(s), \\ -A^{S - |\mathcal{S}_{\mathcal{M}, \pi_E}|} & \text{if } s \in \mathcal{S}_{\mathcal{M}, \pi_E}, a \neq \pi_E(s), \\ -A^{S - |\mathcal{S}_{\mathcal{M}, \pi_E}| - 1}(A-1) & \text{otherwise} \end{cases} \\
&\overset{(11)}{\propto} \begin{cases} 1 & \text{if } s \in \mathcal{S}_{\mathcal{M}, \pi_E}, a = \pi_E(s), \\ 0 & \text{if } s \in \mathcal{S}_{\mathcal{M}, \pi_E}, a \neq \pi_E(s), \\ \frac{1}{A} & \text{otherwise,} \end{cases}
\end{aligned}$$

where at (8) we used that $T_{p,\pi}(\overline{V}_\pi, \overline{A}_\pi)(s,a) = 0$ at all $s \notin \mathcal{S}_{\mathcal{M}, \pi_E}$, for the policies $\pi \in [\pi_E]_{\mathcal{S}_{\mathcal{M}, \pi_E}}$ that prescribe $\pi(s) = a$, and we also defined $x_\pi := 2^{S-1} \cdot (C_1^{\mathrm{OPT}} k_\pi)^S \cdot (C_2^{\mathrm{OPT}})^{S(A-1)+1}$ for necessity of space, at (9) we use that $(k_\pi)^S = 1/|\det(W_{p,\pi})|$, and we divided everything by $2^{S-1} \cdot (C_1^{\mathrm{OPT}})^S \cdot (C_2^{\mathrm{OPT}})^{S(A-1)+1}$ (recall that

we provide a rescaled result), at (10) we use that $[\pi_E]_{\mathcal{S}_{\mathcal{M},\pi_E}}$ contains exactly $A^{S-|\mathcal{S}_{\mathcal{M},\pi_E}|}$ policies, while set $\{\pi \in [\pi_E]_{\mathcal{S}_{\mathcal{M},\pi_E}} : \pi(s) \neq a\}$ contains exactly $A^{S-|\mathcal{S}_{\mathcal{M},\pi_E}|-1}(A-1)$ policies, and at (11) we rescale everything by first summing $A^{S-|\mathcal{S}_{\mathcal{M},\pi_E}|}$, and then dividing by it.

This concludes the proof. □

Theorem 4.7 (MCE)**.** *Let $C_2^{MCE} \geqslant \lambda \log \frac{1}{\pi_{\min}}$. Then, for any MDP\R $\mathcal{M}$ in which $\mathcal{S}_{\mathcal{M},\pi_E} = \mathcal{S}$, the reward centroid $r^{MCE}_{\mathcal{M},\pi_E}$ after rescaling is, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$:*

$$r^{MCE}_{\mathcal{M},\pi_E}(s,a) = \log \pi_E(a|s). \tag{17}$$

Proof. The condition on C_2^{MCE} is made for the same reason as in the proof Proposition 4.3.

Define $\eta(s,a) = \lambda \log \pi_E(a|s)$ for all s,a. First, we remark that, since set $\mathscr{R}^{\mathrm{MCE}}_{\mathcal{M}} \cap \mathcal{R}^{\mathrm{MCE}}_{\mathcal{M},\pi_E}$ is S-dimensional (immediate from Proposition 4.3), then the volume elements below are S-dimensional. We can write:

$$\begin{aligned}
r^{\mathrm{MCE}}_{\mathcal{M},\pi_E} &\overset{(1)}{:=} \int_{\mathscr{R}^{\mathrm{MCE}}_{\mathcal{M}} \cap \mathcal{R}^{\mathrm{MCE}}_{\mathcal{M},\pi_E}} r dr \\
&\overset{(2)}{=} \int_{\mathscr{R}^{\mathrm{MCE}}_{\mathcal{M}} \cap \mathcal{R}^{\mathrm{MCE},\mathcal{S}}_{\mathcal{M},\pi_E}} r dr \\
&\overset{(3)}{=} \int_{\mathcal{V}'_{\pi_E}} \sqrt{|\det(J^{\intercal}_{p,\eta} J_{p,\eta})|} U_{p,\eta}(V) dV \\
&\overset{(4)}{=} \sqrt{|\det(J^{\intercal}_{p,\eta} J_{p,\eta})|} \int_{\mathcal{V}'_{\pi_E}} U_{p,\eta}(V) dV \\
&\overset{(5)}{\propto} \int_{\mathcal{V}'_{\pi_E}} U_{p,\eta}(V) dV \\
&\overset{(6)}{=} U_{p,\eta}\left(\int_{\mathcal{V}'_{\pi_E}} V dV \right) \\
&\overset{(7)}{=} U_{p,\eta}(0) \\
&\overset{(8)}{=} \eta,
\end{aligned}$$

where at (1) we use the definition of $r^{\mathrm{MCE}}_{\mathcal{M},\pi_E}$, at (2) we exploit the fact that, by hypothesis, $\mathcal{S}_{\mathcal{M},\pi_E} = \mathcal{S}$, at (3) we use the change of variables formula in (Boothby, 1986; Stoll, 2007), after having observed that the *rectangular*[18] matrix $U_{p,\eta}$, defined in Eq. (22), along with Proposition B.2, permits to reparametrize the set $\mathscr{R}^{\mathrm{MCE}}_{\mathcal{M}} \cap \mathcal{R}^{\mathrm{MCE},\mathcal{S}}_{\mathcal{M},\pi_E}$ using set $\mathcal{V}'_{\pi_E}$ as:

$$\mathscr{R}^{\mathrm{MCE}}_{\mathcal{M}} \cap \mathcal{R}^{\mathrm{MCE},\mathcal{S}}_{\mathcal{M},\pi_E} = \left\{ r \in \mathbb{R}^{SA} \mid \exists V \in \mathcal{V}'_{\pi_E} : r = U_{p,\eta}(V) \right\},$$

where

$$\mathcal{V}'_{\pi_E} := [-C_1^{\mathrm{MCE}}, +C_1^{\mathrm{MCE}}]^S.$$

Specifically, we use symbol $J_{p,\eta} \in \mathbb{R}^{SA \times S}$ to denote the Jacobian of the operator $U_{p,\eta}$, and we omit the dependence on V since, as observed in the proof of Proposition 4.3, it is constant in V. At (4), we bring it outside because it does not depend on V, at (5) we drop it as we will rescale, at (6) we use the linearity of $U_{p,\eta}$, at (7) we use that $\int_{\mathcal{V}'_{\pi_E}} V dV = 0 \in \mathbb{R}^S$, as the domain is symmetric, and at (8) we use the defintion of operator $U_{p,\eta}$.

The result can be obtained by rescaling the computed reward by λ. □

Theorem 4.8 (BIRL)**.** *Let $C_2^{BIRL} \geqslant \beta \log \frac{1}{A\pi_{\min}}$. Then, for any MDP\R $\mathcal{M}$ in which $\mathcal{S}_{\mathcal{M},\pi_E} = \mathcal{S}$, the reward*

[18] The set $\mathscr{R}^{\mathrm{MCE}}_{\mathcal{M}} \cap \mathcal{R}^{\mathrm{MCE},\mathcal{S}}_{\mathcal{M},\pi_E}$ is S-dimensional but is embedded into $\mathbb{R}^{SA}$, so we cannot apply Proposition 6.1.3 of (Cohn, 2013).

centroid $r^{BIRL}_{\mathcal{M},\pi_E}$ *after rescaling is,* $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$*:*

$$r^{BIRL}_{\mathcal{M},\pi_E}(s,a) = \log \frac{\pi_E(a|s)}{\max_{a' \in \mathcal{A}} \pi_E(a'|s)}. \tag{18}$$

Proof. The condition on C_2^{BIRL} is made for the same reason as in the proof Proposition 4.4.

The proof is analogous to that of Theorem 4.7, with the only difference that we define $\eta(s,a) = \beta \log(\pi_E(a|s)/\max_{a' \in \mathcal{A}} \pi_E(a'|s))$ at all s, a, and that we use Proposition B.3 in place of Proposition B.2. □

E.3 Proofs for Additional Results on the Centroids

Proposition 4.9. *Let* $m \in \{OPT, MCE, BIRL\}$ *be the model of behavior between* π_E *and* r_E *in an MDP*$\backslash R$ $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_0, p, \gamma)$*. Let* $\pi^* \in \arg\max_\pi V^\pi(s_0; p, r^m_{\mathcal{M},\pi_E})$ *be an optimal policy in* $\mathcal{M}$ *under* $r^m_{\mathcal{M},\pi_E}$*. Then,* π^* *is optimal also under* r_E*:* $V^{\pi^*}(s_0; p, r_E) = V^*(s_0; p, r_E)$*.*

Proof. Regarding MCE and BIRL, thanks to Propositions B.2 and B.3, it is immediate to notice that $r^{\text{MCE}}_{\mathcal{M},\pi_E}$ and $r^{\text{BIRL}}_{\mathcal{M},\pi_E}$ are obtained from r_E through reward shaping (Ng et al., 1999) (the state-only function is the value function). Therefore, they induce the same optimal policies.

Regarding OPT, observe that $V^{\pi_E}(s_0; p, r^{\text{OPT}}_{\mathcal{M},\pi_E}) = \frac{1}{1-\gamma}$, which is the maximum possible, thus $\pi^* = \pi_E$. Moreover, observe also that π_E is the unique optimal policy in $\mathcal{M}_{r^{\text{OPT}}_{\mathcal{M},\pi_E}}$ as playing a non-expert's action gives a reward smaller than 1. For these reasons, since π_E is optimal for r_E (by hypothesis of OPT), then the result follows. □

The next result shows that, in the absence of additional constraints, reward $r^{\text{BIRL}}_{\mathcal{M},\pi_E}$ transfers to new $\mathcal{M}'$ the optimal policy w.r.t. r_E in $\mathcal{M}$. Moreover, $r^{\text{OPT}}_{\mathcal{M},\pi_E}$ does so as well when $\mathcal{S}_{\mathcal{M},\pi_E} = \mathcal{S}$.

Proposition E.1. *Assume that* π_E *and* r_E *are related by BIRL in an MDP*$\backslash R$ $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_0, p, \gamma)$*, and that* $\mathcal{S}_{\mathcal{M},\pi_E} = \mathcal{S}$*. Then, for any MDP*$\backslash R$ $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, \gamma', s_0', p')$*, it holds that:*

$$\arg\max_{\pi \in \Pi} V^\pi(s_0; p', r^{BIRL}_{\mathcal{M},\pi_E}) = \arg\max_{\pi \in \Pi} V^\pi(s_0; p, r_E).$$

Moreover, if π_E *is deterministic and, still,* $\mathcal{S}_{\mathcal{M},\pi_E} = \mathcal{S}$*, then the same result holds also for OPT.*

Proof. Regarding the first statement, observe that reward $r^{\text{BIRL}}_{\mathcal{M},\pi_E}$, assuming that $\mathcal{S}_{\mathcal{M},\pi_E} = \mathcal{S}$, associates to at least one action in every possible state the reward 0. Indeed, if we take $\overline{a} \in \arg\max_{a' \in \mathcal{A}} \pi_E(a'|s)$, we have:

$$\log \frac{\pi_E(\overline{a}|s)}{\max_{a' \in \mathcal{A}} \pi_E(a'|s)} = \log 1 = 0.$$

Therefore, any greedy policy π^* that plays such action/s $\overline{a}$ achieves, for any possible p' and γ':

$$V^{\pi^*}(s_0; p, r^{\text{BIRL}}_{\mathcal{M},\pi_E}) = 0.$$

Since playing at least once an action different from $\overline{a}$ gives a smaller expected return (since its corresponding reward is $\leqslant 0$), then we have that such policy π^* is optimal in any $\mathcal{M}'$. In particular, π^* is optimal also in $\mathcal{M}$, and thanks to Proposition 4.9 this means that it is optimal also w.r.t. r_E. This concludes the proof of the first claim.

For the second claim, simply observe that, if $\mathcal{S}_{\mathcal{M},\pi_E} = \mathcal{S}$, then:

$$r^{\text{OPT}}_{\mathcal{M},\pi_E}(s,a) = \mathbb{1}\{a = \pi_E(s)\},$$

that makes π_E the optimal policy in any $\mathcal{M}'$. Since, by hypothesis of OPT, π_E is optimal also w.r.t. r_E, then the result follows. □

E.4 Additional Centroid Calculations for OPT

We now focus on OPT and we show that the centroid of the prior $w^{\text{OPT}}_{\mathcal{M}}$, after rescaling is 0, which is an additional interesting property of this prior.

Proposition E.2. *For any MDP\R $\mathcal{M}$, define:*

$$r_{\mathcal{M}}^{OPT} := \int_{\mathbb{R}^{SA}} w_{\mathcal{M}}^{OPT}(r) r dr.$$

Then, after rescaling, it holds that: $r_{\mathcal{M}}^{OPT} = 0$.

Proof. The proof is completely analogous to that of Theorem 4.6, with the difference that, instead of summing over the policies in $[\pi_E]_{\mathcal{S}_{\mathcal{M},\pi_E}}$, we sum over all the possible deterministic policies $\Pi^{\mathrm{D}} := \mathcal{A}^S$. Then, all the state-action pairs s, a are treated equally, and we obtain:

$$\begin{aligned} r_{\mathcal{M}}^{\mathrm{OPT}}(s,a) &= \sum_{\pi \in \Pi^{\mathrm{D}}} |\det(W_{p,\pi})| T_{p,\pi}(\overline{V}_\pi, \overline{A}_\pi)(s,a) \\ &= - \sum_{\pi \in \Pi^{\mathrm{D}} : \pi(s) \neq a} |\det(W_{p,\pi})| 2^{S-1} \cdot (C_1^{\mathrm{OPT}} k_\pi)^S \cdot (C_2^{\mathrm{OPT}})^{S(A-1)+1} \\ &= - \sum_{\pi \in \Pi^{\mathrm{D}} : \pi(s) \neq a} 2^{S-1} \cdot (C_1^{\mathrm{OPT}})^S \cdot (C_2^{\mathrm{OPT}})^{S(A-1)+1} \\ &= 2^{S-1} \cdot (C_1^{\mathrm{OPT}})^S \cdot (C_2^{\mathrm{OPT}})^{S(A-1)+1} \cdot A^{S-1}(A-1), \end{aligned}$$

which is the same for every s, a. Thus, by subtracting $2^{S-1} \cdot (C_1^{\mathrm{OPT}})^S \cdot (C_2^{\mathrm{OPT}})^{S(A-1)+1} \cdot A^{S-1}(A-1)$ by $r_{\mathcal{M}}^{\mathrm{OPT}}$ elementwise, we get the zero vector:

$$r_{\mathcal{M}}^{\mathrm{OPT}}(s,a) = 0 \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

This concludes the proof. □

We now show what happens if we replace prior $w_{\mathcal{M}}^{\mathrm{OPT}}$ with a new prior $w_{\mathcal{M},q}$ that is not uniform on $\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}}$ anymore, but depends on a vector $q \in \Delta^{\Pi^{\mathrm{D}}}$ in the same way:

$$w_{\mathcal{M},q}(r) := \begin{cases} q(\pi)/\mathrm{vol}(\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}}) & \text{if } r \in \mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}} \wedge \pi \in \Pi^*(p,r) \wedge |\Pi^*(p,r)|, \\ 0 & \text{otherwise.} \end{cases}$$

Simply put, we are assigning different weights on the rewards in $\mathscr{R}_{\mathcal{M}}^{\mathrm{OPT}}$ depending on the induced optimal policy (we neglect rewards that induce more than a single optimal policy because such set has zero volume). It can be shown that:

Theorem E.3 (OPT expert)**.** *For any MDP\R $\mathcal{M}$ and deterministic expert's policy π_E, define:*

$$r_{\mathcal{M},\pi_E,q}^{OPT} := \int_{\mathcal{R}_{\mathcal{M},\pi_E}^{OPT}} w_{\mathcal{M},q}(r) r dr.$$

Then, after rescaling, it holds that:

$$r_{\mathcal{M},\pi_E,q}^{OPT}(s,a) = \begin{cases} 1 & \text{if } s \in \mathcal{S}_{\mathcal{M},\pi_E}, \pi_E(s) = a, \\ 0 & \text{if } s \in \mathcal{S}_{\mathcal{M},\pi_E}, \pi_E(s) \neq a, \\ \frac{\sum_{\pi \in [\pi_E]_{\mathcal{S}_{\mathcal{M},\pi_E}} : \pi(s) \neq a} q(\pi)}{\sum_{\pi \in [\pi_E]_{\mathcal{S}_{\mathcal{M},\pi_E}}} q(\pi)} & \text{if } s \notin \mathcal{S}_{\mathcal{M},\pi_E}. \end{cases}$$

Proof. Again, the proof is completely analogous to that of Theorem 4.6. The only difference is that, after the first passages, we end up with:

$$\sum_{\pi \in [\pi_E]_{\mathcal{S}_{\mathcal{M},\pi_E}}} r_{\mathcal{M},\pi_E,q}^{\mathrm{OPT}} = \sum_{\pi \in [\pi_E]_{\mathcal{S}_{\mathcal{M},\pi_E}}} q(\pi) |\det(W_{p,\pi})| T_{p,\pi}(\overline{V}_\pi, \overline{A}_\pi),$$

as $w_{\mathcal{M},q}$ is not "uniform" over the policies. For this reason, later on in the proof, we obtain that, for all

$(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$r^{\text{OPT}}_{\mathcal{M},\pi_E,q}(s,a) \propto \begin{cases} 0 & \text{if } s \in \mathcal{S}_{\mathcal{M},\pi_E}, a = \pi_E(s), \\ -\sum_{\pi \in [\pi_E]_{\mathcal{S}_{\mathcal{M},\pi_E}}} q(\pi) & \text{if } s \in \mathcal{S}_{\mathcal{M},\pi_E}, a \neq \pi_E(s), \\ -\sum_{\pi \in [\pi_E]_{\mathcal{S}_{\mathcal{M},\pi_E}} : \pi(s) \neq a} q(\pi) & \text{otherwise,} \end{cases}$$

from which, after rescaling, we get the result. □

F Sample Complexity

This appendix is organized into four parts. In Appendix F.1, F.2 and F.3 we provide the proofs of, respectively, Theorems 5.1, 5.2 and 5.3. Finally, in Appendix F.4 we provide the proof of Proposition 5.4.

F.1 Proof of Theorem 5.1

Estimating $r^{\text{OPT}}_{\mathcal{M},\pi_E}$ correctly is equivalent to correctly estimating the support $\mathcal{S}_{\mathcal{M},\pi_E}$ and the set of state-action pairs $\mathcal{Z}_{\mathcal{M},\pi_E}$ visited by π_E in $\mathcal{M}$, which can be written as:

$$\mathcal{S}_{\mathcal{M},\pi_E} = \Big\{ s \in \mathcal{S} \;\Big|\; \exists t \in \mathbb{N} : \; (p^{\pi_E})^t(s|s_0) > 0 \Big\},$$
$$\mathcal{Z}_{\mathcal{M},\pi_E} := \Big\{ (s,a) \in \mathcal{S} \times \mathcal{A} \;\Big|\; s \in \mathcal{S}_{\mathcal{M},\pi_E} \wedge \pi_E(s) = a \Big\},$$

where $p^{\pi_E} \in \mathbb{R}^{S \times S}$ is the right stochastic matrix that, to each pair of states s, s', associates $p(s'|s, \pi_E(s))$, thus $(p^{\pi_E})^t(s|s_0)$ represents the probability of being in state s at stage t when starting from s_0 (Puterman, 1994). As reported in Algorithm 1, our estimators for sets $\mathcal{S}_{\mathcal{M},\pi_E}$ and $\mathcal{Z}_{\mathcal{M},\pi_E}$ are:

$$\widehat{\mathcal{S}} := \Big\{ s \in \mathcal{S} \;\Big|\; \exists i \in [\![N]\!], \exists h \in [\![H]\!] : \; s^i_h = s \Big\},$$
$$\widehat{\mathcal{Z}} := \Big\{ (s,a) \in \mathcal{S} \times \mathcal{A} \;\Big|\; \exists i \in [\![N]\!], \exists h \in [\![H]\!] : \; s^i_h = s \wedge a^i_h = a \Big\}.$$

We want to find a minimum number of samples NH that guarantees that, w.h.p., $\widehat{\mathcal{S}} = \mathcal{S}_{\mathcal{M},\pi_E}$ and $\widehat{\mathcal{Z}} = \mathcal{Z}_{\mathcal{M},\pi_E}$. To this aim, we must first show that taking $H \geqslant S$ is sufficient for guaranteeing that the dataset $\mathcal{D}_E$, made of trajectories long H-stages, can contain all the states and state-action pairs in the support, i.e., $p^{\min,H}_{\mathcal{M},\pi_E} > 0$:

Lemma F.1. *If $H \geqslant S$, then $p^{\min,H}_{\mathcal{M},\pi_E} > 0$.*

Proof. Observe that, since $H \geqslant S$, then:

$$\begin{aligned} p^{\min,H}_{\mathcal{M},\pi_E} &:= \min_{s \in \mathcal{S}_{\mathcal{M},\pi_E}} \mathbb{P}_{\mathcal{M},\pi_E}\big(\exists t \in [\![H]\!] : \; s_t = s\big) \\ &\geqslant \min_{s \in \mathcal{S}_{\mathcal{M},\pi_E}} \mathbb{P}_{\mathcal{M},\pi_E}\big(\exists t \in [\![S]\!] : \; s_t = s\big) \\ &\geqslant \min_{s \in \mathcal{S}_{\mathcal{M},\pi_E}} \min_{t \in [\![S]\!]} (p^{\pi_E})^t(s|s_0). \end{aligned}$$

Therefore, the result can be proved by showing that, for every state $s \in \mathcal{S}_{\mathcal{M},\pi_E}$, there exists a stage $t \in [\![S]\!]$ for which the probability $(p^{\pi_E})^t(s|s_0)$ of being in s at t is strictly greater than 0. But this is exactly the claim of Lemma F.2. This concludes the proof. □

Lemma F.2. *For every state $s \in \mathcal{S}_{\mathcal{M},\pi_E}$, there exists a $t \in [\![S]\!]$ such that $(p^{\pi_E})^t(s|s_0) > 0$.*

Proof. By hypothesis, since $s \in \mathcal{S}_{\mathcal{M},\pi_E}$, then there always exists a $t \in \mathbb{N}$ such that: $(p^{\pi_E})^t(s|s_0) > 0$. If $t \leqslant S$, then the result follows immediately, so we restrict to the case where $t > S$. By computing matrix $(p^{\pi_E})^t$, we can write:

$$(p^{\pi_E})^t(s|s_0) = \sum_{s^{(2)}, s^{(3)}, \ldots, s^{(t-1)} \in \mathcal{S}} p^{\pi_E}(s^{(2)}|s_0) p^{\pi_E}(s^{(3)}|s^{(2)}) \cdot \ldots \cdot p^{\pi_E}(s|s^{(t-1)}) > 0,$$

which is strictly positive as long as at least one specific path $s_0, \overline{s}^{(2)}, \dots, \overline{s}^{(t-1)}, s$ has strictly positive probability, which is the hypothesis made by contradiction:

$$(p^{\pi_E})^t(s|s_0) > 0 \implies \\ \exists \overline{s}^{(2)}, \overline{s}^{(3)}, \dots, \overline{s}^{(t-1)} : \; p^{\pi_E}(\overline{s}^{(2)}|s_0)p^{\pi_E}(\overline{s}^{(3)}|\overline{s}^{(2)}) \cdot \dots \cdot p^{\pi_E}(s|\overline{s}^{(t-1)}) > 0.$$

But this means that all these probabilities are strictly positive:

$$p^{\pi_E}(\overline{s}^{(i)}|\overline{s}^{(i-1)}) > 0 \; \forall i \in \{2, t\},$$

where $\overline{s}^{(1)} = s_0, \overline{s}^{(t)} = s$.

Now, if all the states $\overline{s}^{(i)}$ are distinct, then $t \leqslant S$, because there are at most S states. Otherwise, there exist $i \neq i'$ such that $\overline{s}^{(i)} = \overline{s}^{(i')}$, and so we can get rid of all the terms $p^{\pi_E}(\overline{s}^{(j)}|\overline{s}^{(j-1)})$ for $j \in \{i+1, i'\}$. Then, iterating this reasoning, we will end up with all distinct states, representing a path long $t' \leqslant S$ for which $(p^{\pi_E})^{t'}(s|s_0) > 0$.

This concludes the proof. □

We are now ready to upper bound the sample complexity necessity for estimating $\mathcal{S}_{\mathcal{M},\pi_E}$ and $\mathcal{Z}_{\mathcal{M},\pi_E}$:

Lemma F.3. *Let $\delta \in (0,1)$ and $H \geqslant S$. If π_E is deterministic, then it holds that:*

$$\mathbb{P}_{\mathcal{M},\pi_E}\Big(\widehat{\mathcal{S}} = \mathcal{S}_{\mathcal{M},\pi_E} \wedge \widehat{\mathcal{Z}} = \mathcal{Z}_{\mathcal{M},\pi_E}\Big) \geqslant 1 - \delta,$$

with a number of trajectories:

$$N \leqslant \frac{\log \frac{|\mathcal{S}_{\mathcal{M},\pi_E}|}{\delta}}{p^{\min,H}_{\mathcal{M},\pi_E}}.$$

Proof. If π_E is deterministic, then observe that $\widehat{\mathcal{S}} = \mathcal{S}_{\mathcal{M},\pi_E}$ implies condition $\widehat{\mathcal{Z}} = \mathcal{Z}_{\mathcal{M},\pi_E}$. Thus, we will focus on the former equality. We bound the opposite event:

$$\begin{aligned}
\mathbb{P}_{\mathcal{M},\pi_E}\Big(\widehat{\mathcal{S}} \neq \mathcal{S}_{\mathcal{M},\pi_E}\Big) &= \mathbb{P}_{\mathcal{M},\pi_E}\Big(\exists s \in \mathcal{S}_{\mathcal{M},\pi_E}, \forall i \in [\![N]\!], \nexists h \in [\![H]\!] : s^i_h = s\Big) \\
&\overset{(1)}{\leqslant} \sum_{s \in \mathcal{S}_{\mathcal{M},\pi_E}} \mathbb{P}_{\mathcal{M},\pi_E}\Big(\forall i \in [\![N]\!], \nexists h \in [\![H]\!] : s^i_h = s\Big) \\
&\overset{(2)}{=} \sum_{s \in \mathcal{S}_{\mathcal{M},\pi_E}} \prod_{i=1}^{N} \mathbb{P}_{\mathcal{M},\pi_E}\Big(\nexists h \in [\![H]\!] : s^i_h = s\Big) \\
&= \sum_{s \in \mathcal{S}_{\mathcal{M},\pi_E}} \prod_{i=1}^{N} \Big(1 - \mathbb{P}_{\mathcal{M},\pi_E}\Big(\exists h \in [\![H]\!] : s^i_h = s\Big)\Big) \\
&\overset{(3)}{\leqslant} \sum_{s \in \mathcal{S}_{\mathcal{M},\pi_E}} \prod_{i=1}^{N} (1 - p^{\min,H}_{\mathcal{M},\pi_E}) \\
&= |\mathcal{S}_{\mathcal{M},\pi_E}|(1 - p^{\min,H}_{\mathcal{M},\pi_E})^N,
\end{aligned}$$

where at (1) we apply a union bound, at (2) we use the fact that different trajectories are independent of each other, at (3) we lower bound $\mathbb{P}_{\mathcal{M},\pi_E}\Big(\exists h \in [\![H]\!] : s^i_h = s\Big)$ using $p^{\min,H}_{\mathcal{M},\pi_E}$, which is > 0 because of $H \geqslant S$ and Lemma F.1.

Forcing this probability to be at most δ, and solving w.r.t. N, we get:

$$\begin{aligned}
|\mathcal{S}_{\mathcal{M},\pi_E}|(1 - p^{\min,H}_{\mathcal{M},\pi_E})^N \leqslant \delta \iff N \log(1 - p^{\min,H}_{\mathcal{M},\pi_E}) &\leqslant -\log \frac{|\mathcal{S}_{\mathcal{M},\pi_E}|}{\delta} \\
\iff N &\geqslant \frac{\log \frac{|\mathcal{S}_{\mathcal{M},\pi_E}|}{\delta}}{\log \frac{1}{1 - p^{\min,H}_{\mathcal{M},\pi_E}}}.
\end{aligned}$$

By using that, for any $x \in (0,1)$, we have: $\log \frac{1}{1-x} \geqslant x$, we get the result. □

There are some interesting observations that can be done on this sample complexity result. Specifically, note that there is a trade-off in the choice of H. If H is large, then $p^{\min,H}_{\mathcal{M},\pi_E}$ is also large, and so N gets small. However, at the same time, increasing H we are using more samples since the trajectories are longer. The number of samples is:

$$NH \leqslant \frac{H}{p^{\min,H}_{\mathcal{M},\pi_E}} \log \frac{|\mathcal{S}_{\mathcal{M},\pi_E}|}{\delta}.$$

This expression can be optimized by H depending on the problem at stake, by solving:

$$\operatorname*{arg\,min}_{H\in\mathbb{N}:H\geqslant S} \frac{H}{p^{\min,H}_{\mathcal{M},\pi_E}} \log \frac{|\mathcal{S}_{\mathcal{M},\pi_E}|}{\delta} = \operatorname*{arg\,min}_{H\in\mathbb{N}:H\geqslant S} \frac{H}{p^{\min,H}_{\mathcal{M},\pi_E}}.$$

Finally, we are ready to prove the main theorem.

Theorem 5.1. *Let $\delta \in (0,1)$ and $H \geqslant S$. If π_E is deterministic, then, with probability at least $1-\delta$, it holds that $\widehat{r}^{OPT} = r^{OPT}_{\mathcal{M},\pi_E}$ provided that:*

$$NH \geqslant \frac{H \log \frac{|\mathcal{S}_{\mathcal{M},\pi_E}|}{\delta}}{p^{\min,H}_{\mathcal{M},\pi_E}}.$$

Proof. Simply, apply Lemma F.3, and realize from the definition of $r^{\mathrm{OPT}}_{\mathcal{M},\pi_E}$ that $\widehat{\mathcal{S}} = \mathcal{S}_{\mathcal{M},\pi_E} \wedge \widehat{\mathcal{Z}} = \mathcal{Z}_{\mathcal{M},\pi_E}$ suffices for having $\widehat{r}^{\mathrm{OPT}} = r^{\mathrm{OPT}}_{\mathcal{M},\pi_E}$. □

F.2 Proof of Theorem 5.2

We begin by proving how the error in the estimation of the policy propagates to the estimation of the centroid:

Lemma F.4. *Assume that $\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \pi_E(a|s) \geqslant \pi'_{\min}$. Let $\widehat{\pi}$ be the estimate of policy computed in Algorithm 2. Then, it holds that:*

$$\|\widehat{r}^{MCE} - r^{MCE}_{\mathcal{M},\pi_E}\|_\infty \leqslant \frac{1}{\pi_{\min}} \|\widehat{\pi} - \pi_E\|_\infty.$$

Proof. We can write:

$$\begin{aligned}
\|\widehat{r}^{\mathrm{MCE}} - r^{\mathrm{MCE}}_{\mathcal{M},\pi_E}\|_\infty &:= \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Big|\widehat{r}^{\mathrm{MCE}}(s,a) - r^{\mathrm{MCE}}_{\mathcal{M},\pi_E}(s,a)\Big| \\
&= \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Big|\log \widehat{\pi}(a|s) - \log \pi_E(a|s)\Big| \\
&= \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Big|\log \frac{\widehat{\pi}(a|s)}{\pi_E(a|s)}\Big| \\
&\overset{(1)}{\leqslant} \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\big|\widehat{\pi}(a|s) - \pi_E(a|s)\big|}{\min\{\widehat{\pi}(a|s), \pi_E(a|s)\}} \\
&\overset{(2)}{\leqslant} \frac{1}{\pi'_{\min}} \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big|\widehat{\pi}(a|s) - \pi_E(a|s)\big|,
\end{aligned}$$

where at (1) we use that $\log(1+x) \leqslant x$ for $x \geqslant 0$. Indeed, when $\widehat{\pi}(a|s) \geqslant \pi_E(a|s)$, then $\widehat{\pi}(a|s)/\pi_E(a|s) \geqslant 1$, and so: $\log \frac{\widehat{\pi}(a|s)}{\pi_E(a|s)} = \log\Big(1 + \frac{\widehat{\pi}(a|s)}{\pi_E(a|s)} - 1\Big) \leqslant (\widehat{\pi}(a|s) - \pi_E(a|s))/\pi_E(a|s)$. Doing similarly in the opposite case, we get the minimum at denominator. At (2) we use that $\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \pi_E(a|s) \geqslant \pi'_{\min}$ by hypothesis and the definition of $\widehat{\pi}$. □

The next step consists in upper bounding the estimation error of π_E:

Lemma F.5. *Let $\epsilon, \delta \in (0,1)$ and assume that $\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \pi_E(a|s) \geqslant \pi'_{\min}$. Then, it holds that:*

$$\mathbb{P}_{\mathcal{M},\pi_E}\Big(\|\widehat{\pi} - \pi_E\|_\infty \leqslant \epsilon\Big) \geqslant 1-\delta,$$

with a number of trajectories:

$$N \leqslant \frac{16 \log \frac{4SA}{\delta} \log \frac{2SA}{\delta}}{\epsilon^2 p^{\min,H}_{\mathcal{M},\pi_E}}.$$

Proof. For all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we can write:

$$\begin{aligned}
\left|\widehat{\pi}(a|s) - \pi_E(a|s)\right| &= \left|\max\left\{\pi'_{\min}, \frac{N(s,a)}{\max\{1, N(s)\}}\right\} - \pi_E(a|s)\right| \\
&\overset{(1)}{=} \left|\max\left\{\pi'_{\min}, \frac{N(s,a)}{\max\{1, N(s)\}}\right\} - \max\left\{\pi'_{\min}, \pi_E(a|s)\right\}\right| \\
&\overset{(2)}{\leqslant} \max\left\{\left|\pi'_{\min} - \pi'_{\min}\right|, \left|\frac{N(s,a)}{\max\{1, N(s)\}} - \pi_E(a|s)\right|\right\} \\
&= \left|\frac{N(s,a)}{\max\{1, N(s)\}} - \pi_E(a|s)\right|, \\
&\overset{(3)}{\leqslant} \sqrt{\frac{2 \log \frac{4}{\delta}}{\max\{1, N(s)\}}} \\
&\overset{(4)}{\leqslant} \sqrt{\frac{16 \log \frac{4}{\delta} \log \frac{2}{\delta}}{N p^{\min,H}_{\mathcal{M},\pi_E}}},
\end{aligned}$$

where at (1) we use that $\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \pi_E(a|s) \geqslant \pi'_{\min}$ by hypothesis, at (2) we use that the max operator is 1-Lipschitz, at (3) we use that, when $N(s) = 0$, then $|\pi_E(a|s)| \leqslant 1 \leqslant \sqrt{2 \log \frac{4}{\delta}}$, while when $N(s) \geqslant 1$, we apply the double-sided Azuma-Hoeffding's inequality (that must be applied because $N(s)$ is a random variable itself), which holds w.p. $1 - \delta/2$. At (4) we note that each $N(s) \sim \mathrm{Bin}(N, \mathbb{P}_{\mathcal{M},\pi_E}(\bigcup_{t=0}^{H}\{s_t = s\}))$ is Binomially distributed, so we can apply Lemma A.1 in (Xie et al., 2021) w.p. $\delta/2$ to get that: $\frac{1}{\max\{1,N(s)\}} \leqslant \frac{8 \log \frac{2}{\delta}}{N \mathbb{P}_{\mathcal{M},\pi_E}(\bigcup_{t=0}^{H}\{s_t = s\})} \leqslant \frac{8 \log \frac{2}{\delta}}{N p^{\min,H}_{\mathcal{M},\pi_E}}$, where we used the definition of $p^{\min,H}_{\mathcal{M},\pi_E}$ in Eq. (19).

By applying a union bound, we have that the above bound for s, a holds w.p. $1 - \delta$. In addition, the error above is bounded by ϵ as long as:

$$\sqrt{\frac{16 \log \frac{4}{\delta} \log \frac{2}{\delta}}{N p^{\min,H}_{\mathcal{M},\pi_E}}} \leqslant \epsilon \iff N \geqslant \frac{16 \log \frac{4}{\delta} \log \frac{2}{\delta}}{\epsilon^2 p^{\min,H}_{\mathcal{M},\pi_E}}.$$

The proof is concluded through the application of a union bound over all the $(s,a) \in \mathcal{S} \times \mathcal{A}$. □

We conclude with the main result:

Theorem 5.2. *Let $\epsilon, \delta \in (0,1)$ and $H \geqslant S$. Under the conditions of Theorem 4.7, if $\pi_{\min} \geqslant \pi'_{\min}$, then, with probability $1 - \delta$, we have $\|\widehat{r}^{MCE} - r^{MCE}_{\mathcal{M},\pi_E}\|_\infty \leqslant \epsilon$ provided that:*

$$NH \geqslant \frac{H \log^2 \frac{4SA}{\delta}}{p^{\min,H}_{\mathcal{M},\pi_E}} \cdot \frac{16}{\epsilon^2 \pi'_{\min}}.$$

Proof. Combining Lemma F.4 with Lemma F.5, we get the result:

$$N \leqslant \frac{16 \log \frac{4SA}{\delta} \log \frac{2SA}{\delta}}{\epsilon^2 p^{\min,H}_{\mathcal{M},\pi_E} \pi'_{\min}}.$$

□

F.3 Proof of Theorem 5.3

The proof of Theorem 5.3 is very similar to that of Theorem 5.2. We begin by proving how the error in the estimation of the policy propagates to the estimation of the centroid:

Lemma F.6. *Assume that* $\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \pi_E(a|s) \geqslant \pi'_{\min}$. *Let* $\widehat{\pi}$ *be the estimate of policy computed in Algorithm 3. Then, it holds that, for all* $(s,a) \in \mathcal{S}\times\mathcal{A}$:

$$|\widehat{r}^{BIRL}(s,a) - r^{BIRL}_{\mathcal{M},\pi_E}(s,a)| \leqslant \frac{2}{\pi'_{\min}}\|\widehat{\pi} - \pi_E\|_\infty + \mathbb{1}\{N(s)=0\}\log\frac{1}{\pi'_{\min}}.$$

Proof. For all the states $s \in \mathcal{S}$ with $N(s) > 0$, for all $a \in \mathcal{A}$, we can write:

$$\begin{aligned}
|\widehat{r}^{\text{BIRL}}(s,a) - r^{\text{BIRL}}_{\mathcal{M},\pi_E}(s,a)| &= \Big|\log\frac{\widehat{\pi}(a|s)}{\max_{a'}\widehat{\pi}(a'|s)} - \log\frac{\pi_E(a|s)}{\max_{a'}\pi_E(a'|s)}\Big| \\
&\leqslant \Big|\log\frac{\widehat{\pi}(a|s)}{\pi_E(a|s)}\Big| + \Big|\log\frac{\max_{a'}\widehat{\pi}(a'|s)}{\max_{a'}\pi_E(a'|s)}\Big| \\
&\overset{(1)}{\leqslant} 2\max_{a'\in\mathcal{A}}\Big|\log\frac{\widehat{\pi}(a'|s)}{\pi_E(a'|s)}\Big| \\
&\overset{(2)}{\leqslant} 2\max_{a'\in\mathcal{A}}\frac{\big|\widehat{\pi}(a'|s) - \pi_E(a'|s)\big|}{\min\{\widehat{\pi}(a'|s), \pi_E(a'|s)\}} \\
&\overset{(3)}{\leqslant} \frac{2}{\pi'_{\min}}\max_{a'\in\mathcal{A}}\big|\widehat{\pi}(a'|s) - \pi_E(a'|s)\big|,
\end{aligned}$$

where at (1) we use that $|\max_x f(x) - \max_x g(x)| \leqslant \max_x |f(x) - g(x)|$, at (2) we use that $\log(1+x) \leqslant x$ for $x \geqslant 0$. Indeed, when $\widehat{\pi}(a|s) \geqslant \pi_E(a|s)$, then $\widehat{\pi}(a|s)/\pi_E(a|s) \geqslant 1$, and so: $\log\frac{\widehat{\pi}(a|s)}{\pi_E(a|s)} = \log\Big(1 + \frac{\widehat{\pi}(a|s)}{\pi_E(a|s)} - 1\Big) \leqslant (\widehat{\pi}(a|s) - \pi_E(a|s))/\pi_E(a|s)$. Doing similarly in the opposite case, we get the minimum at denominator. At (2) we use that $\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \pi_E(a|s) \geqslant \pi'_{\min}$ by hypothesis and the definition of $\widehat{\pi}$.

Similarly, for all the remaining states $s \in \mathcal{S}$ satisfying $N(s) = 0$, for all $a \in \mathcal{A}$, we can write:

$$\begin{aligned}
|\widehat{r}^{\text{BIRL}}(s,a) - r^{\text{BIRL}}_{\mathcal{M},\pi_E}(s,a)| &= \Big|\log\pi'_{\min} - \log\frac{\pi_E(a|s)}{\max_{a'}\pi_E(a'|s)}\Big| \\
&\leqslant \log\frac{1}{\pi'_{\min}} + \Big|\log\frac{\pi_E(a|s)}{\max_{a'}\pi_E(a'|s)}\Big| \\
&\overset{(4)}{=} \log\frac{1}{\pi'_{\min}} + \Big|\log\frac{\widehat{\pi}(a|s)}{\max_{a'}\widehat{\pi}(a'|s)} - \log\frac{\pi_E(a|s)}{\max_{a'}\pi_E(a'|s)}\Big|,
\end{aligned}$$

where at (4) we use that $\log(\widehat{\pi}(a|s)/\max_{a'}\widehat{\pi}(a'|s)) = \log(\pi'_{\min}/\pi'_{\min}) = 0$ when $N(s) = 0$.

Then, carrying out the same passages as for $N(s) > 1$, we get the result. □

Now, we can upper bound $\|\widehat{\pi}^E - \pi_E\|_\infty$.

Lemma F.7. *Let* $\epsilon, \delta \in (0,1)$ *and assume that* $\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \pi_E(a|s) \geqslant \pi'_{\min}$. *Then, it holds that:*

$$\mathbb{P}_{\mathcal{M},\pi_E}\Big(\|\widehat{\pi} - \pi_E\|_\infty \leqslant \epsilon\Big) \geqslant 1 - \delta,$$

with a number of trajectories:

$$N \leqslant \frac{16\log\frac{4SA}{\delta}\log\frac{2SA}{\delta}}{\epsilon^2 p^{\min,H}_{\mathcal{M},\pi_E}}.$$

Proof. Observe that Algorithms 3 and 2 estimate π_E in the same manner. Thus, the proof of this lemma is the same as that of Lemma F.5. □

Finally, we can prove the main result:

Theorem 5.3. *Let* $\epsilon, \delta \in (0,1)$ *and* $H \geqslant S$. *Under the conditions of Theorem 4.8, if* $\pi_{\min} \geqslant \pi'_{\min}$, *then, with probability* $1-\delta$, *we have* $\|\widehat{r}^{BIRL} - r^{BIRL}_{\mathcal{M},\pi_E}\|_\infty \leqslant \epsilon$ *provided that:*

$$NH \geqslant \frac{H\log^2\frac{8SA}{\delta}}{p^{\min,H}_{\mathcal{M},\pi_E}} \cdot \frac{33}{\epsilon^2\pi'_{\min}}.$$

Proof. First, observe that, by applying the same reasoning adopted for proving Lemma F.3, it can be shown that, with probability at least $1-\delta$, $\mathbb{1}\{N(s)=0\}=0$ for all $s \in \mathcal{S}$, using a number of trajectories:

$$N \leqslant \frac{\log \frac{S}{\delta}}{p^{\min,H}_{\mathcal{M},\pi_E}},$$

where we also used that $\mathcal{S}_{\mathcal{M},\pi_E} = \mathcal{S}$ by hypothesis.

Then, combining this result with Lemma F.6 and Lemma F.7 through a union bound, we get the result with a number of trajectories bounded by:

$$N \leqslant \frac{32 \log \frac{8SA}{\delta} \log \frac{4SA}{\delta}}{\epsilon^2 p^{\min,H}_{\mathcal{M},\pi_E} \pi'_{\min}} + \frac{\log \frac{2S}{\delta}}{p^{\min,H}_{\mathcal{M},\pi_E}} \leqslant \frac{33 \log \frac{8SA}{\delta} \log \frac{4SA}{\delta}}{\epsilon^2 p^{\min,H}_{\mathcal{M},\pi_E} \pi'_{\min}}.$$

□

F.4 Proof of Proposition 5.4

Proposition 5.4. *For any MDP\R $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, s'_0, p', \gamma')$, cost $c \in \mathbb{R}^{SA}$ and budget $k \geqslant 0$, it holds that:* $|\max_{\pi \in \Pi_{c,k}} V^\pi(s'_0; p', \widehat{r}^m) - \max_{\pi \in \Pi_{c,k}} V^\pi(s'_0; p', r^m_{\mathcal{M},\pi_E})| \leqslant \|\widehat{r}^m - r^m_{\mathcal{M},\pi_E}\|_\infty/(1-\gamma)$.

Proof. We can write:

$$\begin{aligned}\Big|\max_{\pi \in \Pi_{c,k}} V^\pi(s'_0; p', \widehat{r}^m) - \max_{\pi \in \Pi_{c,k}} V^\pi(s'_0; p', r^m_{\mathcal{M},\pi_E})\Big| &\overset{(1)}{\leqslant} \max_{\pi \in \Pi_{c,k}} \Big|V^\pi(s'_0; p', \widehat{r}^m) - V^\pi(s'_0; p', r^m_{\mathcal{M},\pi_E})\Big| \\ &\overset{(2)}{\leqslant} \frac{1}{1-\gamma} \|\widehat{r}^m - r^m_{\mathcal{M},\pi_E}\|_\infty,\end{aligned}$$

where at (1) we use the property of the maximum operator that $|\max_x f(x) - \max_x g(x)| \leqslant \max_x |f(x) - g(x)|$, at (2) we use Holder's inequality after having realized that $V^\pi(s'_0; p', \widehat{r}^m)$ and $V^\pi(s'_0; p', r^m_{\mathcal{M},\pi_E})$ can be written as the dot product of some vector $d \in \mathbb{R}^{SA}$ and $\widehat{r}^m$ or $r^m_{\mathcal{M},\pi_E}$, where d represents the occupancy measure of π in $\mathcal{M}'$ (Puterman, 1994), and, in particular, it sums to $1/(1-\gamma)$. □

G Experimental Details

In this appendix, we provide additional experimental details and simulations that have been dropped from the main paper due to reasons of space. In Appendix G.1, we provide a formal explanation of how we defined and computed $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ in the experiments. In Appendix G.2, we report the results of the experiments conducted for MCE analogous to those provided in Figs. 2 and 3 in Section 6. Next, in Appendix G.3, we provide some additional considerations on the difference between the policies $\pi^m_{\mathcal{M}}$ found through our methods and $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$. Finally, we provide a comparison with directly transferring the expert's policy π_E to the new environment $\mathcal{M}'$ and with a "best-case" approach (Appendix G.4), and we conclude with some additional simulations in Appendix G.5.

Before diving into the presentation of the numerous sections of this appendix, we mention that each experiment has been conducted in just a few seconds on a laptop with processor "AMD Ryzen 5 5500U with Radeon Graphics 2.10 GHz" with 8GB of RAM. Moreover, we remark that each policy $\pi^m_{\mathcal{M}}$ is defined as any policy in:

$$\pi^m_{\mathcal{M},\pi_E,\mathcal{M}'} \in \arg\max_{\pi' \in \Pi_{c,k}} V^{\pi'}(s'_0; p', r^m_{\mathcal{M},\pi_E}),$$

and we mention that, for the simulations involving MCE and BIRL, we set $\pi_{\min} = 10^{-6}$.

G.1 Definition of $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$

We computed the policy $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ as the policy with state-action occupancy measure $d_{\mathcal{M}',\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}}$ in $\mathcal{M}'$ closest to that of the expert $d_{\mathcal{M},\pi_E}$ in $\mathcal{M}$. To measure the distance between the occupancy measures, we adopted the ℓ_1-norm. Formally, we computed $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ by first solving the mathematical program:

$$\min_{d \in \mathbb{R}^{SA}} \|d - d_{\mathcal{M},\pi_E}\|_1$$

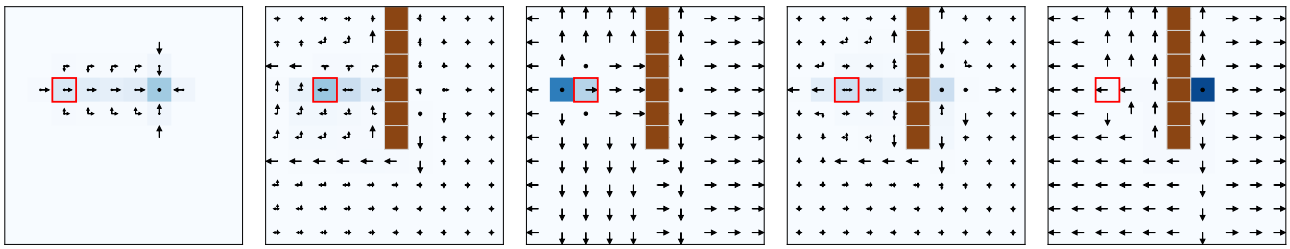

Figure 9: From left to right: (a) π_E in $\mathcal{M}$ with $\gamma = 0.9$, (b)-(c) $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ and $\pi^{\text{MCE}}_{\mathcal{M}}$ in $\mathcal{M}'$ with $\gamma' = 0.7$, (d)-(e) $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ and $\pi^{\text{MCE}}_{\mathcal{M}}$ in $\mathcal{M}'$ with $\gamma' = 0.99$.

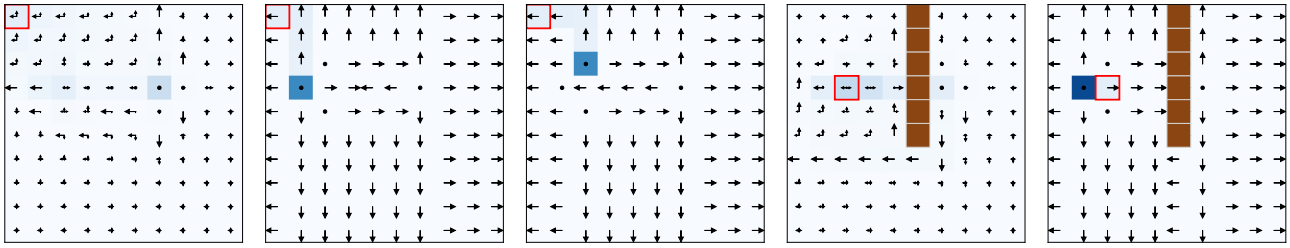

Figure 10: In this series of images, we consider the π_E in Fig. 3 (a), where $\gamma = 0.9$. From left to right: (a) $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ with $\gamma' = 0.9$, (b) $\pi^{\text{BIRL}}_{\mathcal{M}}$ with $\gamma' = 0.9$, (c) $\pi^{\text{MCE}}_{\mathcal{M}}$ with $\gamma' = 0.9$, (d) $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ with constraints and $\gamma' = 0.9$, (e) $\pi^{\text{MCE}}_{\mathcal{M}}$ with constraints and $\gamma' = 0.9$.

$$
\begin{aligned}
\text{s.t.:} \quad & \sum_{a \in \mathcal{A}} d(s,a) = (1-\gamma')\mathbb{1}\{s = s'_0\} + \gamma' \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p'(s|s',a') d(s',a') \qquad \forall s \in \mathcal{S}, \\
& \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d(s,a) c(s,a) \leqslant k, \\
& d(s,a) \geqslant 0 \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \\
& \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d(s,a) = 1,
\end{aligned}
$$

using a Linear Programming (LP) solver,[19] and then we found $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ explicitly from its state-action occupancy measure d as (see (Puterman, 1994)):

$$
\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}(a|s) = \frac{d(s,a)}{\sum_{a' \in \mathcal{A}} d(s,a')}.
$$

G.2 Results for MCE

In Fig. 9, we provide the analogous of the experiments reported in Fig. 2 and 3 for MCE. We remark that the expert's policy π_E in Fig. 9 (a) is the same as the policy used for BIRL, plotted in Fig. 3 (a). Looking at the images, we realize that $\pi^{\text{MCE}}_{\mathcal{M}}$ is quite similar to $\pi^{\text{BIRL}}_{\mathcal{M}}$.

G.3 Additional Considerations on the Difference between $\pi^{m}_{\mathcal{M},\pi_E,\mathcal{M}'}$ and $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$

The difference between our policies $\pi^{m}_{\mathcal{M},\pi_E,\mathcal{M}'}$ and the policy $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ that tries to match the occupancy measure of the expert can be better visualized by looking at Fig. 10.

Let π_E be the expert's policy in Fig. 3 (a), with a support $\mathcal{S}_{\mathcal{M},\pi_E} \subset \mathcal{S}$ strictly. If we consider the problem of generalizing to an MDP\R $\mathcal{M}'$ with an initial state s'_0 falling outside $\mathcal{S}_{\mathcal{M},\pi_E}$, then both the policy $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ (Fig. 10 (a)) and our policies $\pi^{\text{MCE}}_{\mathcal{M}}$ (Fig. 10 (b)) and $\pi^{\text{BIRL}}_{\mathcal{M}}$ (Fig. 10 (c)) try to get back to the support $\mathcal{S}_{\mathcal{M},\pi_E}$. However, they do so differently. As shown in Fig. 10 (a), $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ cares mostly on getting back to the states in the support that have been visited the most by the expert, i.e., those with more intense blue in Fig. 3 (a).

[19] The ℓ_1-norm minimization problem can be recast as a LP problem (Boyd and Vandenberghe, 2004).

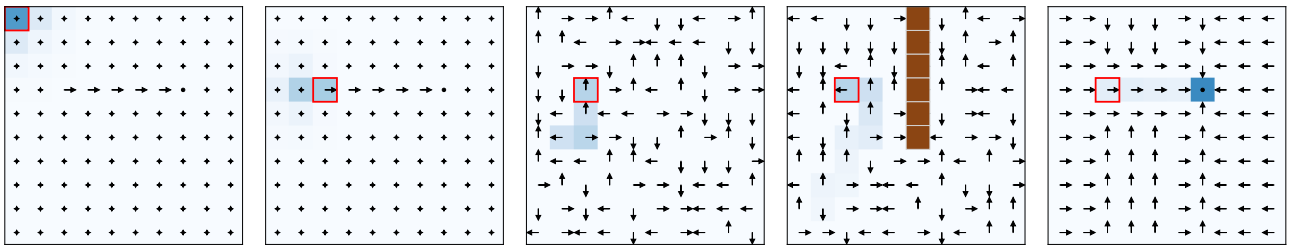

Figure 11: From left to right: (a) $\pi^{\text{BC}}_{\mathcal{M},\pi_E}$ for the expert's policy π_E defined in Fig. 2 (a). We consider $\mathcal{M}'$ with $\gamma' = 0.7$ and $s'_0 \neq s_0$. (b) $\pi^{\text{BC}}_{\mathcal{M},\pi_E}$ for the expert's policy π_E defined in Fig. 2 (a). We consider $\mathcal{M}'$ with $\gamma' = 0.7$ and $s'_0 = s_0$. (c)-(d) Policy π' obtained through the "best-case" approach described in Appendix G.4, respectively in absence and presence of constraints. (e) Illustration of Proposition 4.9.

Instead, both $\pi^{\text{MCE}}_{\mathcal{M}}$ and $\pi^{\text{BIRL}}_{\mathcal{M}}$ are satisfied by an *arbitrary* state in $\mathcal{S}_{\mathcal{M},\pi_E}$, as shown by Fig. 10 (b)-(c). This behavior can be justified mathematically by observing that Algorithms 2 and 3 assign to all the state-action pairs outside $\mathcal{S}_{\mathcal{M},\pi_E}$ the minimum possible reward values, namely:

$$r^{\text{MCE}}_{\mathcal{M},\pi_E}(s,a) = r^{\text{BIRL}}_{\mathcal{M},\pi_E}(s,a) = \log \pi_{\min} \approx -14,$$

where we replaced the value $\pi_{\min} = 10^{-6}$ used in the simulations. As a consequence, $\pi^{\text{MCE}}_{\mathcal{M}}$ and $\pi^{\text{BIRL}}_{\mathcal{M}}$ always try to get back to the states in the support $\mathcal{S}_{\mathcal{M},\pi_E}$, whose reward values is necessarily larger than -14. However, the rewards constructed by Algorithms 2 and 3 do not depend by the probability with which the expert visited the various states in its support, i.e., $d_{\mathcal{M},\pi_E}$, and so, for $\pi^{\text{MCE}}_{\mathcal{M}}$ and $\pi^{\text{BIRL}}_{\mathcal{M}}$, it is irrelevant the state $s \in \mathcal{S}_{\mathcal{M},\pi_E}$ in which they re-play the action/s demonstrated by the expert.

As another example, consider now the presence of constraints in $\mathcal{M}'$ and look at Fig. 10 (d)-(e). Here, even though $\gamma' = 0.9$ is relatively small, $\pi^{\text{MIMIC}}_{\mathcal{M},\pi_E,\mathcal{M}'}$ still tries to go around the obstacle, because its objective is to match the expert's behavior over the whole support $\mathcal{S}_{\mathcal{M},\pi_E}$ (Fig. 10 (d)). Instead, as shown in Fig. 10 (e), $\pi^{\text{MCE}}_{\mathcal{M}}$ is not interested to the states beyond the obstacle if they are too distant (in terms of γ'), but it can be satisfied with states closer to the initial state s'_0. Observe that, if we increase $\gamma' = 0.99$, then also $\pi^{\text{MCE}}_{\mathcal{M}}$ goes around the obstacle, as shown in Fig. 9 (e).

G.4 Comparison with Policy Transfer and a "Best-case" Approach

We now show how generalization to new environments and constraints is carried out by other approaches in this specific problem setting.

Policy transfer. The first "baseline" that we consider here is $\pi^{\text{BC}}_{\mathcal{M},\pi_E}$ (Arora and Doshi, 2021), which is defined as:

$$\pi^{\text{BC}}_{\mathcal{M},\pi_E}(a|s) := \begin{cases} \pi_E(a|s) & \text{if } s \in \mathcal{S}_{\mathcal{M},\pi_E}, \\ \frac{1}{A} & \text{otherwise.} \end{cases}$$

This policy is called "BC" because it represents the result of Behavioral Cloning (Pomerleau, 1988; Osa et al., 2018) on π_E. As shown in Fig. 11 (a)-(b), directly transferring this policy to the new environment $\mathcal{M}'$ might induce quite meaningless behaviors in $\mathcal{M}'$ regardless of whether $s'_0 \neq s_0$ (Fig. 11 (a)) or not (Fig. 11 (b)). Moreover, note that this policy cannot even be applied in general when there are constraints in $\mathcal{M}'$, because it might not satisfy them.

A "best-case" approach. As a second "baseline", we consider the policy π' obtained after carrying out (constrained) planning in $\mathcal{M}'$ with an *arbitrary* reward r in the feasible set $\mathcal{R}^{\text{OPT}}_{\mathcal{M},\pi_E}$:

$$\pi' \in \arg\max_{\pi'' \in \Pi_{c,k}} V^{\pi''}(s'_0; p', r).$$

We have constructed this reward starting from the knowledge of the policy π_E (see Fig. 2 (a)) in its support $\mathcal{S}_{\mathcal{M},\pi_E}$ as follows. First, we have sampled at random a deterministic policy $\overline{\pi}$ from $[\pi_E]_{\mathcal{S}_{\mathcal{M},\pi_E}}$. Next, based on Proposition B.1, we have sampled uniformly at random the values of $V \in [-1,+1]^S$ and $A \in [-0.5,+0.5]^{S(A-1)}$

(the numbers 1 and 0.5 are meaningless), and constructed r as:

$$r(s,a) = V(s) - \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a)V(s') - \begin{cases} 0 & \text{if } a = \overline{\pi}(s), \\ A(s,a) & \text{otherwise.} \end{cases}$$

Note that this approach is representative of all the "best-case" approaches described in Section 3.2. Fig. 11 (c)-(d) show the resulting policy π' in a new environment $\mathcal{M}'$ respectively in absence and presence of constraints. As clear from the images, the policies π' and π_E are rather uncorrelated.

G.5 Other Simulations

In this appendix, we present the result of a simulation aimed to exemplify Proposition 4.9. Specifically, we focused on BIRL and we computed the policy $\pi_{\mathcal{M}}^{\text{BIRL}}$ using $\mathcal{M}' = \mathcal{M}$, i.e., as the policy obtained through our approach in the same environment $\mathcal{M}$ where the expert demonstrated π_E (see Fig. 3 (a)). The result is plotted in Fig. 11 (e). In this setting, $\pi_{\mathcal{M}}^{\text{BIRL}}$ can be interpreted as a "more aggressive" π_E, in the sense that it aims to achieve the same objective as the stochastic expert's policy π_E, but deterministically.