Identifiability and Generalizability in Constrained Inverse Reinforcement Learning

Andreas Schlaginhaufen [1] **Maryam Kamgarpour** [1]

Abstract

Two main challenges in Reinforcement Learning (RL) are designing appropriate reward functions and ensuring the safety of the learned policy. To address these challenges, we present a theoretical framework for Inverse Reinforcement Learning (IRL) in constrained Markov decision processes. From a convex-analytic perspective, we extend prior results on reward identifiability and generalizability to both the constrained setting and a more general class of regularizations. In particular, we show that identifiability up to potential shaping (Cao et al., 2021) is a consequence of entropy regularization and may generally no longer hold for other regularizations or in the presence of safety constraints. We also show that to ensure generalizability to new transition laws and constraints, the true reward must be identified up to a constant. Additionally, we derive a finite sample guarantee for the suboptimality of the learned rewards, and validate our results in a gridworld environment.

1. Introduction

Reinforcement Learning (RL) has been successfully applied to many artificial intelligence tasks such as playing the games of Chess, Go, and Starcraft (Silver et al., 2016), control of humanoid robots (Akkaya et al., 2019), or fine-tuning of large language models (Stiennon et al., 2020). However, two of the key challenges in bringing RL to the real world include designing suitable reward functions for the problem at hand and guaranteeing safety of the learned policy.

While a surge of recent work addresses safe RL, past work has focused on the case in which the reward function is known. In many cases, such as autonomous driving or human-robot interactions, the reward functions are a priori unknown and are hard to design. The framework of inverse reinforcement learning (IRL) addresses learning reward functions based on expert demonstrations. However, most past work on IRL is not incorporating safety explicitly. While safety can be learned implicitly by learning a reward that penalizes unsafe behavior, *incorporating safety in IRL is essential in ensuring that the learned rewards can be transferred to new tasks efficiently*.

Related Work Safe RL for Markov Decision Processes (MDPs) has been extensively studied and various notions of safety have been considered (Garcıa & Fernández, 2015). A well-established framework for safety is Constrained Markov Decision Processes (CMDPs) (Altman, 1999). In this setting, safety is defined through a set of cost functions on the state and actions, whose expectations should be bounded by a predefined threshold. The CMDP approach has been broadly applied to safe RL (Achiam et al., 2017; Chow et al., 2018; Turchetta et al., 2020).

The above safe RL works all assume that the reward is known or can be evaluated along the MDP trajectories. Learning rewards from a data set of expert demonstrations – i.e. the concept of IRL – was first introduced by Russell (1998). Whereas imitation learning (Pomerleau, 1988; Syed & Schapire, 2007; Ho & Ermon, 2016; Garg et al., 2021) attempts to directly recover the expert's policy, the goal of IRL is to learn the expert's latent reward function. The main motivation for IRL is that the reward, being independent of the transition law, provides the most succinct and transferable description of a task (Ng et al., 2000).

A fundamental challenge in IRL is that in general many rewards can generate a given optimal behavior, making it difficult to recover the true underlying reward function. In particular, Ng et al. (1999) show that the set of optimal policies is always invariant under the so-called potential shaping transformations. Additionally, the non-uniqueness of the optimal policy corresponding to some reward can lead to trivial solutions to the IRL problem. For instance, for a constant reward all policies are optimal, but such a reward is most likely not informative for the expert's task. Various methods have been proposed to deal with the above two

[1]SYCAMORE Lab, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland. Correspondence to: Andreas Schlaginhaufen <andreas.schlaginhaufen@epfl.ch>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

degeneracies. These include approaches based on margin maximization (Abbeel & Ng, 2004; Ratliff et al., 2006) or Bayesian reasoning (Ramachandran & Amir, 2007; Choi & Kim, 2012). One promising approach is Maximum Causal Entropy IRL (MCE-IRL) (Ziebart et al., 2010; Zhou et al., 2017), which ensures uniqueness of the optimal policy via entropy regularization and has led to state-of-the-art imitation learning methods (Ho & Ermon, 2016; Garg et al., 2021). MCE-IRL provably recovers the expert's true reward up to potential shaping transformations Cao et al. (2021). Moreover, the MCE-IRL framework has been extended to more general policy regularizations Jeon et al. (2020), but the question of identifiability up to potential shaping transformations remains unanswered in this more general setting.

While both IRL and safe RL problems have been studied extensively, less work has focused on safety aspects in IRL. Robustness to transition laws has been addressed by (Fu et al., 2017; Viano et al., 2021). Motivated by safety and risk considerations, Majumdar et al. (2017) suggest that the expert may not be simply minimizing an expected cumulative cost (corresponding to a constraint), but some general coherent risk measure of that cost. Moreover, Tschiatschek et al. (2019) provide a learner-aware MCE-IRL framework in which the learner has their own preference constraints, such as safety. Most recently, Malik et al. (2021) assume a known reward and address the problem of learning only the constraints from demonstrations, whereas Ding & Xue (2022) consider IRL with combinatorial constraints. These recent works incorporate safety in various approaches but do not focus on identifiability and generalizability aspects.

Contribution Approaching the CMDP problem as a linear program in the so-called occupancy measure, we present a constrained IRL framework for arbitrary convex regularizations of the occupancy measure. In this general setting, we then address the questions of identifiability and generalizability to new transition laws and constraints. In particular, we first show that arbitrary (strictly) convex policy regularizations (Geist et al., 2019) yield a (strictly) convex regularization of the occupancy measure, and are hence naturally incorporated in our framework (Proposition 3.1). Our first main result (Theorem 4.5) is a complete characterization of the set of rewards for which a given expert is optimal. Notably, we show that identifiability up to potential shaping transformations in MCE-IRL is a consequence of entropy regularizations and generally no longer holds for other regularizations (such as e.g. sparse Tsallis entropy (Lee et al., 2018a)), nor in the constrained setting. Our second main result (Theorem 4.12) shows that generalizability to new transition laws and safety constraints is only possible if the expert's reward is recovered up to a constant. Furthermore, we provide a verifiable sufficient condition for generalizability. Our proof techniques, based on convex analysis, unify and generalize past work. In Section 5, we address the finite sample setting and provide a novel result for the number of expert demonstrations needed to recover a reward whose optimal policy is close to the expert's policy. In Section 6, we experimentally verify our results in a gridworld environment.

2. Background

Notation We use $\mathbb{N}, \mathbb{R}$, and $\mathbb{R}_+$ to denote the set of natural, real, and non-negative real numbers, respectively. For an arbitrary set $\mathcal{X}$ we denote $2^{\mathcal{X}}$ for the set of all subsets of $\mathcal{X}$, and for a finite set $\mathcal{Y}$ we denote $\Delta_{\mathcal{Y}}$ for the probability simplex over $\mathcal{Y}$. For two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^l$ we notate $\boldsymbol{a} \leq \boldsymbol{b}$ for the element-wise comparison. Furthermore, we denote $\boldsymbol{I}_l$ for the identity matrix in $\mathbb{R}^l$, we let $\mathbf{1}_l \in \mathbb{R}^l$ be the all-one vector, and $\|\cdot\|$ indicates a general norm on $\mathbb{R}^l$. We also frequently use $\|\cdot\|_p$ with $p \in \{1, 2, \infty\}$ for the p-norms. Given two matrices $\boldsymbol{A}, \boldsymbol{B}$ with compatible dimensions we denote $\begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\end{bmatrix}$ for their concatenation. For a set of vectors $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_q\} \subset \mathbb{R}^l$ we denote $\mathbf{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_q)$ for their linear span, and we denote $\mathbf{span}(\boldsymbol{A})$ for the column span of some matrix $\boldsymbol{A} \in \mathbb{R}^{l \times q}$. Similarly, we use $\mathbf{cone}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_q) := \{\sum_{i=1}^q c_i \boldsymbol{v}_i : c_i \geq 0\}$ to denote the conic hull of $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_q\}$. Moreover, the Minkowski sum of two sets $\mathcal{X}, \mathcal{Y}$ is denoted as $\mathcal{X} + \mathcal{Y} := \{x + y : x \in \mathcal{X}, y \in \mathcal{Y}\}$ and as $x + \mathcal{Y}$ if $\mathcal{X} = \{x\}$. It holds $\mathcal{X} + \emptyset = \emptyset$, as well as, $\mathbf{span}\,\emptyset = \mathbf{cone}\,\emptyset = \mathbf{0}$, where $\mathbf{0}$ is the zero vector. For a set-valued mapping $g : \mathcal{X} \to 2^{\mathcal{Y}}$ we denote by $g(A) := \bigcup_{x \in A} g(x)$ the image of $A \subseteq \mathcal{X}$ under g. We often encounter functions $\boldsymbol{h} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^l$ with finite domain $\mathcal{S} \times \mathcal{A} = \{s_1, \dots, s_n\} \times \{a_1, \dots, a_m\}$. Here, we use the vector notation

$$\begin{aligned}\boldsymbol{h} := \big[&\boldsymbol{h}(s_1, a_1), \boldsymbol{h}(s_2, a_1), \dots \boldsymbol{h}(s_n, a_1), \\ &\boldsymbol{h}(s_1, a_2), \dots, \boldsymbol{h}(s_n, a_2), \dots, \boldsymbol{h}(s_n, a_m)\big]^\top \in \mathbb{R}^{nm \times l}.\end{aligned} \tag{1}$$

Similarly, we identify functions $\mathcal{S} \to \mathbb{R}^l$ and $\mathcal{A} \to \mathbb{R}^l$ with matrices in $\mathbb{R}^{n \times l}$ and $\mathbb{R}^{m \times l}$. The interior $\mathbf{int}\,\mathcal{X}$, the relative interior $\mathbf{relint}\,\mathcal{X}$, the relative boundary $\mathbf{relbd}\,\mathcal{X}$, and the normal cone $N_{\mathcal{X}}(\boldsymbol{x})$ of some set $\mathcal{X} \subseteq \mathbb{R}^l$; as well as the subdifferential $\partial g(\boldsymbol{x})$ of some function $g : \mathcal{X} \to \mathbb{R}$ are for completeness defined in Appendix A.

Constrained Markov Decision Processes We consider CMDPs (Altman, 1999) defined by a tuple $M = (\mathcal{S}, \mathcal{A}, \boldsymbol{P}, \boldsymbol{\nu}_0, \boldsymbol{r}, \boldsymbol{\Psi}, \boldsymbol{b}, \gamma)$. Here, $\mathcal{S}$ and $\mathcal{A}$, with $|\mathcal{S}| = n$ and $|\mathcal{A}| = m > 1$, denote the finite state and action spaces, $\boldsymbol{\nu}_0 \in \Delta_{\mathcal{S}}$ the initial state distribution, $\boldsymbol{P} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ a Markovian transition law, $\boldsymbol{r} \in \mathbb{R}^{nm}$ a reward, and $\gamma \in (0, 1)$ a discount factor. $\boldsymbol{\Psi} := \begin{bmatrix}\boldsymbol{\Psi}_1, \dots, \boldsymbol{\Psi}_k\end{bmatrix} \in \mathbb{R}^{nm \times k}$ is a matrix of safety constraint costs and $\boldsymbol{b} \in \mathbb{R}^k$ is the corresponding threshold. Starting from some initial state $s_0 \sim \boldsymbol{\nu}_0$ the agent can at each step in time t, choose an action $a_t \in \mathcal{A}$, will arrive in some state $s_{t+1} \sim \boldsymbol{P}(\cdot|s_t, a_t)$, and receives reward $\boldsymbol{r}(s_t, a_t)$ and safety cost $\boldsymbol{\Psi}(s_t, a_t)$. The (regularized)

CMDP problem is then defined as follows

$$\max_{\boldsymbol{\pi}\in\Pi} \ (1-\gamma)\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\left[\boldsymbol{r}(s_t,a_t)-\Omega\left(\boldsymbol{\pi}(\cdot|s_t)\right)\right]\right] \quad (2)$$
$$\text{s.t} \quad (1-\gamma)\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\boldsymbol{\Psi}(s_t,a_t)\right]\leq \boldsymbol{b}.$$

Here, $\Pi := \{\tilde{\boldsymbol{\pi}} : \mathcal{S}\to\Delta_{\mathcal{A}}\}$ is the set of Markov policies, $\mathbb{E}_{\boldsymbol{\pi}}$ denotes the expectation with respect to the probability measure $\mathbb{P}^{\boldsymbol{\pi}}_{\boldsymbol{\nu}_0}$, induced by the initial state distribution $\boldsymbol{\nu}_0$ and the policy $a_t\sim\boldsymbol{\pi}(\cdot|s_t)$, on the sample space $(\mathcal{S}\times\mathcal{A})^{\infty} := \{(s_0,a_0,s_1,a_1,\dots) : s_i\in\mathcal{S}, a_i\in\mathcal{A},\ i\in\mathbb{N}\}$. The factor $(1-\gamma)$ is introduced for convenience. Furthermore, $\Omega : \Delta_{\mathcal{A}}\to\mathbb{R}$ is a convex regularization, which if strictly convex ensures uniqueness of the optimal policy (Geist et al., 2019).

A widely used regularization is the negative Shannon entropy[1] $\Omega = -\beta H$, with $\beta>0$ and

$$H:\Delta_{\mathcal{A}}\to\mathbb{R}_+,\ \boldsymbol{d}\mapsto H(\boldsymbol{d}) = -\sum_a \boldsymbol{d}(a)\log\boldsymbol{d}(a). \quad (3)$$

For entropy regularization, the optimal policy can (under Assumption 3.2) be shown to be non-vanishing (see Appendix B.1). Thus, it is often used to foster exploration during optimization (Neu et al., 2017; Haarnoja et al., 2018). Other regularizations such as the sparse Tsallis entropy (Lee et al., 2018b) lead to more sparse optimal policies. Next, we continue with a convex reformulation of problem (2).

3. Convex Viewpoint

Convex Reformulation While the optimization problem (2) is in general non-convex (Agarwal et al., 2019), it admits a convex reformulation. To see this, we introduce the state-action occupancy measure $\boldsymbol{\mu}^{\boldsymbol{\pi}}\in\Delta_{\mathcal{S}\times\mathcal{A}}$ defined by

$$\boldsymbol{\mu}^{\boldsymbol{\pi}}(s,a) := (1-\gamma)\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{P}^{\boldsymbol{\pi}}_{\boldsymbol{\nu}_0}(s_t=s,a_t=a)\right]. \quad (4)$$

For any function $\boldsymbol{h}:\mathcal{S}\times\mathcal{A}\to\mathbb{R}^l$ this allows us to rewrite $(1-\gamma)\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\boldsymbol{h}(s_t,a_t)\right] = \boldsymbol{h}^{\top}\boldsymbol{\mu}^{\boldsymbol{\pi}}$. As shown by Puterman (1994), the set of valid occupancy measures $\mathcal{M} := \{\boldsymbol{\mu}^{\boldsymbol{\pi}} : \boldsymbol{\pi}\in\Pi\}\subseteq\Delta_{\mathcal{S}\times\mathcal{A}}$ is characterized by the Bellman flow constraints

$$\mathcal{M} = \left\{\boldsymbol{\mu}\in\mathbb{R}^{nm}_+ : (\boldsymbol{E}-\gamma\boldsymbol{P})^{\top}\boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0\right\}, \quad (5)$$

where $\boldsymbol{E} := \begin{bmatrix}\boldsymbol{I}_n & \dots & \boldsymbol{I}_n\end{bmatrix}^{\top}\in\mathbb{R}^{nm\times n}$, $\boldsymbol{P}\in\mathbb{R}^{nm\times n}$ is the transition law in matrix form following our convention introduced in (1), and $(\boldsymbol{E}^{\top}\boldsymbol{\mu})(s) = \sum_a\boldsymbol{\mu}(s,a)$ is the so-called state occupancy measure. We refer to states with zero state occupancy measure as unvisited.

[1] Here, we use the convention $0\log 0 = 0$, which is standard in information theory (Cover, 1999) in order to continuously extend H onto the non-negative orthant.

As stated by Puterman (1994) there is a one-to-one mapping $T:\mathcal{M}\to\Pi, \boldsymbol{\mu}\mapsto\boldsymbol{\pi}^{\boldsymbol{\mu}}$ defined via

$$\boldsymbol{\pi}^{\boldsymbol{\mu}}(a|s) := \begin{cases}\boldsymbol{\mu}(s,a)/(\boldsymbol{E}^{\top}\boldsymbol{\mu})(s) & , (\boldsymbol{E}^{\top}\boldsymbol{\mu})(s)>0\\ 1/|\mathcal{A}| & , \text{ otherwise.}\end{cases} \quad (6)$$

The choice $\boldsymbol{\pi}^{\boldsymbol{\mu}}(a|s) = 1/|\mathcal{A}|$ for unvisited states is arbitrary. Without regularization, the above one-to-one mapping proves equivalence of the CMDP problem (2) to a linear program in the occupancy measure. Our next result shows that for arbitrary (strictly) convex regularizations Ω, the objective of (2) is still (strictly) convex in the occupancy measure.

Proposition 3.1. *Let $f(\boldsymbol{\mu}) = \mathbb{E}_{(s,a)\sim\boldsymbol{\mu}}\left[\Omega\left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s)\right)\right]$.*
(a) If Ω is convex, then so is f.
(b) If Ω is strictly convex, then so is f.

For the proof of Proposition 3.1 we refer to Appendix B.2. To the best of our knowledge, Proposition 3.1 is a novel result that has previously only been shown for special cases such as the Shannon entropy (Ziebart et al., 2010).[2]

Due to the one-to-one mapping T and Proposition 3.1, the regularized CMDP problem (2) is equivalent to

$$\max_{\boldsymbol{\mu}\in\mathcal{F}} \ \boldsymbol{r}^{\top}\boldsymbol{\mu} - f(\boldsymbol{\mu}), \quad \text{(P)}$$

where f is defined as in Proposition 3.1 and the set of feasible occupancy measures is given by

$$\mathcal{F} := \left\{\boldsymbol{\mu}\in\mathbb{R}^{nm} : \boldsymbol{\mu}\in\mathcal{M},\ \boldsymbol{\Psi}^{\top}\boldsymbol{\mu}\leq\boldsymbol{b}\right\}\subseteq\Delta_{\mathcal{S}\times\mathcal{A}}. \quad (7)$$

For the rest of the paper, we will be focusing on problem (P). We let $f:\mathcal{X}\to\mathbb{R}$ be an arbitrary convex continuous regularization and $\mathcal{X}\subseteq\mathbb{R}^{nm}$ a closed convex set with $\Delta_{\mathcal{S}\times\mathcal{A}}\subseteq\mathcal{X}$. This formulation includes unregularized CMDPs (Altman, 1999), policy regularization (Geist et al., 2019), as well as many other regularizations such as entropy regularization in the occupancy measure $f(\boldsymbol{\mu}) = -\beta H(\boldsymbol{\mu})$.

If the feasible set $\mathcal{F}$ is non-empty and f is strictly convex, it follows from compactness of $\mathcal{F}$ that problem (P) admits a unique optimal solution. However, we will always state explicitly whether f is assumed to be strictly convex or not. For the further analysis, it will be convenient to define the (set-valued) solution map $\mathsf{RL}_{\mathcal{F}}:\mathbb{R}^{nm}\to 2^{\Delta_{\mathcal{S}\times\mathcal{A}}}$ via

$$\mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}) := \operatorname*{argmax}_{\boldsymbol{\mu}\in\mathcal{F}} \boldsymbol{r}^{\top}\boldsymbol{\mu} - f(\boldsymbol{\mu}), \quad (8)$$

and analogously $\mathsf{RL}_{\mathcal{M}}$ for the unconstrained MDP problem.

Strong Duality Next, we show that for strictly convex regularization, the CMDP problem (P) is equivalent to an unconstrained MDP problem with a modified reward. To see this, we consider the Lagrangian dual problem of (P) obtained via relaxation of the safety constraint

$$\min_{\boldsymbol{\xi}\geq\boldsymbol{0}}\max_{\boldsymbol{\mu}\in\mathcal{M}} \boldsymbol{r}^{\top}\boldsymbol{\mu} - f(\boldsymbol{\mu}) + \boldsymbol{\xi}^{\top}\left(\boldsymbol{b}-\boldsymbol{\Psi}^{\top}\boldsymbol{\mu}\right). \quad \text{(D)}$$

[2] Jeon et al. (2020) mention that strict convexity of a policy regularizer is not always guaranteeing strict convexity in the occupancy measure. However, Proposition 3.1 shows the contrary.

Assumption 3.2 (Slater's condition)**.** Let $\mathbf{relint}\,\mathcal{F} \neq \emptyset$[3].

Assumption 3.3 (Strict convexity)**.** Let $f : \mathcal{X} \to \mathbb{R}$ be strictly convex.

Under Assumption 3.2 above, the optimal values of (P) and (D) coincide and the dual optimum is attained (Altman, 1999). Under the additional assumption of strict convexity, we show that the CMDP problem (P) is equivalent to an unconstrained MDP problem.

Proposition 3.4 (Strong duality)**.** *If Assumption 3.2 and 3.3 hold, the dual optimum of* (D) *is attained for some* $\boldsymbol{\xi}^* \geq \mathbf{0}$*, and* (P) *is equivalent to an unconstrained MDP problem of reward* $\boldsymbol{r} - \boldsymbol{\Psi}\boldsymbol{\xi}^*$*. In other words, it holds*

$$\mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}) = \mathsf{RL}_{\mathcal{M}}(\boldsymbol{r} - \boldsymbol{\Psi}\boldsymbol{\xi}^*). \tag{9}$$

Note that without strict convexity, we have $\mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}) \subseteq \mathsf{RL}_{\mathcal{M}}(\boldsymbol{r} - \boldsymbol{\Psi}\boldsymbol{\xi}^*)$. The proof of Proposition 3.4 and a simple counterexample of why (9) does not hold in the unregularized case are provided in Appendix B.3 and B.4, respectively. As a consequence of (9), we may be tempted to ignore safety constraints for IRL and recover the modified reward $\boldsymbol{r} - \boldsymbol{\Psi}\boldsymbol{\xi}^*$ via standard unconstrained IRL. However, as we discuss before Section 5, such a reward is not guaranteed to generalize to different transition laws and safety constraints.

4. Constrained IRL

Problem Formulation Given a CMDP without reward $M \setminus \boldsymbol{r} = (\mathcal{S}, \mathcal{A}, \boldsymbol{P}, \boldsymbol{\nu}_0, \boldsymbol{\Psi}, \boldsymbol{b}, \gamma)$ and a data set of demonstrations $\mathcal{D} = \left\{(s_t^i, a_t^i)_{t=0}^T\right\}_{i=1}^N$ from some expert $\boldsymbol{\mu}^{\mathrm{E}}$, constrained IRL aims to recover a reward from some reward class $\mathcal{R} \subseteq \mathbb{R}^{nm}$ for which the expert is optimal. Unless stated otherwise, we let the reward class $\mathcal{R}$ be an arbitrary convex set. Clearly, the above IRL problem only has a solution under the following realizability assumption.

Assumption 4.1 (Realizability)**.** Assume the expert is optimal for some $\boldsymbol{r}^{\mathrm{E}} \in \mathcal{R}$ i.e. $\boldsymbol{\mu}^{\mathrm{E}} \in \mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}^{\mathrm{E}})$.

Next, we show that the constrained IRL problem can be formulated as an optimization problem. We first consider an idealized setting, where we are given access to the expert's true occupancy measure $\boldsymbol{\mu}^{\mathrm{E}}$ rather than to the demonstrations $\mathcal{D}$, and address the finite sample setting in Section 5.

Min-Max Formulation It is well-known that in the absence of constraints, the IRL problem can be captured as a min-max optimization problem (Ziebart et al., 2010; Ho & Ermon, 2016). In Proposition 4.2 below we show that the same is true for constrained IRL.

[3]Except for Theorem 4.12, our results continue to hold for the slightly weaker version of Slater's condition $\mathbf{relint}(\mathcal{X}) \cap \mathcal{F} \neq \emptyset$, since the feasible set $\mathcal{F}$ is polyhedral.

Proposition 4.2. *If Assumption 4.1 holds, then the rewards optimizing*

$$\min_{\boldsymbol{r} \in \mathcal{R}} \max_{\boldsymbol{\mu} \in \mathcal{F}} \; \boldsymbol{r}^\top \left(\boldsymbol{\mu} - \boldsymbol{\mu}^E\right) - f(\boldsymbol{\mu}), \tag{IRL}$$

are exactly those rewards in $\mathcal{R}$ *for which the expert occupancy measure is optimal in problem* (P)*.*

The proof of Proposition 4.2 can be found in Appendix C.2. The intuition behind (IRL) is to seek a reward $\boldsymbol{r} \in \mathcal{R}$ for which the suboptimality of the expert is minimized. If Assumption 4.1 fails – that is, $\mathcal{R}$ does not contain a reward for which the expert is optimal – then (IRL) finds a reward for which the expert is least suboptimal. Motivated by Proposition 4.2, we define the (set-valued) IRL solution map $\mathsf{IRL}_{\mathcal{R},\mathcal{F}} : \mathbb{R}^{nm} \to 2^{\mathcal{R}}$ via

$$\mathsf{IRL}_{\mathcal{R},\mathcal{F}}(\boldsymbol{\mu}^{\mathrm{E}}) := \operatorname*{argmin}_{\boldsymbol{r} \in \mathcal{R}} \max_{\boldsymbol{\mu} \in \mathcal{F}} \; \boldsymbol{r}^\top \left(\boldsymbol{\mu} - \boldsymbol{\mu}^{\mathrm{E}}\right) - f(\boldsymbol{\mu}). \tag{10}$$

Additionally, we let $\mathsf{IRL}_{\mathcal{F}} := \mathsf{IRL}_{\mathbb{R}^{nm},\mathcal{F}}$ for the unrestricted reward class $\mathcal{R} = \mathbb{R}^{nm}$. Analogously, $\mathsf{IRL}_{\mathcal{R},\mathcal{M}}$ and $\mathsf{IRL}_{\mathcal{M}}$ are the solution maps for unconstrained IRL. Equipped with the above definitions, we can rewrite Proposition 4.2 as

$$\mathsf{IRL}_{\mathcal{R},\mathcal{F}}(\boldsymbol{\mu}^{\mathrm{E}}) = \left\{\boldsymbol{r} \in \mathcal{R} : \boldsymbol{\mu}^{\mathrm{E}} \in \mathsf{RL}_{\mathcal{F}}(\boldsymbol{r})\right\}. \tag{11}$$

Furthermore, two simple consequences of Proposition 4.2 are summarized in the following corollary.

Corollary 4.3. *If Assumption 4.1 holds, then*

(a) $\mathsf{IRL}_{\mathcal{R},\mathcal{F}}(\boldsymbol{\mu}^E) = \mathsf{IRL}_{\mathcal{F}}(\boldsymbol{\mu}^E) \cap \mathcal{R}$.

If additionally Assumption 3.3 holds, then

(b) $(\mathsf{RL}_{\mathcal{F}} \circ \mathsf{IRL}_{\mathcal{R},\mathcal{F}})(\boldsymbol{\mu}^E) = \left\{\boldsymbol{\mu}^E\right\}$.

Here, (a) states that once we know $\mathsf{IRL}_{\mathcal{F}}(\boldsymbol{\mu}^{\mathrm{E}})$ we can recover $\mathsf{IRL}_{\mathcal{R},\mathcal{F}}(\boldsymbol{\mu}^{\mathrm{E}})$ for any realizable expert via intersection with the reward class $\mathcal{R}$, and (b) shows that if the regularization f is strictly convex, we uniquely recover the expert from the learned reward. In contrast, without strict convexity, there may be trivial solutions to the IRL problem that do not provide any insight into the expert's behavior. For example, let $f = 0$ and $\mathbf{0} \in \mathcal{R}$. Then, all occupancy measures are optimal for $\mathbf{0}$ and hence we have $(\mathsf{RL}_{\mathcal{F}} \circ \mathsf{IRL}_{\mathcal{R},\mathcal{F}})(\boldsymbol{\mu}^{\mathrm{E}}) = \mathcal{F}$ for any expert occupancy measure $\boldsymbol{\mu}^{\mathrm{E}} \in \mathcal{F}$. A popular strictly convex regularization in IRL is the negative Shannon entropy (3) leading to the widely used MCE-IRL algorithm (Ziebart et al., 2010). Moreover, other regularizations considered in the IRL literature are sparse Tsallis entropy (Lee et al., 2018a) or exponential policy regularization (Jeon et al., 2020).

Next, we will explicitly characterize the set of rewards $\mathsf{IRL}_{\mathcal{R},\mathcal{F}}(\boldsymbol{\mu}^{\mathrm{E}})$ that can be recovered via constrained IRL. To this end, we will first consider the unrestricted reward class $\mathcal{R} = \mathbb{R}^{nm}$ and recover $\mathsf{IRL}_{\mathcal{R},\mathcal{F}}(\boldsymbol{\mu}^{\mathrm{E}})$ via Corollary 4.3(b).

Identifiability Trivially, the optimal occupancy measure in (P) is invariant to constant shifts of the reward. Furthermore, Ng et al. (1999) show that if the reward is allowed to depend on the consecutive state $s' \sim \boldsymbol{P}(\cdot|s,a)$, the so-called *potential shaping transformations* $\bar{\boldsymbol{r}}(s,a,s') \mapsto \bar{\boldsymbol{r}}(s,a,s') + \boldsymbol{\eta}(s) - \gamma\boldsymbol{\eta}(s')$ leave the optimal occupancy measure in (P) invariant. Using the conversion $\boldsymbol{r}(s,a) := \mathbb{E}_{s'\sim\boldsymbol{P}(\cdot|s,a)}\bar{\boldsymbol{r}}(s,a,s')$, potential shaping reduces in our setting to $\boldsymbol{r} \mapsto \boldsymbol{r} + \Delta_{\boldsymbol{r}}$ with

$$\begin{aligned}\Delta_{\boldsymbol{r}} \in \mathcal{U} := \big\{\Delta_{\boldsymbol{r}} \in \mathbb{R}^{nm} :\ \Delta_{\boldsymbol{r}}(s,a) = \boldsymbol{\eta}(s) \\ - \gamma\mathbb{E}_{s'\sim\boldsymbol{P}(\cdot|s,a)}\boldsymbol{\eta}(s'),\ \boldsymbol{\eta} \in \mathbb{R}^n\big\} \\ = \mathbf{span}(\boldsymbol{E} - \gamma\boldsymbol{P}).\end{aligned} \tag{12}$$

As a consequence, we expect $\mathsf{IRL}_{\mathcal{F}}(\boldsymbol{\mu}^{\mathrm{E}}) \supseteq \boldsymbol{r}^{\mathrm{E}} + \mathcal{U}$. A recent result by Cao et al. (2021) and Skalse et al. (2022) shows that for standard unconstrained MCE-IRL it holds

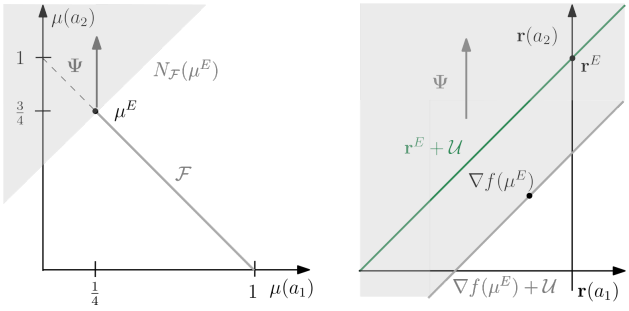
$$\mathsf{IRL}_{\mathcal{M}}(\boldsymbol{\mu}^{\mathrm{E}}) = \boldsymbol{r}^{\mathrm{E}} + \mathcal{U} = \beta \log \boldsymbol{\pi}^{\boldsymbol{\mu}^{\mathrm{E}}} + \mathcal{U}, \tag{13}$$

with $\beta(\log \boldsymbol{\pi}^{\boldsymbol{\mu}^{\mathrm{E}}})(s,a) = \beta \log \boldsymbol{\pi}^{\boldsymbol{\mu}^{\mathrm{E}}}(a|s) = \nabla f(\boldsymbol{\mu}^{\mathrm{E}})(s,a)$. In other words, for $\mathcal{R} = \mathbb{R}^{nm}$ the expert's reward can be identified up to potential shaping in MCE-IRL. Example 4.4 below shows that for MCE-IRL identifiability up to potential shaping is lost when there are active safety constraints[4].

Example 4.4. Consider a single state CMDP with $\mathcal{A} = \{a_1, a_2\}$ and the constraint $\boldsymbol{\mu}(a_2) \leq 3/4$ i.e. $\boldsymbol{\Psi} = \begin{bmatrix}0, 1\end{bmatrix}^\top$ and $b = 3/4$. In this simplified setting it holds $\boldsymbol{\mu}^{\boldsymbol{\pi}} = \boldsymbol{\pi}$, $\mathcal{M} = \Delta_{\mathcal{A}}$, and $\mathcal{U} = \mathbf{span}(\mathbf{1}_2)$. Let the expert $\boldsymbol{\mu}^{\mathrm{E}}$ be optimal for $\boldsymbol{r}^{\mathrm{E}} = [0, 2]^\top$ and the entropy regularization $f(\boldsymbol{\mu}) = \mathbb{E}_{a\sim\boldsymbol{\mu}} \log \boldsymbol{\mu}(a)$. Then, by solving the optimality conditions we can show that $\boldsymbol{\mu}^{\mathrm{E}} = [1/4, 3/4]^\top$ and $\nabla f(\boldsymbol{\mu}^{\mathrm{E}}) \approx [-0.39, 0.71]^\top$. However, as illustrated in Figure 1, we have $\boldsymbol{r}^{\mathrm{E}} \notin \mathsf{IRL}_{\mathcal{M}}(\boldsymbol{\mu}^{\mathrm{E}}) = \nabla f(\boldsymbol{\mu}^{\mathrm{E}}) + \mathcal{U}$. In fact, starting from any $\boldsymbol{r} \in \mathsf{IRL}_{\mathcal{M}}(\boldsymbol{\mu}^{\mathrm{E}})$ we can – due to the active safety constraint – increase $\boldsymbol{r}(a_2)$ without affecting optimality of $\boldsymbol{\mu}^{\mathrm{E}}$. Hence, the expert is optimal for all $\boldsymbol{r} \in \nabla f(\boldsymbol{\mu}^{\mathrm{E}}) + \mathcal{U} + \mathbf{cone}\,\boldsymbol{\Psi}$.

Our main result of this section shows that more generally identifiability up to potential shaping is lost whenever there are active inequality constraints. In particular, under Assumption 3.2 an occupancy measure $\boldsymbol{\mu} \in \mathcal{F}$ is optimal for some reward if and only if this reward is contained in the Minkowski sum of the subdifferential of f and the normal cone to $\mathcal{F}$ at $\boldsymbol{\mu}$. Moreover, the normal cone decomposes into the linear subspace of potential shaping transformation $\mathcal{U}$ and additional conic combinations of the gradients of active inequality constraints. The latter become nonzero when $\boldsymbol{\mu}$ lies on the relative boundary of the feasible set. In this case, we may have $\mathsf{IRL}_{\mathcal{F}}(\boldsymbol{\mu}^{\mathrm{E}}) \supset \boldsymbol{r}^{\mathrm{E}} + \mathcal{U}$ and $\boldsymbol{r}^{\mathrm{E}} \notin \mathsf{IRL}_{\mathcal{M}}(\boldsymbol{\mu}^{\mathrm{E}})$.

[4]We say that an inequality constraint is active if it is satisfied with equality.



(a) Feasible set. (b) Rewards.

Figure 1: (a) illustrates the feasible set $\mathcal{F}$ and the normal cone $N_{\mathcal{F}}(\boldsymbol{\mu}^{\mathrm{E}})$. (b) shows $\mathsf{IRL}_{\mathcal{M}}(\boldsymbol{\mu}^{\mathrm{E}}) = \nabla f(\boldsymbol{\mu}^{\mathrm{E}}) + \mathcal{U}$ and $\mathsf{IRL}_{\mathcal{F}}(\boldsymbol{\mu}^{\mathrm{E}}) = \nabla f(\boldsymbol{\mu}^{\mathrm{E}}) + \mathcal{U} + \mathbf{cone}\,\boldsymbol{\Psi}$.

Theorem 4.5. *Let Assumption 3.2 hold and consider $\boldsymbol{\mu} \in \mathcal{F}$. Let $\mathcal{I}(\boldsymbol{\mu})$ and $\mathcal{J}(\boldsymbol{\mu})$ denote the set of indices of active inequality constraints under $\boldsymbol{\mu}$ i.e. $\boldsymbol{\Psi}_i^\top \boldsymbol{\mu} = \boldsymbol{b}_i$ and $\boldsymbol{\mu}(s,a) = 0$ if and only if $i \in \mathcal{I}(\boldsymbol{\mu})$ and $(s,a) \in \mathcal{J}(\boldsymbol{\mu})$. Then,*

$$\boldsymbol{\mu} \in \mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}) \iff \boldsymbol{r} \in \partial f(\boldsymbol{\mu}) + N_{\mathcal{F}}(\boldsymbol{\mu}), \tag{14}$$

where $N_{\mathcal{F}}(\boldsymbol{\mu}) = \mathcal{U} + \mathcal{C}(\boldsymbol{\mu}) + \mathcal{E}(\boldsymbol{\mu})$ with

$$\begin{aligned}\mathcal{C}(\boldsymbol{\mu}) &:= \mathbf{cone}\left(\{\boldsymbol{\Psi}_i\}_{i\in\mathcal{I}(\boldsymbol{\mu})}\right), \\ \mathcal{E}(\boldsymbol{\mu}) &:= \mathbf{cone}\left(\{-\boldsymbol{e}_{s,a}\}_{(s,a)\in\mathcal{J}(\boldsymbol{\mu})}\right).\end{aligned}$$

Here, $\boldsymbol{e}_{s,a} \in \mathbb{R}^{nm}$ denote the standard unit vectors with $\boldsymbol{e}_{s,a}(s',a') = 1$ if $(s,a) = (s',a')$ and $\boldsymbol{e}_{s,a}(s',a') = 0$ otherwise.

Remark 4.6. Note that in the differentiable case, i.e. if $\partial f(\boldsymbol{\mu}) = \{\nabla f(\boldsymbol{\mu})\}$, the right-hand-side in (14) reduces to the standard first-order optimality condition

$$\nabla h(\boldsymbol{\mu})^\top (\boldsymbol{\mu}' - \boldsymbol{\mu}) \leq 0, \forall \boldsymbol{\mu}' \in \mathcal{F}, \tag{15}$$

for maximization of $h(\boldsymbol{\mu}) = \boldsymbol{r}^\top\boldsymbol{\mu} - f(\boldsymbol{\mu})$ over the set $\mathcal{F}$. Furthermore, for entropy regularization we have $\partial f(\boldsymbol{\mu}) = \emptyset$ for $\boldsymbol{\mu} \in \mathbf{relbd}\,\mathcal{M}$ (see Corollary B.1), which by condition (14) ensures that the optimal occupancy measure $\boldsymbol{\mu}$ lies in the relative interior of $\mathcal{M}$ and hence $\mathcal{E}(\boldsymbol{\mu}) = \mathbf{0}$.

The proof of Theorem 4.5 rests on the optimality conditions for the CMDP problem (P) and is provided in Appendix C.3. Moreover, we provide an extension to state-action-state rewards in Appendix C.4. In light of the above result and Proposition 4.2, the rewards recovered via constrained IRL are characterized as follows:

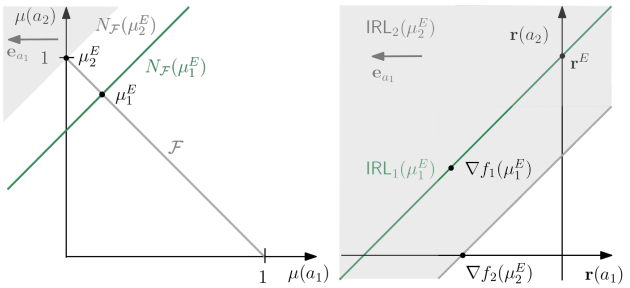
Corollary 4.7. *Let Assumption 3.2 and 4.1 hold. Then,*

$$\mathsf{IRL}_{\mathcal{F}}(\boldsymbol{\mu}^E) = \partial f(\boldsymbol{\mu}^E) + \mathcal{U} + \mathcal{C}(\boldsymbol{\mu}^E) + \mathcal{E}(\boldsymbol{\mu}^E). \tag{16}$$

Corollary 4.7 shows that whenever $\boldsymbol{\mu}^{\mathrm{E}} \in \mathbf{relbd}\,\mathcal{F}$, then $\mathcal{C}(\boldsymbol{\mu}^{\mathrm{E}})$ or $\mathcal{E}(\boldsymbol{\mu}^{\mathrm{E}})$ is nonzero. From this, we observe two

points. First, in the case of active safety constraint, we lose identifiability up to potential shaping i.e. (13) no longer holds. Second, in the case in which the feasible set is $\mathcal{M}$ (no safety constraints), the expert's reward is identifiable up to potential shaping if $\boldsymbol{\mu}^{\mathrm{E}}$ lies in $\mathbf{relint}\,\mathcal{M}$. While the entropy regularization ensures that any optimal occupancy measure lies in $\mathbf{relint}\,\mathcal{M}$, the following example shows that this is not the case for a 2-norm regularization.

Example 4.8. Consider the same MDP as in Example 4.4, but without constraint. Let again $\boldsymbol{r}^{\mathrm{E}} = [0, 2]^\top$, and let $\boldsymbol{\mu}_1^{\mathrm{E}}$ be optimal for $\boldsymbol{r}^{\mathrm{E}}$ with regularization $f_1(\boldsymbol{\mu}) = \mathbb{E}_{a\sim\boldsymbol{\mu}} \log \boldsymbol{\mu}(a)$, and $\boldsymbol{\mu}_2^{\mathrm{E}}$ for $f_2(\boldsymbol{\mu}) = \|\boldsymbol{\mu}\|_2^2/2$. Then, it can be shown that $\boldsymbol{\mu}_1^{\mathrm{E}} = [0.12, 0.88]^\top, \boldsymbol{\mu}_2^{\mathrm{E}} = [0, 1]^\top$ and $\nabla f_1(\boldsymbol{\mu}_1^{\mathrm{E}}) \approx [-1.13, 0.87]^\top, \nabla f_2(\boldsymbol{\mu}_2^{\mathrm{E}}) = [0, 1]^\top$. While for f_1 it holds $\mathsf{IRL}_{1,\mathcal{M}}(\boldsymbol{\mu}_1^{\mathrm{E}}) = \boldsymbol{r}^{\mathrm{E}} + \mathcal{U}$, for f_2 the active non-negativity constraint $\boldsymbol{\mu}_2^{\mathrm{E}}(a_1) \geq 0$ allows us to decrease $\boldsymbol{r}(a_1)$ without affecting optimality of $\boldsymbol{\mu}_2^{\mathrm{E}}$ (see Figure 2). Moreover, according to Theorem 4.4 we have $\mathsf{IRL}_{2,\mathcal{M}}(\boldsymbol{\mu}_2^{\mathrm{E}}) = \nabla f(\boldsymbol{\mu}_2^{\mathrm{E}}) + \mathcal{U} + \mathbf{cone}(-\boldsymbol{e}_{a_1})$.



(a) Feasible set. (b) Rewards.

Figure 2: (a) illustrates the feasible set $\mathcal{F}$ and the normal cones $N_{\mathcal{F}}(\boldsymbol{\mu}_1^{\mathrm{E}}), N_{\mathcal{F}}(\boldsymbol{\mu}_2^{\mathrm{E}})$. (b) shows $\mathsf{IRL}_{1,\mathcal{M}}(\boldsymbol{\mu}_1^{\mathrm{E}})$ under f_1 (in green), and $\mathsf{IRL}_{2,\mathcal{M}}(\boldsymbol{\mu}_2^{\mathrm{E}})$ under f_2 (in gray).

More generally, the optimal occupancy measure is guaranteed to lie in $\mathbf{relint}\,\mathcal{M}$ if the gradient of the regularization becomes unbounded when approaching $\mathbf{relbd}\,\mathcal{M}$. This is formalized in Assumption 4.9 and Corollary 4.10 below.

Assumption 4.9. Let $f : \mathcal{X} \to \mathbb{R}$ be such that:

(a) f is differentiable throughout $\mathbf{int}\,\mathcal{X}$,

(b) $\lim_{k\to\infty} \|\nabla f(\boldsymbol{\mu}_k)\| = \infty$ if $(\boldsymbol{\mu}_k)_{k\in\mathbb{N}}$ is a sequence in $\mathbf{int}\,\mathcal{X}$ converging to a point $\boldsymbol{\mu} \in \mathbf{relbd}\,\mathcal{M}$.

Corollary 4.10. *Let Assumptions 3.2, 4.1, 4.9 hold. Then, we have* $\mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}) \subset \mathbf{relint}\,\mathcal{M}$ *for any* $\boldsymbol{r} \in \mathbb{R}^{nm}$ *and*

$$\mathsf{IRL}_{\mathcal{F}}(\boldsymbol{\mu}^E) = \nabla f(\boldsymbol{\mu}^E) + \mathcal{U} + \mathcal{C}(\boldsymbol{\mu}^E). \tag{17}$$

The proof of Corollary 4.10 is provided in Appendix C.5. Assumption 4.9 is satisfied for the entropy regularization (3) or relative entropy regularization (see Corollary B.1). However, certainly it is not satisfied for many other choices of regularization such as $f = 0$, the sparse Tsallis entropy (Lee et al., 2018b), or the 2-norm regularization.

Generalizability As for our initial goal of learning a reward generalizing to new transition laws and constraints, the result of Corollary 4.7 is problematic, since the set $\mathsf{IRL}_{\mathcal{F}}(\boldsymbol{\mu}^{\mathrm{E}})$ depends on both the transition law and the constraints. To make this dependency explicit, we will throughout this section denote $\mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}$, $\mathsf{IRL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}$ for the solution maps corresponding to the transition law $\boldsymbol{P}$ and constraint threshold $\boldsymbol{b}$. Additionally, we let $\mathfrak{P} \subset \mathbb{R}^{nm\times n}$ be the set of all transition laws. Moreover, we introduce the following notion of generalizability.

Definition 4.11 (Generalizability)**.** Fix some transition law and constraint threshold $(\boldsymbol{P}_0, \boldsymbol{b}_0)$. We say that IRL generalizes to $\mathcal{P} \subseteq \mathfrak{P}$ and $\mathcal{B} \subseteq \mathbb{R}^k$, if

$$\mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}(\boldsymbol{r}) = \mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}(\boldsymbol{r}'),\ \forall \boldsymbol{P} \in \mathcal{P}, \forall \boldsymbol{b} \in \mathcal{B}, \tag{18}$$

for any pair of rewards $\boldsymbol{r}, \boldsymbol{r}' \in \mathsf{IRL}_{\mathcal{R},\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{\mu}^{\mathrm{E}})$.

Definition 4.11 requires all rewards recovered via IRL to yield the same optimal occupancy measures for all $\boldsymbol{P} \in \mathcal{P}$ and $\boldsymbol{b} \in \mathcal{B}$. The subsequent result shows that, under Assumption 4.9, generalization to a neighborhood of new transition laws and arbitrary constraint thresholds is possible if and only if the expert's reward is recovered up to a constant.

Theorem 4.12. *Let Assumptions 3.2, 3.3, 4.1, 4.9 hold for* $(\boldsymbol{P}_0, \boldsymbol{b}_0)$ *and let* $\boldsymbol{\mu}^E \in \mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{r}^E)$ *for some* $\boldsymbol{r}^E \in \mathcal{R}$*. Consider an arbitrary neighborhood* $\mathcal{O}_{\boldsymbol{P}_0} \subseteq \mathbb{R}^{nm\times n}$ *of* $\boldsymbol{P}_0$*. Then, IRL generalizes to* $\mathcal{P} = \mathcal{O}_{\boldsymbol{P}_0} \cap \mathfrak{P}$ *and* $\mathcal{B} = \mathbb{R}^k$ *if and only if*

$$\mathsf{IRL}_{\mathcal{R},\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{\mu}^E) \subseteq \boldsymbol{r}^E + \mathbf{span}(\mathbf{1}_{nm}). \tag{19}$$

Note that since $\mathcal{B} = \mathbb{R}^k$, Theorem 4.12 considers generalizability to all possible constraint thresholds – in particular to the unconstrained setting.[5] The main idea of the proof of Theorem 4.12 is to show that in any neighborhood of $\boldsymbol{P}_0$ we can find $\boldsymbol{P}_1, \boldsymbol{P}_2 \in \mathfrak{P}$ such that only the rewards in $\boldsymbol{r}^{\mathrm{E}} + \mathbf{span}(\mathbf{1}_{nm})$ generalize to both $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$. To the best of our knowledge, this is a novel result – even in the context of unconstrained IRL. For more details about the proof and a brief discussion about the connection to recent results on identifiability from multiple experts (Cao et al., 2021; Rolland et al., 2022), we refer to Appendix C.6.

In light of Theorem 4.12, the following corollary shows that for the unrestricted reward class, IRL is not generalizing to a neighborhood of new transition laws and arbitrary constraints.

Corollary 4.13. *Let Assumption 3.2 and 4.1 hold and let* $n > 1$ *and* $\mathcal{R} = \mathbb{R}^{nm}$*. Moreover, let* $\mathcal{B}$ *and* $\mathcal{P}$ *be defined as in Theorem 4.12. Then, IRL is not generalizing to* $\mathcal{P}$ *and* $\mathcal{B}$*.*

[5] For the unrestricted reward class $\mathcal{R} = \mathbb{R}^{nm}$, we expect the same result to hold even if $\mathcal{B}$ is only a neighborhood of constraint thresholds. However, a rigorous proof would require continuity of $\boldsymbol{P} \mapsto \mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P}}(\boldsymbol{r})$, which we leave open to future work.

Corollary 4.13 is a consequence of $\dim \mathcal{U} = n$, which implies that $\mathsf{IRL}_{\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{\mu}^{\mathrm{E}}) \supset \boldsymbol{r}^{\mathrm{E}} + \mathbf{span}(\mathbf{1}_{nm})$. Below, we show that the above problem can be resolved by a suitable restriction of the reward class. In particular, if the rank condition in Proposition 4.14 holds, then the reward class intersects the space spanned by potential shaping transformations and the safety cost only at the origin, which ensures that the expert's reward can be identified exactly.

Proposition 4.14. *Let Assumptions 3.2, 3.3, 4.1, 4.9 hold. Moreover, let* $\boldsymbol{\mu}^E \in \mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{r}^E)$ *for some* $\boldsymbol{r}^E \in \mathcal{R}$ *and*

$$\mathcal{R} \subseteq \left\{ \boldsymbol{r}_{\boldsymbol{w}} = \boldsymbol{\Phi}\boldsymbol{w} : \boldsymbol{\Phi} \in \mathbb{R}^{mn \times d}, \boldsymbol{w} \in \mathbb{R}^d \right\}. \tag{20}$$

Then, if for $\boldsymbol{\Xi} := \left[\boldsymbol{E} - \gamma \boldsymbol{P}_0, \boldsymbol{\Psi}\right]$ *it holds that*

$$\mathbf{rank}\left[\boldsymbol{\Phi}, \boldsymbol{\Xi}\right] - (\mathbf{rank}\,\boldsymbol{\Phi} + \mathbf{rank}\,\boldsymbol{\Xi}) = 0, \tag{21}$$

then we have $\mathsf{IRL}_{\mathcal{R},\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{\mu}^E) = \left\{\boldsymbol{r}^E\right\}$.

The proof of Proposition 4.14 is provided in Appendix C.7. Observe that for a known transition law, condition (21) can easily be verified.

Ignoring the Constraints As for strictly convex regularization, we have the equivalence $\mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}) = \mathsf{RL}_{\mathcal{M}}(\boldsymbol{r} - \boldsymbol{\Psi}\boldsymbol{\xi}^*)$ (Proposition 3.4), we may ignore safety constraints in IRL and recover the modified reward $\boldsymbol{r} - \boldsymbol{\Psi}\boldsymbol{\xi}^*$ via unconstrained IRL. However, this comes with two caveats. First, the reward class needs to be sufficiently expressive to implicitly account for the safety constraints. Second, as illustrated in Example 4.4, $\mathsf{IRL}_{\mathcal{M}}(\boldsymbol{\mu}^{\mathrm{E}})$ may not contain the expert's reward in case of active safety constraints and hence fail to generalize to the unconstrained setting.

5. Finite Sample Setting

Practical Inverse Reinforcement Learning In practice, we only have access to a finite data set of demonstrations $\mathcal{D} = \left\{(s_t^i, a_t^i)_{t=0}^T\right\}_{i=1}^N$. In this case, the expert occupancy measure $\boldsymbol{\mu}^{\mathrm{E}}$ can be estimated via (Abbeel & Ng, 2004)

$$\hat{\boldsymbol{\mu}}_{\mathcal{D}}^{\mathrm{E}}(s,a) := \frac{(1-\gamma)}{N} \sum_{i=1}^{N} \sum_{t=0}^{T} \gamma^t \mathbb{1}(s_t^i = s, a_t^i = a), \tag{22}$$

where $\mathbb{1}$ is an indicator function. Swapping the order of minimization and maximization using Sion's min-max theorem (Sion, 1958), we can interpret the resulting min-max problem as the dual of an occupancy measure matching problem (Syed & Schapire, 2007; Syed et al., 2008; Ho & Ermon, 2016)

$$\begin{aligned} &\min_{\boldsymbol{r}\in\mathcal{R}} \max_{\boldsymbol{\mu}\in\mathcal{F}} \ \boldsymbol{r}^\top \left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\mathcal{D}}^{\mathrm{E}}\right) - f(\boldsymbol{\mu}) \\ &= -\max_{\boldsymbol{r}\in\mathcal{R}} \min_{\boldsymbol{\mu}\in\mathcal{F}} \ \left[-\boldsymbol{r}^\top \left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\mathcal{D}}^{\mathrm{E}}\right) + f(\boldsymbol{\mu})\right] \\ &= -\min_{\boldsymbol{\mu}\in\mathcal{F}} \left[\delta_{\mathcal{R}}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_{\mathcal{D}}^{\mathrm{E}}) + f(\boldsymbol{\mu})\right]. \end{aligned} \tag{23}$$

Here, $\delta_{\mathcal{R}}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_{\mathcal{D}}^{\mathrm{E}}) := \max_{\boldsymbol{r}\in\mathcal{R}} \boldsymbol{r}^\top \left(\hat{\boldsymbol{\mu}}_{\mathcal{D}}^{\mathrm{E}} - \boldsymbol{\mu}\right)$ is an integral probability metric (Müller, 1997) measuring the distance from $\boldsymbol{\mu}$ to the empirical expert occupancy measure. For different choices of $\mathcal{R}$ different distance measures arise. The choice $\mathcal{R} = \mathbb{R}^{nm}$ yields a characteristic function with $\delta_{\mathcal{R}}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_{\mathcal{D}}^{\mathrm{E}}) = 0$ if $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}_{\mathcal{D}}^{\mathrm{E}}$, and $\delta_{\mathcal{R}}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_{\mathcal{D}}^{\mathrm{E}}) = \infty$ otherwise (Boyd et al., 2004). For the bounded linear feature classes $\mathcal{R}^{\|\cdot\|} := \left\{\boldsymbol{r}_{\boldsymbol{w}} = \boldsymbol{\Phi}\boldsymbol{w} : \boldsymbol{\Phi} \in \mathbb{R}^{nm\times d}, \|\boldsymbol{w}\| \le 1\right\}$ we get $\delta_{\mathcal{R}}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_{\mathcal{D}}^{\mathrm{E}}) = \left\|\boldsymbol{\Phi}^\top(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\mathcal{D}}^{\mathrm{E}})\right\|_*$, where $\|\cdot\|_*$ denotes the dual norm to $\|\cdot\|$. Thus, for $\|\cdot\|_2$ we recover feature expectation matching in the 2-norm (Abbeel & Ng, 2004) and for $\|\cdot\|_1$ in the ∞-norm (Syed et al., 2008). Other choices of $\mathcal{R}$ lead to other distance measures such as the Wasserstein-1 distance or the maximum mean discrepancy (for an overview see (Xiao et al., 2019; Sun et al., 2019; Swamy et al., 2021)). In the following, we focus on the choice

$$\mathcal{R}^{\|\cdot\|_1} := \left\{\boldsymbol{r}_{\boldsymbol{w}} = \boldsymbol{\Phi}\boldsymbol{w} : \boldsymbol{\Phi} \in \mathbb{R}^{nm\times d}, \|\boldsymbol{w}\|_1 \le 1\right\}. \tag{24}$$

We will see that bounding the reward class as above enables us to derive a bound for the sample complexity of IRL.

Sample Complexity The subsequent result shows that if the reward class is bounded, we can bound the suboptimality of solutions obtained by the finite sample problem (23) with respect to the idealized problem (IRL).

Theorem 5.1. *Let Assumption 4.1 hold and* $\boldsymbol{\mu}^E \in \mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}^E)$ *for some* $\boldsymbol{r}^E \in \mathcal{R} := \mathcal{R}^{\|\cdot\|_1}$. *Let* $\hat{\boldsymbol{\mu}} \in \mathsf{RL}_{\mathcal{F}} \circ \mathsf{IRL}_{\mathcal{R},\mathcal{F}}(\hat{\boldsymbol{\mu}}_{\mathcal{D}}^E)$ *and* $R := \max_{s,a} \|\boldsymbol{\Phi}(s,a)\|_\infty$. *Choosing*

$$N = \left\lceil \frac{32R^2}{\varepsilon^2} \log\left(\frac{2d}{\delta}\right) \right\rceil \text{ and } T = \left\lceil \log\left(\frac{\varepsilon}{8R}\right) / \log(\gamma) \right\rceil. \tag{25}$$

It holds with probability at least $1-\delta$

$$J(\boldsymbol{\mu}^E, \boldsymbol{r}^E) - J(\hat{\boldsymbol{\mu}}, \boldsymbol{r}^E) \le \varepsilon, \tag{26}$$

where $J(\boldsymbol{\mu}, \boldsymbol{r}) := \boldsymbol{r}^\top \boldsymbol{\mu} - f(\boldsymbol{\mu})$. *Moreover, if*

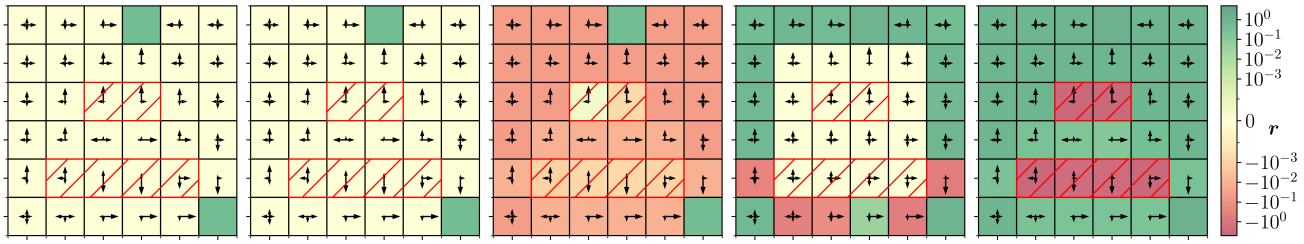
(a) f *is* L*-strongly convex with respect to the norm* $\|\cdot\|$, *it holds with probability at least* $1-\delta$

$$\left\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^E\right\| \le \sqrt{\frac{2\varepsilon}{L}}. \tag{27}$$

(b) $f(\boldsymbol{\mu}) = -\beta\, \mathbb{E}_{(s,a)\sim\boldsymbol{\mu}}\left[H\left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s)\right)\right]$ *with* $\beta > 0$, *it holds with probability at least* $1-\delta$

$$\mathbb{E}_{(s,a)\sim\boldsymbol{\mu}^E}\left[\left\|\boldsymbol{\pi}^{\hat{\boldsymbol{\mu}}}(\cdot|s) - \boldsymbol{\pi}^E(\cdot|s)\right\|_1\right] \le \sqrt{\frac{2\varepsilon}{\beta}}. \tag{28}$$

The first result of Theorem 5.1 shows that by collecting enough expert trajectories with large enough time horizon, constrained IRL recovers with high probability an occupancy measure which is only ε-suboptimal under the expert's reward. A similar result has been shown by Syed &



(*a*) Expert (*b*) $\mathsf{IRL}_{\mathcal{R}_1,\mathcal{F}}$ (*c*) $\mathsf{IRL}_{\mathcal{R}_2,\mathcal{F}}$ (*d*) $\mathsf{IRL}_{\mathcal{R}_1,\mathcal{M}}$ (*e*) $\mathsf{IRL}_{\mathcal{R}_2,\mathcal{M}}$

Figure 3: Comparing rewards and policies recovered via constrained and unconstrained IRL for the two reward classes $\mathcal{R}_1$ and $\mathcal{R}_2$. The color indicates the reward, arrows the policies, and the two red hatched rectangles the constrained states. (*a*) depicts $\boldsymbol{r}^{\mathrm{E}}$ and $\boldsymbol{\pi}^{\mathrm{E}}$, (*b*)-(*c*) the rewards and policies learned from constrained IRL, and (*d*)-(*e*) the rewards and policies learned from unconstrained IRL.

Schapire (2007) for unconstrained unregularized IRL. Second, we show that under strong convexity the recovered occupancy measure is close to the expert's one. Moreover, although the regularization is not strongly convex in MCE-IRL[6], we are still recovering a policy that is close to the expert's policy – at least under the support of the expert occupancy measure. To the best of our knowledge, closeness to the expert occupancy measure (or policy) is novel and only holds in the regularized setting.

6. Experimental Results

Setup To validate our results, we consider a gridworld environment (Sutton & Barto, 2018) with 36 states (the grid cells) and 4 actions (up, down, left, right).[7] The agent has a 90% chance of reaching the desired location when taking an action and a 10% chance of ending up in a random neighboring grid cell. We choose the entropy regularization $f(\boldsymbol{\mu}) = -\mathbb{E}_{(s,a)\sim\boldsymbol{\mu}} H(\boldsymbol{\pi}^{\boldsymbol{\mu}})$, and consider rewards that are only state-dependent, namely, two linear reward classes

$$\mathcal{R}_1 := \left\{\boldsymbol{\Phi}_1 \boldsymbol{w} : \boldsymbol{w} \in \mathbb{R}^{20}\right\} \text{ and } \mathcal{R}_2 := \left\{\boldsymbol{\Phi}_2 \boldsymbol{w} : \boldsymbol{w} \in \mathbb{R}^{36}\right\}. \tag{29}$$

$\mathcal{R}_1$ has a single reward feature for every state on the boundary, and $\mathcal{R}_2$ has reward features for all states. That is, $\boldsymbol{\Phi}_1 = [\boldsymbol{E}_{i_1}, \dots, \boldsymbol{E}_{i_{20}}]$ and $\boldsymbol{\Phi}_2 = \boldsymbol{E}$ and, where $i_1, \dots, i_{20}$ are the indices corresponding to states on the boundary of the gridworld and $\boldsymbol{E}$ is the matrix as defined in (5). The rank condition of Corollary 4.7 is satisfied for the smaller reward class $\mathcal{R}_1$, but not for $\mathcal{R}_2$. The expert's reward $\boldsymbol{r}^{\mathrm{E}}$ is depicted in Figure 3(a). It is zero everywhere except for the two green grid cells where $\boldsymbol{r}^{\mathrm{E}}(s,\cdot) = 0.5$. Furthermore, there are two safety constraints indicated by the red-hatched rectangles. The two rectangular constraints are enforced separately via $\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2$ which are one on the constrained cells and zero everywhere else. The constraint threshold is $\boldsymbol{b}_0 = 0.02 \cdot \mathbf{1}_2$, where feasibility is checked via the LP solver `linprog` provided by (Virtanen et al., 2020).

[6]While the entropy itself is 1-strongly convex in the 1-norm, the resulting regularization f is in general not satisfying this property. However, it is still strictly convex as shown in Proposition 3.1.

[7]The code to all our experiments is available at: https://github.com/andrschl/cirl

Table 1: Comparing generalization of the learned rewards for different constraint thresholds. *Train* indicates the constrained setting with threshold $\boldsymbol{b}_0$ (as used in training), and *test* the generalization to the unconstrained setting (by setting $\boldsymbol{b}_1$ large).

Method	Train ($\boldsymbol{b}_0$)		Test ($\boldsymbol{b}_1 \gg \boldsymbol{b}_0$)	
	$\Delta\boldsymbol{\mu}$	ΔJ	$\Delta\boldsymbol{\mu}$	ΔJ
$\mathsf{IRL}_{\mathcal{R}_1,\mathcal{F}}$	9.6E-9	5.3E-15	1.3E-7	9.2E-14
$\mathsf{IRL}_{\mathcal{R}_2,\mathcal{F}}$	1.7E-6	2.1E-10	2.1E-2	1.7E-3
$\mathsf{IRL}_{\mathcal{R}_1,\mathcal{M}}$	9.5E-2	1.0E-1	2.4E-1	2.9E-1
$\mathsf{IRL}_{\mathcal{R}_2,\mathcal{M}}$	9.3E-3	1.0E-2	2.8E-1	5.9E-1

Algorithm As an algorithm for the min-max problem (IRL) we use a primal-dual gradient-descent-ascent method (Daskalakis & Panageas, 2018) in the policy space (instead of occupancy measure space). In particular, we update the reward parameters and dual variables for the constraints via a (projected) gradient descent step, and the policy via an entropy-regularized natural policy gradient (Cen et al., 2022) step. The algorithm is provided in Appendix E.

Generalizability Learning from the *true expert occupancy measure*, we compare the rewards recovered for constrained vs. unconstrained IRL and $\mathcal{R}_1$ vs. $\mathcal{R}_2$. Figure 3 illustrates the rewards and policies recovered during IRL. Furthermore, Table 1 above summarizes occupancy measure errors and suboptimality for generalization to the same constrained setting as in training ($\boldsymbol{b}_0$) and an unconstrained test setting (with $\boldsymbol{b}_1 \gg \boldsymbol{b}_0$). Here, the occupancy measure error and suboptimality are defined via $\Delta\boldsymbol{\mu} = ||\boldsymbol{\mu}^{\mathrm{E},\boldsymbol{b}} - \boldsymbol{\mu}^{\boldsymbol{b}}||_1$ and $\Delta J = J(\boldsymbol{\mu}^{\mathrm{E},\boldsymbol{b}}, \boldsymbol{r}^{\mathrm{E}}) - J(\boldsymbol{\mu}^{\boldsymbol{b}}, \boldsymbol{r}^{\mathrm{E}})$, with $\boldsymbol{\mu}^{\mathrm{E},\boldsymbol{b}} \in \mathsf{RL}^{\boldsymbol{b}}_{\mathcal{F}}(\boldsymbol{r}^{\mathrm{E}})$ and $\boldsymbol{\mu}^{\boldsymbol{b}} \in \mathsf{RL}^{\boldsymbol{b}}_{\mathcal{F}}(\hat{\boldsymbol{r}})$, where $\hat{\boldsymbol{r}}$ indicates the reward recovered via IRL and $\boldsymbol{b} \in \{\boldsymbol{b}_1, \boldsymbol{b}_2\}$. As depicted in Figures 3(b)-(c) the reward is almost perfectly identified for constrained IRL with the reward class $\mathcal{R}_1$, whereas there is a small mismatch

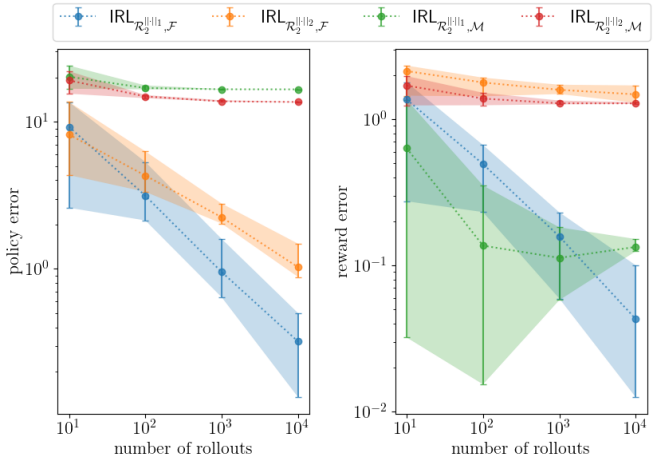


Figure 4: Compare constrained and unconstrained IRL for the two reward classes $\mathcal{R}_2^{\|\cdot\|_1}$ and $\mathcal{R}_2^{\|\cdot\|_2}$ and learning from different amount of expert data. The circles indicate the median, and the shaded areas the 0.1 and 0.9 quantiles for 10 independent realizations of the expert data.

to the expert's reward for $\mathcal{R}_2$. This is in line with our identifiability result in Corollary 4.7, since only $\mathcal{R}_1$ satisfies the rank condition (21). In contrast, the rewards recovered from unconstrained IRL, as shown in Figures 3(d)-(e), substantially deviate from the expert's reward, since they need to implicitly account for the safety constraints. Table 1 shows that constrained IRL with $\mathcal{R}_1$ clearly outperforms the other methods – especially in terms of generalization to the unconstrained setting.

Learning from Expert Data To verify the sample complexity result of Theorem 5.1, we compare constrained and unconstrained IRL for the two reward classes $\mathcal{R}_2^{\|\cdot\|_1}$ and [8]

$$\mathcal{R}_2^{\|\cdot\|_2} := \left\{ \boldsymbol{r}_{\boldsymbol{w}} = \boldsymbol{\Phi}\boldsymbol{w} : \boldsymbol{\Phi} \in \mathbb{R}^{nm \times d}, \|\boldsymbol{w}\|_2 \leq 1/\sqrt{2} \right\}.$$

To this end, we solve the min-max problem (23) for the above reward classes and the feasible sets $\mathcal{F}$ and $\mathcal{M}$. Figure 4 shows the policy and reward errors for 10 independent realizations of the expert demonstrations containing $N \in \{10, 100, 1000, 10000\}$ trajectories of length $T = 10000$. Here, the policy error is $\left\|\boldsymbol{\pi}^{\mathrm{E}} - \hat{\boldsymbol{\pi}}\right\|_1$, where $\hat{\boldsymbol{\pi}}$ is the policy recovered via IRL, and the reward error is defined as the distance of the recovered reward $\hat{\boldsymbol{r}}$ to the line $\boldsymbol{r}^{\mathrm{E}} + \mathbf{span}(\mathbf{1}_{nm})$. As predicted by Theorem 5.1, the policy error is converging towards zero for an increasing number of trajectories. Moreover, the policy error is quite large for unconstrained IRL, which we expect to be due to non-realizability of the expert in this setting (as we need to implicitly account for the constraints). On the other hand, the recovered rewards are – for both constrained and unconstrained IRL – much closer to the expert's reward for the reward class $\mathcal{R}_2^{\|\cdot\|_1}$. We expect the reason for this to be the sparsity induced by the projection onto the 1-norm ball (Tibshirani, 1996), which helps to recover the expert's true reward in this setting. This showcases the importance of the choice of norm when using a bounded linear reward class.

[8] Since in our experiments the recovered reward was always located on the boundary, we choose the bound $1/\sqrt{2}$ here.

7. Limitations and Future Work

For ease of exposition, we limit the scope of this paper to discrete state and action spaces. However, an interesting direction for future research would be to extend our results to the continuous setting, where the CMDP problem can be formulated as an infinite-dimensional convex optimization problem involving the occupancy measure (Altman, 1999). Furthermore, our results are based on optimal solutions and rewards, but in practical settings, we hardly ever obtain optimal solutions and approximately optimal solutions are the norm. Hence, examining identifiability and generalizability in an approximate setting would be valuable for practical applications and may reveal valuable insights on how to choose the regularization f. Finally, Proposition 4.14 provides a sufficient condition for generalizability, but checking the rank condition (21) requires knowledge of the transition law and the constraints. To alleviate this, it may be helpful to learn a reward from multiple experts with different transition laws and constraints.

8. Summary

In this paper, we present a constrained IRL framework for CMDPs with arbitrary convex regularizations of the occupancy measure. From a convex-analytic viewpoint, we address identifiability and generalizability to new transition laws and constraints. Our results indicate that identifiability of rewards up to potential shaping is contingent on the use of entropy regularizations and that generalizability to new transition laws and constraints is only possible when the expert's reward is identified up to a constant. Based on these insights, we provide a sufficient condition for identifiability and generalizability. Furthermore, we show a novel result on the number of expert trajectories required to recover a reward whose optimal policy is close to the expert's policy. Lastly, we showcase the applicability of our results in a gridworld experiment.

Acknowledgements

Andreas Schlaginhaufen is funded by a PhD fellowship from the Swiss Data Science Center.

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.

Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained policy optimization. In *International conference on machine learning*, pp. 22–31. PMLR, 2017.

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Altman, E. *Constrained Markov decision processes: stochastic modeling*. Routledge, 1999.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Cao, H., Cohen, S., and Szpruch, L. Identifiability in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4): 2563–2578, 2022.

Choi, J. and Kim, K.-E. Nonparametric bayesian inverse reinforcement learning for multiple reward functions. *Advances in Neural Information Processing Systems*, 25, 2012.

Chow, Y., Nachum, O., Duenez-Guzman, E., and Ghavamzadeh, M. A lyapunov-based approach to safe reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Cover, T. M. *Elements of information theory*. John Wiley & Sons, 1999.

Daskalakis, C. and Panageas, I. The limit points of (optimistic) gradient descent in min-max optimization. *Advances in neural information processing systems*, 31, 2018.

Ding, D., Zhang, K., Duan, J., Başar, T., and Jovanović, M. R. Convergence and sample complexity of natural policy gradient primal-dual methods for constrained mdps. *arXiv preprint arXiv:2206.02346*, 2022.

Ding, F. and Xue, Y. X-men: guaranteed xor-maximum entropy constrained inverse reinforcement learning. In *Uncertainty in Artificial Intelligence*, pp. 589–598. PMLR, 2022.

Ding, J. Perturbation analysis for the projection of a point to an affine set. *Linear Algebra and its applications*, 191: 199–212, 1993.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pp. 272–279, 2008.

Fu, J., Luo, K., and Levine, S. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

Garcıa, J. and Fernández, F. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Garg, D., Chakraborty, S., Cundy, C., Song, J., and Ermon, S. Iq-learn: Inverse soft-q learning for imitation. *Advances in Neural Information Processing Systems*, 34: 4028–4039, 2021.

Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169. PMLR, 2019.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018.

Halmos, P. R. *Finite-dimensional vector spaces*. Courier Dover Publications, 2017.

Ho, J. and Ermon, S. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Jeon, W., Su, C.-Y., Barde, P., Doan, T., Nowrouzezahrai, D., and Pineau, J. Regularized inverse reinforcement learning. *arXiv preprint arXiv:2010.03691*, 2020.

Lee, K., Choi, S., and Oh, S. Maximum causal tsallis entropy imitation learning. *Advances in neural information processing systems*, 31, 2018a.

Lee, K., Choi, S., and Oh, S. Sparse markov decision processes with causal sparse tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, 3(3):1466–1473, 2018b.

Majumdar, A., Singh, S., Mandlekar, A., and Pavone, M. Risk-sensitive inverse reinforcement learning via coherent risk models. In *Robotics: Science and Systems*, volume 16, pp. 117, 2017.

Malik, S., Anwar, U., Aghasi, A., and Ahmed, A. Inverse constrained reinforcement learning. In *International Conference on Machine Learning*, pp. 7390–7399. PMLR, 2021.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pp. 6820–6829. PMLR, 2020.

Müller, A. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29 (2):429–443, 1997.

Nemirovski, A. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1): 229–251, 2004.

Neu, G., Jonsson, A., and Gómez, V. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Ng, A. Y., Harada, D., and Russell, S. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287, 1999.

Ng, A. Y., Russell, S. J., et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.

Ong, F. and Lustig, M. Sigpy: a python package for high performance iterative reconstruction. In *Proceedings of the ISMRM 27th Annual Meeting, Montreal, Quebec, Canada*, volume 4819, pp. 5, 2019.

Penrose, R. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pp. 406–413. Cambridge University Press, 1955.

Pomerleau, D. A. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1, 1988.

Puterman, M. L. Markov decision processes: Discrete stochastic dynamic programming, 1994.

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

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

Rockafellar, R. T. *Convex analysis*, volume 18. Princeton university press, 1970.

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

Rolland, P., Viano, L., Schürhoff, N., Nikolov, B., and Cevher, V. Identifiability and generalizability from multiple experts in inverse reinforcement learning. *arXiv preprint arXiv:2209.10974*, 2022.

Russell, S. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 101–103, 1998.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Sion, M. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

Skalse, J., Farrugia-Roberts, M., Russell, S., Abate, A., and Gleave, A. Invariance in policy optimisation and partial identifiability in reward learning. *arXiv preprint arXiv:2203.07475*, 2022.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Sun, W., Vemula, A., Boots, B., and Bagnell, D. Provably efficient imitation learning from observation alone. In *International conference on machine learning*, pp. 6036–6045. PMLR, 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Swamy, G., Choudhury, S., Bagnell, J. A., and Wu, S. Of moments and matching: A game-theoretic framework for closing the imitation gap. In *International Conference on Machine Learning*, pp. 10022–10032. PMLR, 2021.

Syed, U. and Schapire, R. E. A game-theoretic approach to apprenticeship learning. *Advances in neural information processing systems*, 20, 2007.

Syed, U., Bowling, M., and Schapire, R. E. Apprenticeship learning using linear programming. In *Proceedings of the 25th international conference on Machine learning*, pp. 1032–1039, 2008.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Tschiatschek, S., Ghosh, A., Haug, L., Devidze, R., and Singla, A. Learner-aware teaching: Inverse reinforcement learning with preferences and constraints. *Advances in neural information processing systems*, 32, 2019.

Turchetta, M., Kolobov, A., Shah, S., Krause, A., and Agarwal, A. Safe reinforcement learning via curriculum induction. *Advances in Neural Information Processing Systems*, 33:12151–12162, 2020.

Viano, L., Huang, Y.-T., Kamalaruban, P., Weller, A., and Cevher, V. Robust inverse reinforcement learning under transition dynamics mismatch. *Advances in Neural Information Processing Systems*, 34, 2021.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Xiao, H., Herman, M., Wagner, J., Ziesche, S., Etesami, J., and Linh, T. H. Wasserstein adversarial imitation learning. *arXiv preprint arXiv:1906.08113*, 2019.

Ying, D., Ding, Y., and Lavaei, J. A dual approach to constrained markov decision processes with entropy regularization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1887–1909. PMLR, 2022.

Zeng, S., Li, C., Garcia, A., and Hong, M. Maximum-likelihood inverse reinforcement learning with finite-time guarantees. *arXiv preprint arXiv:2210.01808*, 2022.

Zhou, Z., Bloem, M., and Bambos, N. Infinite time horizon maximum causal entropy inverse reinforcement learning. *IEEE Transactions on Automatic Control*, 63(9):2787–2802, 2017.

Ziebart, B. D., Bagnell, J. A., and Dey, A. K. Modeling interaction via the principle of maximum causal entropy. In *ICML*, 2010.

A. Notation

In the following, we briefly recall a few basic definitions from convex analysis (Rockafellar, 1970; Boyd et al., 2004). To this end, we denote $B(\boldsymbol{x}, r) := \left\{\boldsymbol{x} \in \mathbb{R}^l : \|\boldsymbol{x}\|_2 < r\right\}$ for an open ball of radius r and center $\boldsymbol{x}$.

Definition A.1 (Interior)**.** The interior of a set $\mathcal{X} \subseteq \mathbb{R}^l$ is defined as

$$\mathbf{int}\, \mathcal{X} := \{\boldsymbol{x} \in \mathcal{X} : B(\boldsymbol{x}, r) \subseteq \mathcal{X} \text{ for some } r > 0\}. \tag{30}$$

Definition A.2 (Affine hull)**.** The affine hull of a set $\mathcal{X} \subseteq \mathbb{R}^l$ is defined as

$$\mathbf{aff}\, \mathcal{X} := \{\theta_1 \boldsymbol{x}_1 + \ldots + \theta_k \boldsymbol{x}_k : \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathcal{X}, \theta_1 + \ldots + \theta_k = 1\}. \tag{31}$$

Definition A.3 (Relative interior)**.** The relative interior of a set $\mathcal{X} \subseteq \mathbb{R}^l$ is defined as

$$\mathbf{relint}\, \mathcal{X} := \{\boldsymbol{x} \in \mathcal{X} : B(\boldsymbol{x}, r) \cap \mathbf{aff}\, \mathcal{X} \subseteq \mathcal{X} \text{ for some } r > 0\}. \tag{32}$$

Definition A.4 (Relative boundary)**.** The relative boundary of a closed set $\mathcal{X} \subseteq \mathbb{R}^l$ is defined as

$$\mathbf{relbd}\, \mathcal{X} := \mathcal{X} \setminus \mathbf{relint}\, \mathcal{X}. \tag{33}$$

Definition A.5 (Subdifferential)**.** A subgradient of a convex function $f : \mathcal{X} \to \mathbb{R}$ with $\mathcal{X} \subseteq \mathbb{R}^l$ at some point $\boldsymbol{x} \in \mathcal{X}$ is a vector $\boldsymbol{g} \in \mathbb{R}^l$ such that $f(\tilde{\boldsymbol{x}}) \geq f(\boldsymbol{x}) + \boldsymbol{g}^\top (\tilde{\boldsymbol{x}} - \boldsymbol{x})$ for all $\tilde{\boldsymbol{x}} \in \mathcal{X}$. The subdifferential $\partial f(\boldsymbol{x})$ at $\boldsymbol{x} \in \mathcal{X}$ is the set of all subgradients at $\boldsymbol{x}$.

Definition A.6 (Normal cone)**.** The normal cone $N_{\mathcal{X}}(\boldsymbol{x})$ of a convex set $\mathcal{X} \subseteq \mathbb{R}^l$ at some point $\boldsymbol{x} \in \mathcal{X}$ is the set of all $\boldsymbol{h} \in \mathbb{R}^l$ such that $\boldsymbol{h}^\top (\tilde{\boldsymbol{x}} - \boldsymbol{x}) \leq 0$ for all $\tilde{\boldsymbol{x}} \in \mathcal{X}$.

B. Proofs and Comments for Section 2 and 3

B.1. Entropy Regularization

In their work on regularized MDPs Geist et al. (2019) consider a family of regularized MDPs with the objective

$$\max_{\boldsymbol{\pi} \in \Pi} J(\boldsymbol{\pi}, \boldsymbol{r}), \tag{34}$$

where $J(\boldsymbol{\pi}, \boldsymbol{r}) := \mathbb{E}_{(s,a)\sim\boldsymbol{\mu}^{\boldsymbol{\pi}}} \left[\boldsymbol{r}(s,a) - \Omega(\boldsymbol{\pi}(\cdot|s))\right]$ and $\Omega : \Delta_{\mathcal{A}} \to \mathbb{R}$ is strongly convex. Defining the optimal value and q-value function

$$\boldsymbol{v}^*(s) := \max_{\boldsymbol{\pi} \in \Pi} \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t \left[\boldsymbol{r}(s_t, a_t) - \Omega(\boldsymbol{\pi}(\cdot|s_t))\right] \middle| s_0 = s\right] \tag{35}$$

$$\boldsymbol{q}^*(s,a) := \boldsymbol{r}(s,a) + \gamma \mathbb{E}_{s' \sim \boldsymbol{P}(\cdot|s,a)} \boldsymbol{v}^*(s'), \tag{36}$$

the optimal policy can be shown to be $\boldsymbol{\pi}^*(\cdot|s) = \nabla\Omega^*(\boldsymbol{q}^*(s,\cdot))$, where $\nabla\Omega^*$ is the gradient of the convex conjugate $\Omega^*(\boldsymbol{q}^*(s,\cdot)) := \max_{\boldsymbol{d} \in \Delta_{\mathcal{A}}} \boldsymbol{q}^*(s,\cdot)^\top \boldsymbol{d} - \Omega(\boldsymbol{d})$. For the entropy regularization $\Omega(\boldsymbol{d}) = -\beta H(\boldsymbol{d})$ with $\beta > 0$, the optimal policy can be shown to have the soft-max form

$$\boldsymbol{\pi}^*(a|s) = \frac{\exp\left(\boldsymbol{q}^*(s,a)/\beta\right)}{\sum_{a'} \exp\left(\boldsymbol{q}^*(s,a')/\beta\right)}. \tag{37}$$

Accordingly, entropy regularization forces the optimal policy to always assign a non-zero probability to each action regularizing the optimal policy towards the uniform distribution. Similar to unregularized MDPs, the optimal policy in entropy regularized MDPs can be computed via value or policy iteration (Ziebart et al., 2010; Haarnoja et al., 2018).

In the occupancy measure, entropy regularization in the policy takes the form

$$f(\boldsymbol{\mu}) = -\beta\, \mathbb{E}_{(s,a)\sim\boldsymbol{\mu}} \left[H\left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s)\right)\right] = \beta \sum_{s,a} \boldsymbol{\mu}(s,a) \log\left(\frac{\boldsymbol{\mu}(s,a)}{\sum_{a'} \boldsymbol{\mu}(s,a')}\right). \tag{38}$$

In order to incorporate prior knowledge about the expert's policy, we may also consider the relative entropy regularization

$$f(\boldsymbol{\mu}) = \beta\, \mathbb{E}_{(s,a)\sim\boldsymbol{\mu}} \left[D_{\mathrm{KL}} \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s) || \boldsymbol{\pi}_0(\cdot|s) \right) \right], \tag{39}$$

where $D_{\mathrm{KL}} : \Delta_{\mathcal{A}} \times \Delta_{\mathcal{A}} \to \mathbb{R}_+$ with $D_{\mathrm{KL}}\left(\boldsymbol{p}||\boldsymbol{q}\right) = \sum_a \boldsymbol{p}(a) \log(\boldsymbol{p}(a)/\boldsymbol{q}(a))$ is the KL divergence or relative entropy and $\boldsymbol{\pi}_0 \in \Pi$ is some some reference policy. The following corollary shows that, under Slater's condition, entropy and relative entropy regularization in the policy are indeed satisfying Assumption 4.9. By Corollary 4.10 this implies that the optimal occupancy measure lies in the relative interior of $\mathcal{M}$.

Corollary B.1. *Let Assumption 3.2 hold. Then, the regularizations*

$$\begin{aligned} f_1(\boldsymbol{\mu}) &= -\beta\, \mathbb{E}_{(s,a)\sim\boldsymbol{\mu}} \left[H \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s) \right) \right], \\ f_2(\boldsymbol{\mu}) &= \beta\, \mathbb{E}_{(s,a)\sim\boldsymbol{\mu}} \left[D_{KL} \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s) || \boldsymbol{\pi}_0(\cdot|s) \right) \right], \end{aligned} \tag{40}$$

both satisfy Assumption 4.9.

To prove Corollary B.1 we first provide a formula for the gradients in Proposition B.2 below.

Proposition B.2. *Consider a differentiable policy regularization $\Omega_s : \Delta_{\mathcal{A}} \to \mathbb{R}$ that is additionally allowed to depend on the state s. Let $\boldsymbol{\mu} \in \mathbf{relint}\, \Delta_{\mathcal{S}\times\mathcal{A}}$. For $n > 1$ and $f(\boldsymbol{\mu}) = \mathbb{E}_{(s,a)\sim\boldsymbol{\mu}} \left[\Omega_s \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s) \right) \right]$ we have*

$$\frac{\partial f(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}(s',a')} = \Omega_{s'} \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s') \right) + \nabla \Omega_{s'} \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s') \right)(a') - \sum_a \boldsymbol{\pi}^{\boldsymbol{\mu}}(a|s') \nabla \Omega_{s'} \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s') \right)(a). \tag{41}$$

In particular, for $f_1(\boldsymbol{\mu}) = -\beta\, \mathbb{E}_{(s,a)\sim\boldsymbol{\mu}} \left[H \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s) \right) \right]$ and $f_2(\boldsymbol{\mu}) = \beta\, \mathbb{E}_{(s,a)\sim\boldsymbol{\mu}} \left[D_{KL} \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s) || \boldsymbol{\pi}_0(\cdot|s) \right) \right]$ with $\beta > 0$ and $\boldsymbol{\pi}_0 > \mathbf{0}$, we get the following gradients:

(a) For $n = 1$, we have $\nabla f_1(\boldsymbol{\mu}) = \beta \left(\log \boldsymbol{\mu} + \mathbf{1}_m \right)$ and $\nabla f_2(\boldsymbol{\mu}) = \beta \left(\log \frac{\boldsymbol{\mu}}{\boldsymbol{\pi}_0} + \mathbf{1}_m \right)$.

(b) For $n > 1$, we have $\nabla f_1(\boldsymbol{\mu}) = \beta \log \boldsymbol{\pi}^{\boldsymbol{\mu}}$ and $\nabla f_2(\boldsymbol{\mu}) = \beta \log \frac{\boldsymbol{\pi}^{\boldsymbol{\mu}}}{\boldsymbol{\pi}_0}$, where $\frac{\boldsymbol{\pi}^{\boldsymbol{\mu}}}{\boldsymbol{\pi}_0} := \left[\frac{\boldsymbol{\pi}^{\boldsymbol{\mu}}(a_1|s_1)}{\boldsymbol{\pi}_0(a_1|s_1)}, \ldots, \frac{\boldsymbol{\pi}^{\boldsymbol{\mu}}(a_m|s_n)}{\boldsymbol{\pi}_0(a_m|s_n)} \right]^\top$.

Proof of Proposition B.2. We define the state occupancy measure $\boldsymbol{\nu}(s) := \sum_a \boldsymbol{\mu}(s,a)$. The result then follows by naive differentiation. In particular, by the product rule we have

$$\frac{\partial f(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}(s',a')} = \Omega_{s'} \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s') \right)(a') + \boldsymbol{\nu}(s') \frac{\partial \Omega_{s'} \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s') \right)}{\partial \boldsymbol{\mu}(s',a')}. \tag{42}$$

Furthermore, it holds

$$\begin{aligned} \frac{\partial \Omega_{s'} \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s') \right)}{\partial \boldsymbol{\mu}(s',a')} &= \nabla \Omega_{s'} \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s') \right)^\top \frac{\partial \boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s')}{\partial \boldsymbol{\mu}(s',a')} \\ &= \sum_a \nabla \Omega_{s'} \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s') \right)(a) \frac{\delta_{a,a'} - \boldsymbol{\pi}^{\boldsymbol{\mu}}(a|s')}{\boldsymbol{\nu}(s')} \\ &= \frac{1}{\boldsymbol{\nu}(s')} \left(\nabla \Omega_{s'} \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s') \right)(a') - \sum_a \boldsymbol{\pi}^{\boldsymbol{\mu}}(a|s') \nabla \Omega_{s'} \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s') \right)(a) \right), \end{aligned} \tag{43}$$

where we used that $\boldsymbol{\pi}^{\boldsymbol{\mu}}(a|s) = \boldsymbol{\mu}(s,a)/\boldsymbol{\nu}(s)$ and $\delta_{a,a'}$ denotes the Kronecker delta with $\delta_{a,a'} = 1$ if $a = a'$ and $\delta_{a,a'} = 0$ otherwise. Hence,

$$\frac{\partial \boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s')}{\partial \boldsymbol{\mu}(s',a')}(a) = \frac{\boldsymbol{\nu}(s')\delta_{a,a'} - \boldsymbol{\mu}(s',a)}{\boldsymbol{\nu}(s')^2} = \frac{\delta_{a,a'} - \boldsymbol{\pi}^{\boldsymbol{\mu}}(a|s')}{\boldsymbol{\nu}(s')}. \tag{44}$$

Plugging (43) back into (42) yields

$$\frac{\partial f(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}(s',a')} = \Omega_{s'} \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s') \right)(a') + \nabla \Omega_{s'} \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s') \right)(a') - \sum_a \boldsymbol{\pi}^{\boldsymbol{\mu}}(a|s') \nabla \Omega_{s'} \left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s') \right)(a). \tag{45}$$

Now, for the special cases f_1 and f_2 we have $f_1(\boldsymbol{\mu}) = \mathbb{E}_{(s,a)\sim\boldsymbol{\mu}}\left[\Omega_1\left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s)\right)\right]$ and $f_2(\boldsymbol{\mu}) = \mathbb{E}_{(s,a)\sim\boldsymbol{\mu}}\left[\Omega_{2,s}\left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s)\right)\right]$ for $\Omega_1(\boldsymbol{d}) = -\beta H(\boldsymbol{d})$ and $\Omega_{2,s}(\boldsymbol{d}) = \beta D_{\mathrm{KL}}(\boldsymbol{d}||\boldsymbol{\pi}_0(\cdot|s))$, respectively. Moreover, $\nabla\Omega_1(\boldsymbol{d}) = \beta\left(\log \boldsymbol{d} + \mathbf{1}_m\right)$ and $\nabla\Omega_{2,s}(\boldsymbol{d}) = \beta\left(\log \frac{\boldsymbol{d}}{\boldsymbol{\pi}_0(\cdot|s)} + \mathbf{1}_m\right)$. This proves (a), since for $n = 1$ we have $\boldsymbol{\mu} = \boldsymbol{\pi}^{\boldsymbol{\mu}}$ and $f_i = \Omega_i$ for $i = 1, 2$. Moreover, to show (b) we plug the above gradients of the policy regularizations back into the formula (41) which yields

$$\begin{aligned}\frac{\partial f_1(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}(s',a')} &= \Omega_1\left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s')\right) + \nabla\Omega_1\left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s')\right)(a') - \sum_a \boldsymbol{\pi}^{\boldsymbol{\mu}}(a|s')\nabla\Omega_1\left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s')\right)(a) \\ &= -\beta H(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s')) + \beta\left(\log \boldsymbol{\pi}^{\boldsymbol{\mu}}(a'|s') + 1\right) - \sum_a \boldsymbol{\pi}^{\boldsymbol{\mu}}(a|s')\beta\left(\log \boldsymbol{\pi}^{\boldsymbol{\mu}}(a|s') + 1\right) \\ &= \beta \log \boldsymbol{\pi}^{\boldsymbol{\mu}}(a'|s'),\end{aligned} \tag{46}$$

and

$$\begin{aligned}\frac{\partial f_2(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}(s',a')} &= \Omega_{2,s'}\left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s')\right) + \nabla\Omega_{2,s'}\left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s')\right)(a') - \sum_a \boldsymbol{\pi}^{\boldsymbol{\mu}}(a|s')\nabla\Omega_{2,s'}\left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s')\right)(a) \\ &= \beta D_{\mathrm{KL}}(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s')||\boldsymbol{\pi}_0(\cdot|s')) + \beta\left(\log \frac{\boldsymbol{\pi}^{\boldsymbol{\mu}}(a'|s')}{\boldsymbol{\pi}_0(a'|s')} + 1\right) - \sum_a \boldsymbol{\pi}^{\boldsymbol{\mu}}(a|s')\beta\left(\log \frac{\boldsymbol{\pi}^{\boldsymbol{\mu}}(a|s')}{\boldsymbol{\pi}_0(a|s')} + 1\right) \\ &= \beta \log \frac{\boldsymbol{\pi}^{\boldsymbol{\mu}}(a'|s')}{\boldsymbol{\pi}_0(a'|s')},\end{aligned} \tag{47}$$

as desired. □

Before we can proceed with the proof of Corollary B.1, we need to prove the following proposition showing that under Slater's condition the state occupancy measure can only be zero if the policy assigns zero probability to some state action pair.

Proposition B.3. *Let Assumption 3.2 hold. If $\boldsymbol{\nu}(s) := \sum_a \boldsymbol{\mu}(s,a) = 0$ for some $s \in \mathcal{S}$, then $\boldsymbol{\pi}^{\boldsymbol{\mu}}(a'|s') = 0$ for some $(s',a') \in \mathcal{S}\times\mathcal{A}$.*

Proof. We show the contraposition: if $\boldsymbol{\pi}^{\boldsymbol{\mu}} > \mathbf{0}$, then $\boldsymbol{\nu} > \mathbf{0}$. To this end, let $\boldsymbol{\pi}^{\boldsymbol{\mu}} > \mathbf{0}$ and note that due to Assumption 3.2 (Slater's condition) there is some $\bar{\boldsymbol{\mu}} \in \mathcal{M}$ such that $\bar{\boldsymbol{\mu}} > \mathbf{0}$. Hence for any $s \in \mathcal{S}$ it holds that

$$\bar{\boldsymbol{\nu}}(s) = \sum_a \bar{\boldsymbol{\mu}}(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}_{\boldsymbol{\nu}_0}^{\boldsymbol{\pi}^{\bar{\mu}}}(s_t = s) > 0. \tag{48}$$

Therefore, there must exist some $T \in \mathbb{N}$ such that $\mathbb{P}_{\boldsymbol{\nu}_0}^{\boldsymbol{\pi}^{\bar{\mu}}}(s_T = s) > 0$. Let us fix such a T.

If $T = 0$, then $\mathbb{P}_{\boldsymbol{\nu}_0}^{\boldsymbol{\pi}^{\bar{\mu}}}(s_T = s) = \boldsymbol{\nu}_0(s) > 0$. In particular, this implies that $\boldsymbol{\nu}(s) > 0$.

If $T > 0$, we have

$$\mathbb{P}_{\boldsymbol{\nu}_0}^{\boldsymbol{\pi}^{\bar{\mu}}}(s_T = s) = \sum_{\substack{s_0,\ldots,s_{T-1}\\ a_0,\ldots,a_{T-1}}} \boldsymbol{\nu}_0(s_0)\prod_{t=1}^{T}\boldsymbol{\pi}^{\bar{\boldsymbol{\mu}}}(a_{t-1}|s_{t-1})\boldsymbol{P}(s_t|s_{t-1},a_{t-1}) > 0. \tag{49}$$

This implies that there is at least one path $(s_0, a_0, \ldots, s_T)$ with non-zero probability under $\mathbb{P}_{\boldsymbol{\nu}_0}^{\boldsymbol{\pi}^{\bar{\mu}}}$ i.e.

$$\boldsymbol{\nu}_0(s_0)\prod_{t=1}^{T}\boldsymbol{\pi}^{\bar{\boldsymbol{\mu}}}(a_{t-1}|s_{t-1})\boldsymbol{P}(s_t|s_{t-1},a_{t-1}) > 0. \tag{50}$$

Moreover, the above product remains positive under each non-vanishing policy. This implies that $\mathbb{P}_{\boldsymbol{\nu}_0}^{\boldsymbol{\pi}^{\mu}}(s_T = s) > 0$ and thus $\boldsymbol{\nu}(s) > 0$. Since the above proof holds for each $s \in \mathcal{S}$, we have proven that $\boldsymbol{\nu} > \mathbf{0}$. □

Now, we are ready to prove Corollary B.1.

Proof of Corollary B.1. Both regularizations are differentiable in the relative interior of their domain $\mathbb{R}_+^{nm}$ (with the gradients provided in Proposition B.2). Now, let $(\boldsymbol{\mu}_k)_{k\in\mathbb{N}}$ be a sequence in $\mathbf{relint}\,\mathbb{R}_+^{nm}$ converging to some occupancy measure $\boldsymbol{\mu} \in \mathbf{relbd}\,\mathcal{M}$ i.e. we have $\boldsymbol{\mu}_k(s,a) \to \boldsymbol{\mu}(s,a) = 0$ for some $(s,a) \in \mathcal{S}\times\mathcal{A}$. In case $\boldsymbol{\nu}(s) = \sum_a \boldsymbol{\mu}(s,a) > 0$, this implies that $\boldsymbol{\pi}^{\boldsymbol{\mu}}(a|s) = 0$. Moreover, in case $\boldsymbol{\nu}(s) = 0$, the result of Proposition B.3 implies that we have $\boldsymbol{\pi}^{\boldsymbol{\mu}}(a'|s') = 0$ for some other $(s',a') \in \mathcal{S}\times\mathcal{A}$. Now, since the mapping

$$\boldsymbol{\mu} \mapsto \boldsymbol{\pi}^{\boldsymbol{\mu}}(a|s) = \begin{cases} \boldsymbol{\mu}(s,a)/\boldsymbol{\nu}(s) & ,\boldsymbol{\nu}(s) > 0 \\ 1/|\mathcal{A}| & ,\text{ otherwise,} \end{cases} \tag{51}$$

is for all state-action pairs continuous on $\{\boldsymbol{\mu}\in\mathcal{M} : \boldsymbol{\pi}^{\boldsymbol{\mu}}(a|s) = 0\}$, convergence in occupancy measure $\boldsymbol{\mu}_k(s,a) \to \boldsymbol{\mu}(s,a) = 0$ implies $\boldsymbol{\pi}^{\boldsymbol{\mu}_k}(a'|s') \to \boldsymbol{\pi}^{\boldsymbol{\mu}}(a'|s') = 0$ for some $(s',a') \in \mathcal{S}\times\mathcal{A}$. Therefore, since $|\log(x)| \to \infty$ as $x \to 0$, we have $\lim_{k\to\infty} \|\nabla f_i(\boldsymbol{\mu}_k)\| = \infty$ for $i = 1, 2$. □

Next, we provide the proof of Proposition 3.1 showing that policy regularization is a special case of occupancy measure regularization.

B.2. Proof of Proposition 3.1

Proposition 3.1 *Let* $f(\boldsymbol{\mu}) = \mathbb{E}_{(s,a)\sim\boldsymbol{\mu}}\left[\Omega\left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s)\right)\right]$.

(a) If Ω *is convex, then so is* f.

(b) If Ω *is strictly convex, then so is* f.

Proof. Defining $\boldsymbol{\mu}_s := \left[\boldsymbol{\mu}(s,a_1),\dots,\boldsymbol{\mu}(s,a_m)\right]^\top$ and denoting the all-one vector in $\mathbb{R}^m$ by $\mathbf{1}$ we can rewrite

$$f(\boldsymbol{\mu}) = \sum_{s:\mathbf{1}^\top\boldsymbol{\mu}_s>0} \mathbf{1}^\top\boldsymbol{\mu}_s\,\Omega\left(\frac{\boldsymbol{\mu}_s}{\mathbf{1}^\top\boldsymbol{\mu}_s}\right). \tag{52}$$

To prove (strict) convexity consider $\boldsymbol{\mu},\bar{\boldsymbol{\mu}} \in \mathbb{R}_+^{nm}$ with $\boldsymbol{\mu} \neq \bar{\boldsymbol{\mu}}$. It will be convenient to define the sets $\mathcal{V} := \left\{s\in\mathcal{S} : \mathbf{1}^\top\boldsymbol{\mu}_s > 0\right\}$ and $\mathcal{W} := \left\{s\in\mathcal{S} : \mathbf{1}^\top\bar{\boldsymbol{\mu}}_s > 0\right\}$. Let $\alpha \in (0,1)$ and $\bar\alpha := 1-\alpha$, then it follows from $\mathcal{V}\cup\mathcal{W} = (\mathcal{V}\cap\mathcal{W})\cup(\mathcal{V}\setminus\mathcal{W})\cup(\mathcal{W}\setminus\mathcal{V})$ that

$$\begin{aligned} &f(\alpha\boldsymbol{\mu} + \bar\alpha\bar{\boldsymbol{\mu}}) \\ &= \sum_{s\in\mathcal{V}\cup\mathcal{W}} \mathbf{1}^\top\left(\alpha\boldsymbol{\mu}_s + \bar\alpha\bar{\boldsymbol{\mu}}_s\right)\Omega\left(\frac{\alpha\boldsymbol{\mu}_s + \bar\alpha\bar{\boldsymbol{\mu}}_s}{\mathbf{1}^\top\left(\alpha\boldsymbol{\mu}_s + \bar\alpha\bar{\boldsymbol{\mu}}_s\right)}\right) \\ &= \underbrace{\sum_{s\in\mathcal{V}\cap\mathcal{W}} \mathbf{1}^\top\left(\alpha\boldsymbol{\mu}_s + \bar\alpha\bar{\boldsymbol{\mu}}_s\right)\Omega\left(\frac{\alpha\boldsymbol{\mu}_s + \bar\alpha\bar{\boldsymbol{\mu}}_s}{\mathbf{1}^\top\left(\alpha\boldsymbol{\mu}_s + \bar\alpha\bar{\boldsymbol{\mu}}_s\right)}\right)}_{(\Delta)} + \sum_{s\in\mathcal{V}\setminus\mathcal{W}} \alpha\mathbf{1}^\top\boldsymbol{\mu}_s\,\Omega\left(\frac{\alpha\boldsymbol{\mu}_s}{\alpha\mathbf{1}^\top\boldsymbol{\mu}_s}\right) \\ &+ \sum_{s\in\mathcal{W}\setminus\mathcal{V}} \bar\alpha\mathbf{1}^\top\bar{\boldsymbol{\mu}}_s\,\Omega\left(\frac{\bar\alpha\bar{\boldsymbol{\mu}}_s}{\bar\alpha\mathbf{1}^\top\bar{\boldsymbol{\mu}}_s}\right). \end{aligned} \tag{53}$$

From here on we can use (strict) convexity of Ω to bound (Δ) as follows

$$\begin{aligned} (\Delta) &= \sum_{s\in\mathcal{V}\cap\mathcal{W}} \mathbf{1}^\top\left(\alpha\boldsymbol{\mu}_s + \bar\alpha\bar{\boldsymbol{\mu}}_s\right)\Omega\left(\frac{\alpha\mathbf{1}^\top\boldsymbol{\mu}_s}{\mathbf{1}^\top\left(\alpha\boldsymbol{\mu}_s + \bar\alpha\bar{\boldsymbol{\mu}}_s\right)}\frac{\boldsymbol{\mu}_s}{\mathbf{1}^\top\boldsymbol{\mu}_s} + \frac{\bar\alpha\mathbf{1}^\top\bar{\boldsymbol{\mu}}_s}{\mathbf{1}^\top\left(\alpha\boldsymbol{\mu}_s + \bar\alpha\bar{\boldsymbol{\mu}}_s\right)}\frac{\bar{\boldsymbol{\mu}}_s}{\mathbf{1}^\top\bar{\boldsymbol{\mu}}_s}\right) \\ &\overset{(<)}{\leq} \sum_{s\in\mathcal{V}\cap\mathcal{W}} \mathbf{1}^\top\left(\alpha\boldsymbol{\mu}_s + \bar\alpha\bar{\boldsymbol{\mu}}_s\right)\left(\frac{\alpha\mathbf{1}^\top\boldsymbol{\mu}_s}{\mathbf{1}^\top\left(\alpha\boldsymbol{\mu}_s + \bar\alpha\bar{\boldsymbol{\mu}}_s\right)}\Omega\left(\frac{\boldsymbol{\mu}_s}{\mathbf{1}^\top\boldsymbol{\mu}_s}\right) + \frac{\bar\alpha\mathbf{1}^\top\bar{\boldsymbol{\mu}}_s}{\mathbf{1}^\top\left(\alpha\boldsymbol{\mu}_s + \bar\alpha\bar{\boldsymbol{\mu}}_s\right)}\Omega\left(\frac{\bar{\boldsymbol{\mu}}_s}{\mathbf{1}^\top\bar{\boldsymbol{\mu}}_s}\right)\right) \\ &= \sum_{s\in\mathcal{V}\cap\mathcal{W}} \left(\alpha\mathbf{1}^\top\boldsymbol{\mu}_s\,\Omega\left(\frac{\boldsymbol{\mu}_s}{\mathbf{1}^\top\boldsymbol{\mu}_s}\right) + \bar\alpha\mathbf{1}^\top\bar{\boldsymbol{\mu}}_s\,\Omega\left(\frac{\bar{\boldsymbol{\mu}}_s}{\mathbf{1}^\top\bar{\boldsymbol{\mu}}_s}\right)\right). \end{aligned} \tag{54}$$

Plugging this back into (53) yields (strict) convexity as desired

$$
\begin{aligned}
& f(\alpha\boldsymbol{\mu} + \bar{\alpha}\bar{\boldsymbol{\mu}}) \\
& \overset{(<)}{\leq} \sum_{s\in\mathcal{V}\cap\mathcal{W}} \left(\alpha \mathbf{1}^\top \boldsymbol{\mu}_s \Omega\left(\frac{\boldsymbol{\mu}_s}{\mathbf{1}^\top\boldsymbol{\mu}_s}\right) + \bar{\alpha}\mathbf{1}^\top\bar{\boldsymbol{\mu}}_s\Omega\left(\frac{\bar{\boldsymbol{\mu}}_s}{\mathbf{1}^\top\bar{\boldsymbol{\mu}}_s}\right)\right) + \sum_{s\in\mathcal{V}\setminus\mathcal{W}} \alpha\mathbf{1}^\top\boldsymbol{\mu}_s\Omega\left(\frac{\boldsymbol{\mu}_s}{\mathbf{1}^\top\boldsymbol{\mu}_s}\right) \\
& + \sum_{s\in\mathcal{W}\setminus\mathcal{V}} \bar{\alpha}\mathbf{1}^\top\bar{\boldsymbol{\mu}}_s\Omega\left(\frac{\bar{\boldsymbol{\mu}}_s}{\mathbf{1}^\top\bar{\boldsymbol{\mu}}_s}\right) \\
& = \alpha f(\boldsymbol{\mu}) + \bar{\alpha} f(\bar{\boldsymbol{\mu}}),
\end{aligned} \tag{55}
$$

where we used $\mathcal{V} = (\mathcal{V}\cap\mathcal{W}) \cup (\mathcal{V}\setminus\mathcal{W})$ and $\mathcal{W} = (\mathcal{V}\cap\mathcal{W}) \cup (\mathcal{W}\setminus\mathcal{V})$ in the last equality. □

B.3. Proof of Proposition 3.4

Proposition 3.4 *If Assumption 3.2 and 3.3 hold, the dual optimum of (D) is attained for some $\boldsymbol{\xi}^* \geq \mathbf{0}$, and (P) is equivalent to an unconstrained MDP problem of reward $\boldsymbol{r} - \boldsymbol{\Psi}\boldsymbol{\xi}^*$. In other words, it holds*

$$
\mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}) = \mathsf{RL}_{\mathcal{M}}(\boldsymbol{r} - \boldsymbol{\Psi}\boldsymbol{\xi}^*). \tag{56}
$$

Proof. The proof is based on standard Lagrangian duality theory. First, we note that the CMDP problem (P) is a convex optimization problem. Its primal optimum is finite, as the feasible set $\mathcal{F} \subseteq \Delta_{\mathcal{S}\times\mathcal{A}}$ is bounded and the objective is upper bounded by a linear function (since f is convex). From Slater's condition it follows that strong duality holds and the dual optimum is attained by some not necessarily unique $\boldsymbol{\xi}^* \geq \mathbf{0}$ (Boyd et al., 2004).

To show equation (56), note that the primal optimum $\mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}) = \{\boldsymbol{\mu}^*\}$ is unique due to strict convexity of f. Moreover, for each pair $(\boldsymbol{\mu}^*, \boldsymbol{\xi}^*)$ of primal and dual optimal solutions the Lagrangian

$$
L(\boldsymbol{\mu}, \boldsymbol{\xi}) = \boldsymbol{r}^\top\boldsymbol{\mu} - f(\boldsymbol{\mu}) + \boldsymbol{\xi}^\top\left(\boldsymbol{b} - \boldsymbol{\Psi}^\top\boldsymbol{\mu}\right), \tag{57}
$$

has a saddle point at $(\boldsymbol{\mu}^*, \boldsymbol{\xi}^*)$ i.e.

$$
L(\boldsymbol{\mu}, \boldsymbol{\xi}^*) \leq L(\boldsymbol{\mu}^*, \boldsymbol{\xi}^*) \leq L(\boldsymbol{\mu}^*, \boldsymbol{\xi}), \quad \forall \boldsymbol{\mu} \in \mathcal{M}, \boldsymbol{\xi} \geq \mathbf{0}. \tag{58}
$$

We then have $\mathsf{RL}_{\mathcal{M}}(\boldsymbol{r} - \boldsymbol{\Psi}\boldsymbol{\xi}^*) = \operatorname{argmax}_{\boldsymbol{\mu}\in\mathcal{M}} L(\boldsymbol{\mu}, \boldsymbol{\xi}^*) = \{\boldsymbol{\mu}^*\}$ where we again used strict convexity of f for the last equality. □

B.4. Remarks on Strong Duality

Whereas strong duality holds also for unregularized CMDPs Altman (1999), unique recovery of the optimal occupancy measure from an unconstrained RL problem (56) is a consequence of the strictly convex regularization. To illustrate this, consider the following simple example.

Example B.4. Consider a single state MDP with $\mathcal{A} = \{a_1, a_2\}$. The reward is defined via $\boldsymbol{r} = \left[\boldsymbol{r}(a_1), \boldsymbol{r}(a_2)\right]^\top = \left[0, 1\right]^\top$ and there is no regularization. Furthermore, the agent needs to respect the constraints $\boldsymbol{\Psi}^\top\boldsymbol{\mu} = \left[0, 1\right]^\top\boldsymbol{\mu} = \boldsymbol{\mu}(a_2) \leq 3/4$. In this single state setting $\boldsymbol{\mu}^{\boldsymbol{\pi}}(a) = \boldsymbol{\pi}(a)$ and $\mathcal{M} = \Delta_{\mathcal{A}}$. Clearly, the unique primal optimal solution is $\boldsymbol{\mu}^*(a_1) = 1/4$ and $\boldsymbol{\mu}^*(a_2) = 3/4$. This is a key difference to the unconstrained setting where always a deterministic optimal policy exists. Thus, $\boldsymbol{\mu}^*$ cannot be realized as the unique optimum of an unconstrained, unregularized MDP, but only as the convex combination of multiple deterministic solutions. Indeed relaxing the safety constraint yields the Lagrangian $L(\boldsymbol{\mu}, \xi) = (\boldsymbol{r} - \boldsymbol{\Psi}\xi)^\top\boldsymbol{\mu} + \xi b = (1-\xi)\boldsymbol{\mu}(a_2) + 3\xi/4$ and the dual function

$$
g(\xi) = \max_{\boldsymbol{\mu}\in\Delta_{\mathcal{A}}} L(\boldsymbol{\mu}, \xi) = \begin{cases} 1 - \xi/4 & , \xi \leq 1 \\ 3\xi/4 & , \xi > 1. \end{cases} \tag{59}
$$

Thus, there is a unique dual optimum $\xi^* = \operatorname{argmin}_{\xi\geq 0} g(\xi) = 1$, leading to the dual optimal value $1/2$, which is equal to the primal optimum due to strong duality. However, for the reward $\boldsymbol{r} - \boldsymbol{\Psi}\xi = \left[0, 0\right]^\top$ not only $\boldsymbol{\mu}^*$, but all $\boldsymbol{\mu} \in \Delta_{\mathcal{A}}$ are optimal in the unconstrained problem – even those with $\boldsymbol{\mu}(a_2) > 3/4$ that are primal infeasible.

C. Proofs and Comments of Section 4

C.1. Preliminaries from Convex Analysis

Throughout this section, we introduce a few additional tools from convex analysis which turn out to be useful for the proof of Theorem 4.5. In convex analysis it is standard to extend convex functions over the entire space by setting their value to $+\infty$ outside of their domain. This leads us to extended real value functions $h : \mathbb{R}^n \to [-\infty, \infty]$. Their effective domain is defined as $\mathbf{dom}\, h := \{\boldsymbol{x} : h(\boldsymbol{x}) < \infty\}$, and a convex function h is said to be proper if $h > -\infty$ and $\mathbf{dom}\, h \neq \emptyset$. Furthermore, h is referred to as closed if its epigraph $\{(\boldsymbol{x}, y) : \boldsymbol{x} \in \mathbf{dom}\, h, y \geq h(\boldsymbol{x})\}$ is a closed set. For instance, h is closed if it is continuous and $\mathbf{dom}\, h$ is a closed set (Boyd et al., 2004). Next, we introduce the two key tools needed for the proof of Theorem 4.5 – convex conjugates and the Moreau-Rockafeller theorem.

Convex Conjugate The convex conjugate $h^* : \mathbb{R}^n \to [-\infty, \infty]$ of h is defined as

$$h^*(\boldsymbol{y}) := \sup_{\boldsymbol{x} \in \mathbb{R}^n} \boldsymbol{y}^\top \boldsymbol{x} - h(\boldsymbol{x}). \tag{60}$$

If h is closed proper convex, it holds $h^{**} = h$. Moreover, the following optimality conditions hold.

Theorem C.1 (Rockafellar (1970))**.** *For any proper convex function $h^* : \mathbb{R}^n \to [-\infty, \infty]$ it holds*

$$h^*(\boldsymbol{y}) = \boldsymbol{y}^\top \boldsymbol{x} - h(\boldsymbol{x}) \quad \iff \quad \boldsymbol{y} \in \partial h(\boldsymbol{x}). \tag{61}$$

If additionally h is closed, then

$$h^*(\boldsymbol{y}) = \boldsymbol{y}^\top \boldsymbol{x} - h(\boldsymbol{x}) \quad \iff \quad \boldsymbol{y} \in \partial h(\boldsymbol{x}) \quad \iff \quad \boldsymbol{x} \in \partial h^*(\boldsymbol{y}). \tag{62}$$

Moreau-Rockafeller Theorem The following theorem gives sufficient conditions under which the sum of subdifferentials of two functions is equal to the subdifferential of the sum of the two functions.

Theorem C.2 (Rockafellar (1970))**.** *Let h_1, h_2 be proper convex functions on $\mathbb{R}^n$ and let $h := h_1 + h_2$. If $\mathbf{relint}(\mathbf{dom}\, h_1)$ and $\mathbf{relint}(\mathbf{dom}\, h_2)$, have a point in common then*

$$\partial h(\boldsymbol{x}) = \partial h_1(\boldsymbol{x}) + \partial h_2(\boldsymbol{x}), \forall \boldsymbol{x}. \tag{63}$$

If h_1 is polyhedral (i.e. its epigraph is polyhedral) then it is enough if the sets $\mathbf{dom}\, h_1$ and $\mathbf{relint}(\mathbf{dom}\, h_2)$ have a point in common.

C.2. Proof of Proposition 4.2

Proposition 4.2 *If Assumption 4.1 holds, then the rewards optimizing*

$$\min_{\boldsymbol{r} \in \mathcal{R}} \max_{\boldsymbol{\mu} \in \mathcal{F}} \; \boldsymbol{r}^\top \left(\boldsymbol{\mu} - \boldsymbol{\mu}^E\right) - f(\boldsymbol{\mu}), \tag{IRL}$$

are exactly those rewards in $\mathcal{R}$ for which the expert occupancy measure is optimal in problem (P).

Proof. We can rewrite problem (IRL) equivalently as $\min_{\boldsymbol{r} \in \mathcal{R}} \max_{\boldsymbol{\mu} \in \mathcal{F}} L(\boldsymbol{\mu}, \boldsymbol{r})$, where $L(\boldsymbol{\mu}, \boldsymbol{r}) := J(\boldsymbol{\mu}, \boldsymbol{r}) - J(\boldsymbol{\mu}^{\mathrm{E}}, \boldsymbol{r})$ and $J(\boldsymbol{\mu}, \boldsymbol{r}) := \boldsymbol{r}^\top \boldsymbol{\mu} - f(\boldsymbol{\mu})$. For a fixed $\boldsymbol{r}$ it clearly holds $\operatorname{argmax}_{\boldsymbol{\mu} \in \mathcal{F}} L(\boldsymbol{\mu}, \boldsymbol{r}) = \mathsf{RL}_{\mathcal{F}}(\boldsymbol{r})$. Also, we always get the lower bound $\max_{\boldsymbol{\mu} \in \mathcal{F}} L(\boldsymbol{\mu}, \boldsymbol{r}) \geq 0$. This lower bound is achieved if and only if $\boldsymbol{\mu}^{\mathrm{E}} \in \operatorname{argmax}_{\boldsymbol{\mu} \in \mathcal{F}} L(\boldsymbol{\mu}, \boldsymbol{r}) = \mathsf{RL}_{\mathcal{F}}(\boldsymbol{r})$. By Assumption 4.1, there is indeed $\boldsymbol{r}^{\mathrm{E}} \in \mathcal{R}$ such that $\boldsymbol{\mu}^{\mathrm{E}} \in \mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}^{\mathrm{E}})$, and thus $\max_{\boldsymbol{\mu} \in \mathcal{F}} L(\boldsymbol{\mu}, \boldsymbol{r}^{\mathrm{E}}) = 0$. Therefore, any optimal $r^* \in \mathcal{R}$ must achieve $\max_{\boldsymbol{\mu} \in \mathcal{F}} L(\boldsymbol{\mu}, \boldsymbol{r}^*) = 0$, which implies $\boldsymbol{\mu}^{\mathrm{E}} \in \mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}^*)$. Moreover, for any $\boldsymbol{r}^* \in \mathcal{R}$ with $\boldsymbol{\mu}^{\mathrm{E}} \in \mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}^*)$ it needs to hold $\max_{\boldsymbol{\mu} \in \mathcal{F}} L(\boldsymbol{\mu}, \boldsymbol{r}^*) = 0$, which proves optimality of $\boldsymbol{r}^*$. □

C.3. Proof of Theorem 4.5

Theorem 4.5 *Let Assumption 3.2 hold and consider $\boldsymbol{\mu} \in \mathcal{F}$. Let $\mathcal{I}(\boldsymbol{\mu})$ and $\mathcal{J}(\boldsymbol{\mu})$ denote the set of indices of active inequality constraints under $\boldsymbol{\mu}$ i.e. $\boldsymbol{\Psi}_i^\top \boldsymbol{\mu} = \boldsymbol{b}_i$ and $\boldsymbol{\mu}(s, a) = 0$ if and only if $i \in \mathcal{I}(\boldsymbol{\mu})$ and $(s, a) \in \mathcal{J}(\boldsymbol{\mu})$. Then,*

$$\boldsymbol{\mu} \in \mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}) \iff \boldsymbol{r} \in \partial f(\boldsymbol{\mu}) + N_{\mathcal{F}}(\boldsymbol{\mu}), \tag{64}$$

where $N_{\mathcal{F}}(\boldsymbol{\mu}) = \mathcal{U} + \mathcal{C}(\boldsymbol{\mu}) + \mathcal{E}(\boldsymbol{\mu})$ *with*

$$\mathcal{C}(\boldsymbol{\mu}) := \mathbf{cone}\left(\{\boldsymbol{\Psi}_i\}_{i\in\mathcal{I}(\boldsymbol{\mu})}\right),$$
$$\mathcal{E}(\boldsymbol{\mu}) := \mathbf{cone}\left(\{-\boldsymbol{e}_{s,a}\}_{(s,a)\in\mathcal{J}(\boldsymbol{\mu})}\right).$$

Here, $\boldsymbol{e}_{s,a} \in \mathbb{R}^{nm}$ *denote the standard unit vectors with* $\boldsymbol{e}_{s,a}(s',a') = 1$ *if* $(s,a) = (s',a')$ *and* $\boldsymbol{e}_{s,a}(s',a') = 0$ *otherwise.*

Proof. The main idea of the proof is to use Theorem C.1 and C.2 to prove that $\boldsymbol{\mu} \in \mathcal{F}$ is optimal for some $\boldsymbol{r}$ if and only if $\boldsymbol{r} \in \partial f(\boldsymbol{\mu}) + N_{\mathcal{F}}(\boldsymbol{\mu})$. In order to apply Theorem C.1 and C.2 to the constrained MDP problem, we recall that $f : \mathcal{X} \to \mathbb{R}$ is by definition a continuous convex function with $\mathcal{X} \subseteq \mathbb{R}^{nm}$ closed convex. We define the extended real value functions

$$\bar{f} : \mathbb{R}^{nm} \to [-\infty, \infty],\ \boldsymbol{\mu} \mapsto \bar{f}(\boldsymbol{\mu}) := \begin{cases} f(\boldsymbol{\mu}), & \boldsymbol{\mu} \in \mathcal{X}, \\ \infty, & \boldsymbol{\mu} \notin \mathcal{X}, \end{cases} \tag{65}$$

and

$$g_{\mathcal{F}} : \mathbb{R}^{nm} \to [-\infty, \infty],\ \boldsymbol{\mu} \mapsto g_{\mathcal{F}}(\boldsymbol{\mu}) := \bar{f}(\boldsymbol{\mu}) + \delta_{\mathcal{F}}(\boldsymbol{\mu}), \tag{66}$$

where $\delta_{\mathcal{F}}$ is the characteristic function

$$\delta_{\mathcal{F}}(\boldsymbol{\mu}) := \begin{cases} 0 & , \boldsymbol{\mu} \in \mathcal{F} \\ \infty & , \boldsymbol{\mu} \notin \mathcal{F}. \end{cases} \tag{67}$$

Note that since f is continuous and $\mathcal{F}$ closed, $g_{\mathcal{F}}$ is a closed proper convex function. Now, we can rewrite the CMDP problem (P) as

$$\max_{\boldsymbol{\mu}\in\mathcal{F}} \boldsymbol{r}^\top \boldsymbol{\mu} - \bar{f}(\boldsymbol{\mu}) = \max_{\boldsymbol{\mu}\in\mathbb{R}^{nm}} \boldsymbol{r}^\top \boldsymbol{\mu} - g_{\mathcal{F}}(\boldsymbol{\mu}) = g^*_{\mathcal{F}}(\boldsymbol{r}), \tag{68}$$

which is exactly taking the form of the convex conjugate of $g_{\mathcal{F}}$.[9] Therefore, Theorem C.1 yields

$$\boldsymbol{\mu} \in \mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}) = \operatorname*{argmax}_{\boldsymbol{\mu}\in\mathcal{F}} \boldsymbol{r}^\top \boldsymbol{\mu} - g_{\mathcal{F}}(\boldsymbol{\mu}) \iff \boldsymbol{r} \in \partial g_{\mathcal{F}}(\boldsymbol{\mu}). \tag{69}$$

Since Slater's condition is satisfied we have $\mathbf{relint}(\mathbf{dom}\,\delta_{\mathcal{F}}) \cap \mathbf{relint}(\mathbf{dom}\,\bar{f}) = \mathbf{relint}\,\mathcal{F} \cap \mathbf{relint}\,\mathcal{X} = \mathbf{relint}\,\mathcal{F} \neq \emptyset$. Hence, the conditions of Theorem C.2 are satisfied and we get

$$\partial g_{\mathcal{F}}(\boldsymbol{\mu}) = \partial \bar{f}(\boldsymbol{\mu}) + \partial \delta_{\mathcal{F}}(\boldsymbol{\mu}). \tag{70}$$

Using that $\partial \bar{f} = \partial f$ and $\partial \delta_{\mathcal{F}}(\boldsymbol{\mu}) = N_{\mathcal{F}}(\boldsymbol{\mu})$ (Rockafellar, 1970) we arrive at

$$\boldsymbol{r} \in \mathsf{IRL}_{\mathcal{F}}(\boldsymbol{\mu}) \iff \boldsymbol{r} \in \partial f(\boldsymbol{\mu}) + N_{\mathcal{F}}(\boldsymbol{\mu}). \tag{71}$$

To finish the proof we note that for the polyhedron $\mathcal{F}$ the normal cone takes the form

$$N_{\mathcal{F}}(\boldsymbol{\mu}) = \mathbf{span}\,(\boldsymbol{E} - \gamma \boldsymbol{P}) + \mathbf{cone}\left(\{\boldsymbol{\Psi}_i\}_{i\in\mathcal{I}(\boldsymbol{\mu})}\right) + \mathbf{cone}\left(\{-\boldsymbol{e}_{s,a}\}_{(s,a)\in\mathcal{J}(\boldsymbol{\mu})}\right), \tag{72}$$

where $\mathcal{I}(\boldsymbol{\mu})$ and $\mathcal{J}(\boldsymbol{\mu})$ are the sets of active safety and non-negativity constraints, respectively (Rockafellar & Wets, 2009). □

Remark C.3. Note that as a consequence of (69) it holds $\mathsf{IRL}_{\mathcal{F}}(\boldsymbol{\mu}) = \partial g_{\mathcal{F}}(\boldsymbol{\mu})$ and since $g_{\mathcal{F}}$ is closed proper convex Theorem C.1 also implies $\mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}) = \partial g^*_{\mathcal{F}}(\boldsymbol{r})$.

[9] Since the maximum is achieved here, we can replace the supremum with the maximum.

C.4. Identifiability for State-Action-State Rewards

Throughout this paper our focus lies on state-action rewards $\boldsymbol{r} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}, (s,a) \mapsto \boldsymbol{r}(s,a)$ that are naturally arising in the convex analytic approach to CMDPs (see (P) and (Altman, 1999)), and as the dual variables to the occupancy measure matching problem (see (23) and (Ho & Ermon, 2016)). However, some authors (Ng et al., 1999; Sutton & Barto, 2018; Skalse et al., 2022) also consider state-action-state rewards $\bar{\boldsymbol{r}} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}, (s,a,s') \mapsto \bar{\boldsymbol{r}}(s,a,s')$ that are allowed to depend on the consecutive state $s' \sim \boldsymbol{P}(\cdot|s,a)$. As mentioned in Section 4, this adds no generality to the forward CMDP problem, since a CMDP problem with state-action-state reward $\bar{\boldsymbol{r}}$ is equivalent to a CMDP problem with the state-action reward $\boldsymbol{r}(s,a) := \mathbb{E}_{s' \sim \boldsymbol{P}(\cdot|s,a)} \bar{\boldsymbol{r}}(s,a,s')$. Nevertheless, in practice the transition law is typically unknown and it may in certain cases be easier to specify a state-action-state reward. To relate our identifiability results to this setting, we make use of the following vector notation for a state-action-state reward $\bar{\boldsymbol{r}}(s,a,s')$

$$\bar{\boldsymbol{r}} = \left[\bar{\boldsymbol{r}}_{s'_1}^\top, \dots, \bar{\boldsymbol{r}}_{s'_n}^\top\right]^\top \in \mathbb{R}^{n^2 m}, \quad \text{with } \bar{\boldsymbol{r}}_{s'} = \bar{\boldsymbol{r}}(\cdot,\cdot,s') \in \mathbb{R}^{nm}, \tag{73}$$

and define the linear mapping $\mathbf{A} : \mathbb{R}^{n^2 m} \to \mathbb{R}^{nm}$ via $(\boldsymbol{A}\bar{\boldsymbol{r}})(s,a) := \mathbb{E}_{s' \sim \boldsymbol{P}(\cdot|s,a)} \bar{\boldsymbol{r}}(s,a,s')$. Furthermore, we denote $\overline{\mathsf{RL}}_{\mathcal{F}}(\bar{\boldsymbol{r}}) := \mathsf{RL}_{\mathcal{F}}(\boldsymbol{A}\bar{\boldsymbol{r}})$ for the CMDP solution map for state-action-state rewards. The following corollary shows that identifiability of state-action-state rewards can be reduced to identifiability of state-action rewards – and hence to the result of Theorem 4.5.

Corollary C.4. *Let Assumption 3.2 hold and consider $\boldsymbol{\mu} \in \mathcal{F}$. Then,*

$$\boldsymbol{\mu} \in \overline{\mathsf{RL}}_{\mathcal{F}}(\bar{\boldsymbol{r}}) \iff \bar{\boldsymbol{r}} \in \left\{ \begin{bmatrix} \boldsymbol{r} \\ \vdots \\ \boldsymbol{r} \end{bmatrix} \,\middle|\, \boldsymbol{r} \in \partial f(\boldsymbol{\mu}) + \mathcal{U} + \mathcal{C}(\boldsymbol{\mu}) + \mathcal{E}(\boldsymbol{\mu}) \right\} + \mathcal{V}, \tag{74}$$

where $\mathcal{V} = \ker \boldsymbol{A}$ with $\dim \mathcal{V} = n(nm-1)$.

Proof. By Theorem 4.5, we have $\boldsymbol{\mu} \in \overline{\mathsf{RL}}_{\mathcal{F}}(\bar{\boldsymbol{r}}) = \mathsf{RL}_{\mathcal{F}}(\boldsymbol{A}\bar{\boldsymbol{r}})$ if and only if $\boldsymbol{A}\bar{\boldsymbol{r}} \in \partial f(\boldsymbol{\mu}) + \mathcal{U} + \mathcal{C}(\boldsymbol{\mu}) + \mathcal{E}(\boldsymbol{\mu})$. It therefore suffices to show that $\boldsymbol{A}\bar{\boldsymbol{r}} = \boldsymbol{r}$ if and only if $\bar{\boldsymbol{r}} = \bar{\boldsymbol{r}}' + \bar{\boldsymbol{r}}''$ with $\bar{\boldsymbol{r}}' = [\boldsymbol{r}, \dots, \boldsymbol{r}]^\top$ and $\bar{\boldsymbol{r}}'' \in \ker \boldsymbol{A}$. If $\bar{\boldsymbol{r}} = \bar{\boldsymbol{r}}' + \bar{\boldsymbol{r}}''$ with $\bar{\boldsymbol{r}}' = [\boldsymbol{r}, \dots, \boldsymbol{r}]^\top$ and $\bar{\boldsymbol{r}}'' \in \ker \boldsymbol{A}$, then it follows from $\boldsymbol{A}\bar{\boldsymbol{r}}' = \boldsymbol{r}$ that $\boldsymbol{A}\bar{\boldsymbol{r}} = \boldsymbol{r} + \boldsymbol{0}$. Conversely, if $\boldsymbol{A}\bar{\boldsymbol{r}} = \boldsymbol{r}$, then $\bar{\boldsymbol{r}}' = [\boldsymbol{r}, \dots, \boldsymbol{r}]^\top$ also satisfies $\boldsymbol{A}\bar{\boldsymbol{r}}' = \boldsymbol{r}$, and thus $\bar{\boldsymbol{r}} - \bar{\boldsymbol{r}}' \in \ker \boldsymbol{A}$.

Finally, since for $\bar{\boldsymbol{r}} = [\boldsymbol{r}, \dots, \boldsymbol{r}]^\top$ we have $\boldsymbol{A}\bar{\boldsymbol{r}} = \boldsymbol{r}$, the mapping $\boldsymbol{A}$ is surjective and thus $\dim \mathcal{V} = n(nm-1)$. □

From a more abstract perspective, Corollary C.4 makes use of the fact that the image $\mathbf{im}\boldsymbol{A} = \mathbb{R}^{nm}$ of $\boldsymbol{A}$ is isomorphic to the quotient space $\mathbb{R}^{n^2 m} / \ker \boldsymbol{A}$ (Halmos, 2017), where $\mathbb{R}^{n^2 m} / \ker \boldsymbol{A}$ is the set of equivalence classes $[\bar{\boldsymbol{r}}'] := \left\{ \bar{\boldsymbol{r}} \in \mathbb{R}^{n^2 m} : \bar{\boldsymbol{r}} = \bar{\boldsymbol{r}}' + \bar{\boldsymbol{r}}'', \bar{\boldsymbol{r}}'' \in \ker \boldsymbol{A} \right\}$.

For unconstrained MDPs Skalse et al. (2022) discuss invariances of optimal policies to reward transformation for state-action-state rewards. They introduce the additional invariances along the linear subspace $\mathcal{V} = \ker \mathbf{A}$ as *s'-redistribution*. Moreover, they show that identifying a state-action-state reward up to s'-redistribution is not sufficient for generalizability to new environments. In contrast, our result in Theorem 4.12 shows that for state-action rewards identifying the rewards up to potential shaping is not enough for generalizability and that we instead need to recover the expert's reward up to a constant. However, this result is not extending immediately to the state-action-state setting, and one would need to modify the proof of Theorem 4.12 to additionally account for the space $\mathcal{V}$ in order to make a statement about generalizability.

C.5. Proof of Corollary 4.10

Assumption 4.9 Let $f : \mathcal{X} \to \mathbb{R}$ be such that:

(a) f is differentiable throughout $\mathbf{int}\, \mathcal{X}$,

(b) $\lim_{k\to\infty} \|\nabla f(\boldsymbol{\mu}_k)\| = \infty$ if $(\boldsymbol{\mu}_k)_{k\in\mathbb{N}}$ is a sequence in $\mathbf{int}\, \mathcal{X}$ converging to a point $\boldsymbol{\mu} \in \mathbf{relbd}\, \mathcal{M}$.

Corollary 4.10 *Let Assumptions 3.2, 4.1, 4.9 hold. Then, we have $\mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}) \subset \mathbf{relint}\, \mathcal{M}$ for any $\boldsymbol{r} \in \mathbb{R}^{nm}$ and*

$$\mathsf{IRL}_{\mathcal{F}}(\boldsymbol{\mu}^E) = \nabla f(\boldsymbol{\mu}^E) + \mathcal{U} + \mathcal{C}(\boldsymbol{\mu}^E). \tag{75}$$

Proof. Under Assumption 3.2, Theorem 4.5 states that

$$\boldsymbol{\mu} \in \mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}) \iff \boldsymbol{r} \in \partial f(\boldsymbol{\mu}) + N_{\mathcal{F}}(\boldsymbol{\mu}). \tag{76}$$

However, Assumption 4.9 ensures that $f(\boldsymbol{\mu}) = \{\nabla f(\boldsymbol{\mu})\}$ for $\boldsymbol{\mu} \in \mathbf{relint}\, \mathcal{M}$ and $\partial f(\boldsymbol{\mu}) = \emptyset$ for $\boldsymbol{\mu} \in \mathbf{relbd}\, \mathcal{M}$ (see (Rockafellar, 1970, Theorem 25.6.)). Hence,

$$\boldsymbol{\mu} \in \mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}) \subset \mathbf{relbd}\, \mathcal{M} \iff \boldsymbol{r} \in \emptyset + N_{\mathcal{F}}(\boldsymbol{\mu}) = \emptyset. \tag{77}$$

Furthermore, under Assumption 4.1, it follows from differentiability in $\mathbf{relint}\, \mathcal{M}$ and Corollary 4.7 that

$$\mathsf{IRL}_{\mathcal{F}}(\boldsymbol{\mu}^{\mathrm{E}}) = \nabla f(\boldsymbol{\mu}^{\mathrm{E}}) + \mathcal{U} + \mathcal{C}(\boldsymbol{\mu}^{\mathrm{E}}). \tag{78}$$

□

C.6. Proof of Theorem 4.12

Theorem 4.12 *Let Assumption 3.2, 3.3, 4.1, 4.9 be satisfied for $(\boldsymbol{P}_0, \boldsymbol{b}_0)$ and let $\boldsymbol{\mu}^E \in \mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{r}^E)$ for some $\boldsymbol{r}^E \in \mathcal{R}$. Consider an arbitrary neighborhood $\mathcal{O}_{\boldsymbol{P}_0} \subseteq \mathbb{R}^{nm\times n}$ of $\boldsymbol{P}_0$. Then, IRL generalizes to $\mathcal{P} = \mathcal{O}_{\boldsymbol{P}_0} \cap \mathfrak{P}$ and $\mathcal{B} = \mathbb{R}^k$ if and only if*

$$\mathsf{IRL}_{\mathcal{R},\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{\mu}^E) \subseteq \boldsymbol{r}^E + \mathbf{span}(\mathbf{1}_{nm}). \tag{79}$$

Proof. The *if* direction is trivial, since addition of a constant is not changing the set of optimal occupancy measures. Hence, if $\mathsf{IRL}_{\mathcal{R},\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{\mu}^{\mathrm{E}}) \subseteq \boldsymbol{r}^{\mathrm{E}} + \mathbf{span}(\mathbf{1}_{nm})$, then IRL generalizes to any arbitrary set of transition laws and constraint thresholds.

To prove the *only if* direction, we proceed in the following steps:

1. Show that Slater's condition is still satisfied in a sufficiently small neighborhood of $\boldsymbol{P}_0$.
2. Apply Theorem 4.5 to rewrite generalizability as a condition on the rewards.
3. Construct $\bar{\boldsymbol{P}}_1, \bar{\boldsymbol{P}}_2 \in \mathfrak{P}$ such that only $\boldsymbol{r}^{\mathrm{E}} + \mathbf{span}\,(\mathbf{1}_{nm})$ generalize to $\bar{\boldsymbol{P}}_1, \bar{\boldsymbol{P}}_2$.
4. Use $\bar{\boldsymbol{P}}_1, \bar{\boldsymbol{P}}_2$ to construct $\boldsymbol{P}_1, \boldsymbol{P}_2 \in \mathcal{P}$ such that only $\boldsymbol{r}^{\mathrm{E}} + \mathbf{span}\,(\mathbf{1}_{nm})$ generalize to $\boldsymbol{P}_1, \boldsymbol{P}_2$.

Step 1:
By Assumption 3.2 (Slater's condition) there is some occupancy measure $\bar{\boldsymbol{\mu}} \in \mathbf{relint}\, \mathcal{F}^{\boldsymbol{P}_0,\boldsymbol{b}_0}$ where

$$\mathcal{F}^{\boldsymbol{P}_0,\boldsymbol{b}_0} := \left\{\boldsymbol{\mu} \in \mathbb{R}^{nm} : \boldsymbol{\mu} \geq \mathbf{0}, (\boldsymbol{E} - \gamma \boldsymbol{P}_0)^\top \boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0, \boldsymbol{\Psi}^\top \boldsymbol{\mu} \leq \boldsymbol{b}_0\right\}. \tag{80}$$

Note that $\mathcal{F}^{\boldsymbol{P}_0,\boldsymbol{b}_0} \subseteq \mathcal{F}^{\boldsymbol{P}_0,\boldsymbol{b}}$ for $\boldsymbol{b} \geq \boldsymbol{b}_0$. Thus, Slater's condition remains to hold when the constraint threshold is relaxed. Furthermore, for $\boldsymbol{P} \in \mathfrak{P}$ consider the orthogonal projection

$$\mathrm{Proj}^{\boldsymbol{P}}(\bar{\boldsymbol{\mu}}) := \underset{\boldsymbol{\mu}:(\boldsymbol{E}-\gamma\boldsymbol{P})^\top\boldsymbol{\mu}=(1-\gamma)\boldsymbol{\nu}_0}{\operatorname{argmin}} \|\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}\|_2, \tag{81}$$

of $\bar{\boldsymbol{\mu}}$ onto the affine hull of $\mathcal{F}^{\boldsymbol{P},\boldsymbol{b}_0}$. Since $(\boldsymbol{E} - \gamma\boldsymbol{P})$ has full rank, the projection $\mathrm{Proj}^{\boldsymbol{P}}(\bar{\boldsymbol{\mu}})$ is continuous in $\boldsymbol{P}$ (Penrose, 1955; Ding, 1993). Therefore, if $\boldsymbol{P}$ is sufficiently close to $\boldsymbol{P}_0$, we have $\mathrm{Proj}^{\boldsymbol{P}}(\bar{\boldsymbol{\mu}}) > \mathbf{0}$ and $\boldsymbol{\Psi}^\top \mathrm{Proj}^{\boldsymbol{P}}(\bar{\boldsymbol{\mu}}) < \boldsymbol{b}_0$ i.e. $\mathrm{Proj}^{\boldsymbol{P}}(\bar{\boldsymbol{\mu}}) \in \mathbf{relint}\, \mathcal{F}^{\boldsymbol{P},\boldsymbol{b}_0}$. Hence, there exists some neighborhood $\mathcal{O}'_{\boldsymbol{P}_0} \subseteq \mathcal{O}_{\boldsymbol{P}_0}$ of $\boldsymbol{P}_0$ such that $\mathbf{relint}\, \mathcal{F}^{\boldsymbol{P},\boldsymbol{b}_0} \neq \emptyset$ for all $\boldsymbol{P} \in \mathcal{O}'_{\boldsymbol{P}_0} \cap \mathfrak{P}$.

Step 2:
Since Assumption 3.3 (strict convexity) holds, all sets $\mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}(\boldsymbol{r})$ are singleton for any $\boldsymbol{r}, \boldsymbol{P}, \boldsymbol{b}$. Throughout this proof we therefore interpret $\mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}$ as a single-valued mapping and write $\boldsymbol{\mu} = \mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}(\boldsymbol{r})$ instead of $\{\boldsymbol{\mu}\} = \mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}(\boldsymbol{r})$.

Now, let $\mathcal{P}' := \mathcal{O}'_{\boldsymbol{P}_0} \cap \mathfrak{P} \subseteq \mathcal{P}$ and $\mathcal{B}' := \{\boldsymbol{b} \in \mathcal{B} : \boldsymbol{b} \geq \boldsymbol{b}_0\} \subseteq \mathcal{B}$. Due to Assumption 4.1 (realizability) and Proposition 4.2 we have $\boldsymbol{r}^{\mathrm{E}} \in \mathsf{IRL}_{\mathcal{R},\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{\mu}^{\mathrm{E}})$. Hence, the following chain of implications holds:

$$
\begin{aligned}
&\mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}(\boldsymbol{r}) = \mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}(\boldsymbol{r}'),\ \forall \boldsymbol{r}, \boldsymbol{r}' \in \mathsf{IRL}_{\mathcal{R},\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{\mu}^{\mathrm{E}}), \forall \boldsymbol{P} \in \mathcal{P}, \forall \boldsymbol{b} \in \mathcal{B} \qquad (82)\\
&\overset{(i)}{\iff} \mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}(\boldsymbol{r}) = \mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}(\boldsymbol{r}^{\mathrm{E}}),\ \forall \boldsymbol{r} \in \mathsf{IRL}_{\mathcal{R},\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{\mu}^{\mathrm{E}}), \forall \boldsymbol{P} \in \mathcal{P}, \forall \boldsymbol{b} \in \mathcal{B}\\
&\overset{(ii)}{\iff} \mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}(\boldsymbol{r}) = \mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}(\boldsymbol{r}^{\mathrm{E}}),\ \forall \boldsymbol{r} \in \mathsf{IRL}_{\mathcal{R},\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{\mu}^{\mathrm{E}}), \forall \boldsymbol{P} \in \mathcal{P}', \forall \boldsymbol{b} \in \mathcal{B}'\\
&\overset{(iii)}{\iff} \mathsf{IRL}_{\mathcal{R},\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{\mu}^{\mathrm{E}}) \subseteq \left[\partial f(\boldsymbol{\mu}) + \mathcal{U}^{\boldsymbol{P}} + \mathcal{C}^{\boldsymbol{b}}(\boldsymbol{\mu}) + \mathcal{E}(\boldsymbol{\mu})\right]_{\boldsymbol{\mu}=\mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}(\boldsymbol{r}^{\mathrm{E}})},\ \forall \boldsymbol{P} \in \mathcal{P}', \forall \boldsymbol{b} \in \mathcal{B}'\\
&\overset{(iv)}{\iff} \mathsf{IRL}_{\mathcal{R},\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{\mu}^{\mathrm{E}}) \subseteq \bigcap_{\boldsymbol{P} \in \mathcal{P}'} \bigcap_{\boldsymbol{b} \in \mathcal{B}'} \left[\partial f(\boldsymbol{\mu}) + \mathcal{U}^{\boldsymbol{P}} + \mathcal{C}^{\boldsymbol{b}}(\boldsymbol{\mu}) + \mathcal{E}(\boldsymbol{\mu})\right]_{\boldsymbol{\mu}=\mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}(\boldsymbol{r}^{\mathrm{E}})}\\
&\overset{(v)}{\iff} \mathsf{IRL}_{\mathcal{R},\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{\mu}^{\mathrm{E}}) \subseteq \bigcap_{\boldsymbol{P} \in \mathcal{P}'} \bigcap_{\boldsymbol{b} \in \mathcal{B}'} \left[\nabla f(\boldsymbol{\mu}) + \mathcal{U}^{\boldsymbol{P}} + \mathcal{C}^{\boldsymbol{b}}(\boldsymbol{\mu})\right]_{\boldsymbol{\mu}=\mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}(\boldsymbol{r}^{\mathrm{E}})}.
\end{aligned}
$$

Here, (i) holds since $\boldsymbol{r}^{\mathrm{E}} \in \mathsf{IRL}_{\mathcal{R},\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{\mu}^{\mathrm{E}})$, (ii) follows from $\mathcal{P}' \subseteq \mathcal{P}$ and $\mathcal{B}' \subseteq \mathcal{B}$, and (iii) is a consequence of Theorem 4.5 which applies since Assumption 3.2 (Slater's condition) is satisfied for all $\boldsymbol{P} \in \mathcal{P}', \boldsymbol{b} \in \mathcal{B}'$. Moreover, (iv) follows from the definition of the intersection, and (v) from differentiability and Assumption 4.9 which ensures $\mathcal{E}(\boldsymbol{\mu}) = \boldsymbol{0}$.

Next, we recall that $\mathcal{B} = \mathbb{R}^k$. Thus, we may choose a large enough $\bar{\boldsymbol{b}} \in \mathcal{B}'$ such that the set of active safety constraints is empty for any $\boldsymbol{\mu}$ and hence $\mathcal{C}^{\bar{\boldsymbol{b}}}(\boldsymbol{\mu}) = \boldsymbol{0}$. This allows us to further simplify (82) to:

$$
\begin{aligned}
&\mathsf{IRL}_{\mathcal{R},\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{\mu}^{\mathrm{E}}) \subseteq \bigcap_{\boldsymbol{P} \in \mathcal{P}'} \bigcap_{\boldsymbol{b} \in \mathcal{B}'} \left[\nabla f(\boldsymbol{\mu}) + \mathcal{U}^{\boldsymbol{P}} + \mathcal{C}^{\boldsymbol{b}}(\boldsymbol{\mu})\right]_{\boldsymbol{\mu}=\mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\boldsymbol{b}}(\boldsymbol{r}^{\mathrm{E}})} \qquad (83)\\
&\iff \mathsf{IRL}_{\mathcal{R},\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{\mu}^{\mathrm{E}}) \subseteq \bigcap_{\boldsymbol{P} \in \mathcal{P}'} \left[\nabla f(\boldsymbol{\mu}) + \mathcal{U}^{\boldsymbol{P}}\right]_{\boldsymbol{\mu}=\mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P},\bar{\boldsymbol{b}}}(\boldsymbol{r}^{\mathrm{E}})}\\
&\iff \mathsf{IRL}_{\mathcal{R},\mathcal{F}}^{\boldsymbol{P}_0,\boldsymbol{b}_0}(\boldsymbol{\mu}^{\mathrm{E}}) \subseteq \bigcap_{\boldsymbol{P} \in \mathcal{P}'} \left[\boldsymbol{r}^{\mathrm{E}} + \mathcal{U}^{\boldsymbol{P}}\right] = \boldsymbol{r}^{\mathrm{E}} + \bigcap_{\boldsymbol{P} \in \mathcal{P}'} \mathcal{U}^{\boldsymbol{P}}.
\end{aligned}
$$

Therefore, it suffices to show that $\bigcap_{\boldsymbol{P} \in \mathcal{P}'} \mathcal{U}^{\boldsymbol{P}} \subseteq \mathbf{span}\, \mathbf{1}_{nm}$. In particular, it is enough to show that $\mathcal{U}^{\boldsymbol{P}_1} \cap \mathcal{U}^{\boldsymbol{P}_2} = \mathbf{span}\,(\mathbf{1}_{nm})$ for two $\boldsymbol{P}_1, \boldsymbol{P}_2 \in \mathcal{P}'$. To that end, we will continue by first showing that there are $\bar{\boldsymbol{P}}_1, \bar{\boldsymbol{P}}_2 \in \mathfrak{P}$ such that $\mathcal{U}^{\bar{\boldsymbol{P}}_1} \cap \mathcal{U}^{\bar{\boldsymbol{P}}_2} = \mathbf{span}\,(\mathbf{1}_{nm})$.

Step 3: Note that, as shown by Rolland et al. (2022), the condition $\mathcal{U}^{\bar{\boldsymbol{P}}_1} \cap \mathcal{U}^{\bar{\boldsymbol{P}}_2} = \mathbf{span}\,(\mathbf{1}_{nm})$ is equivalent to

$$
\mathbf{rank}\left[\boldsymbol{E} - \gamma \bar{\boldsymbol{P}}_1, \boldsymbol{E} - \gamma \bar{\boldsymbol{P}}_2\right] = 2n - 1. \qquad (84)
$$

This follows from the two facts that (a) for any $\boldsymbol{P} \in \mathfrak{P}$ we have $\mathbf{1}_{nm} \in \mathcal{U}^{\boldsymbol{P}}$ and (b) the condition (84) is equivalent to $\dim(\mathcal{U}^{\bar{\boldsymbol{P}}_1} \cap \mathcal{U}^{\bar{\boldsymbol{P}}_2}) = 1$. Here, (a) holds since

$$
(\boldsymbol{E} - \gamma \boldsymbol{P})\mathbf{1}_n = \begin{bmatrix} (\boldsymbol{I}_n - \gamma \boldsymbol{P}_{a_1})\mathbf{1}_n \\ \vdots \\ (\boldsymbol{I}_n - \gamma \boldsymbol{P}_{a_m})\mathbf{1}_n \end{bmatrix} = \begin{bmatrix} (1-\gamma)\mathbf{1}_n \\ \vdots \\ (1-\gamma)\mathbf{1}_n \end{bmatrix} = (1-\gamma)\mathbf{1}_{nm}, \qquad (85)
$$

where we use that $\mathbf{1}_n$ is for both $\boldsymbol{I}_n$ and $\boldsymbol{P}_{a_i}, i = 1, ..., m$ an eigenvector to the eigenvalue 1. Furthermore, (b) is a consequence of

$$
\dim(\mathbf{span}\, \boldsymbol{A}_1 \cap \mathbf{span}\, \boldsymbol{A}_2) = (\mathbf{rank}\, \boldsymbol{A}_1 + \mathbf{rank}\, \boldsymbol{A}_2) - \mathbf{rank}\left[\boldsymbol{A}_1, \boldsymbol{A}_2\right], \qquad (86)
$$

for the two matrices $\boldsymbol{A}_i = \boldsymbol{E} - \gamma \bar{\boldsymbol{P}}_i, i = 1, 2$. Therefore, the goal for this step is to prove the following claim:

Claim There exist $\bar{\boldsymbol{P}}_1, \bar{\boldsymbol{P}}_2 \in \mathfrak{P}$ such that $\mathbf{rank}\left[\boldsymbol{E} - \gamma \bar{\boldsymbol{P}}_1, \boldsymbol{E} - \gamma \bar{\boldsymbol{P}}_2\right] = 2n - 1$.

Note that since the vector $\left[\mathbf{1}_n^\top, -\mathbf{1}_n^\top\right]^\top$ lies in the kernel of $\left[\boldsymbol{E} - \gamma \bar{\boldsymbol{P}}_1, \boldsymbol{E} - \gamma \bar{\boldsymbol{P}}_2\right]$, the rank cannot be larger than $2n - 1$. Moreover, since the rank of a matrix is larger or equal than the rank of any submatrix, it suffices to prove the claim for $m = 2$. To this end, let $\bar{\boldsymbol{P}}_1, \bar{\boldsymbol{P}}_2 \in \mathfrak{P}$ be defined as follows

$$
\bar{\boldsymbol{P}}_1 = \begin{bmatrix} \boldsymbol{I}_n \\ \boldsymbol{D} \end{bmatrix},\ \bar{\boldsymbol{P}}_2 = \begin{bmatrix} \boldsymbol{D} \\ \boldsymbol{I}_n \end{bmatrix}, \qquad (87)
$$

where

$$\boldsymbol{D} := \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \in \mathbb{R}^{n \times n}. \tag{88}$$

It then holds $\mathbf{rank}(\boldsymbol{I}_n - \boldsymbol{D}) = n - 1$ as is readily seen since $\boldsymbol{I}_n - \boldsymbol{D}$ is an upper triangular matrix in row echelon form. Now, in order to prove that the rank of

$$C := \left[\boldsymbol{E} - \gamma \bar{\boldsymbol{P}}_1, \boldsymbol{E} - \gamma \bar{\boldsymbol{P}}_2\right] \in \mathbb{R}^{2n \times 2n}, \tag{89}$$

equals $2n - 1$, we show that the first $2n - 1$ columns of C are linearly independent. To this end, let $C^{(-2n)} \in \mathbb{R}^{2n \times (2n-1)}$ denote the submatrix obtained by removing the last column of C. Moreover, let $\boldsymbol{z} := \begin{bmatrix} \boldsymbol{x}^\top & \boldsymbol{y}^\top \end{bmatrix}^\top$ with $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{y} \in \mathbb{R}^{n-1}$. The columns of $C^{(-2n)}$ are linearly independent if

$$C^{(-2n)}\boldsymbol{z} = \boldsymbol{0} \implies \boldsymbol{z} = \boldsymbol{0}. \tag{90}$$

Plugging in the definition of $\bar{\boldsymbol{P}}_1, \bar{\boldsymbol{P}}_2$ it holds

$$C^{(-2n)}\boldsymbol{z} = \begin{bmatrix} (1-\gamma)\boldsymbol{I}_n & \boldsymbol{I}_n - \gamma \boldsymbol{D}^{(-n)} \\ \boldsymbol{I}_n - \gamma \boldsymbol{D} & (1-\gamma)\boldsymbol{I}_n^{(-n)} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} (1-\gamma)\boldsymbol{I}_n & \boldsymbol{I}_n - \gamma \boldsymbol{D} \\ \boldsymbol{I}_n - \gamma \boldsymbol{D} & (1-\gamma)\boldsymbol{I}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ 0 \end{bmatrix}, \tag{91}$$

where we again use the notation $\boldsymbol{B}^{(-n)}$ to denote the submatrix of some matrix $\boldsymbol{B}$ obtained when removing the n-th column. Therefore, we can rewrite (90) as

$$\begin{aligned} \boldsymbol{x} + \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix} - \gamma \left(\boldsymbol{x} + \boldsymbol{D} \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix} \right) &= \boldsymbol{0} \\ \boldsymbol{x} + \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix} - \gamma \left(\boldsymbol{D}\boldsymbol{x} + \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix} \right) &= \boldsymbol{0}. \end{aligned} \tag{92}$$

Substituting $\tilde{\boldsymbol{x}} := \boldsymbol{x} + \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix}$ we get

$$\begin{aligned} (1-\gamma)\tilde{\boldsymbol{x}} &= \gamma(\boldsymbol{D} - \boldsymbol{I}_n) \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix} \\ (\boldsymbol{I}_n - \gamma \boldsymbol{D})\tilde{\boldsymbol{x}} &= -\gamma(\boldsymbol{D} - \boldsymbol{I}_n) \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix}. \end{aligned} \tag{93}$$

This implies

$$(1-\gamma)\tilde{\boldsymbol{x}} + (\boldsymbol{I}_n - \gamma \boldsymbol{D})\tilde{\boldsymbol{x}} = 2(\boldsymbol{I}_n - \gamma \tilde{\boldsymbol{D}})\tilde{\boldsymbol{x}} = \boldsymbol{0}, \tag{94}$$

for $\tilde{\boldsymbol{D}} := (\boldsymbol{I}_n + \boldsymbol{D})/2$. Since $\tilde{\boldsymbol{D}}$ is again a row stochastic matrix, $\boldsymbol{I}_n - \gamma \tilde{\boldsymbol{D}}$ is invertible and thus $\tilde{\boldsymbol{x}} = \boldsymbol{0}$. Moreover, since $\mathbf{rank}(\boldsymbol{I}_n - \boldsymbol{D}) = n - 1$ and $(\boldsymbol{I}_n - \boldsymbol{D})\boldsymbol{1}_n = \boldsymbol{0}$, we have $\ker(\boldsymbol{I}_n - \boldsymbol{D}) = \mathbf{span}(\boldsymbol{1}_n)$. In light of

$$(\boldsymbol{D} - \boldsymbol{I}_n) \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix} = \boldsymbol{0}, \tag{95}$$

this implies that $\boldsymbol{y} = \boldsymbol{0}$, which proves the claim.

Step 4:
Equipped with the above claim, the final step of the proof is to show that there are $\boldsymbol{P}_1, \boldsymbol{P}_2 \in \mathcal{P}' \subseteq \mathcal{P}$ such that

$$\mathcal{U}^{\boldsymbol{P}_1} \cap \mathcal{U}^{\boldsymbol{P}_2} = \mathbf{span}(\boldsymbol{1}_{nm}). \tag{96}$$

Analogously to the previous step, we will show that

$$\mathbf{rank}\left[\boldsymbol{E} - \gamma \boldsymbol{P}_1, \boldsymbol{E} - \gamma \boldsymbol{P}_2\right] = 2n - 1. \tag{97}$$

For this purpose, we choose two arbitrary (and possibly equal) $\boldsymbol{P}_{1,0}, \boldsymbol{P}_{2,0} \in \mathcal{P}'$ and define

$$(\boldsymbol{P}_1(\tau), \boldsymbol{P}_2(\tau)) := (1-\tau)(\boldsymbol{P}_{1,0}, \boldsymbol{P}_{2,0}) + \tau(\bar{\boldsymbol{P}}_1, \bar{\boldsymbol{P}}_2) \in \mathfrak{P} \times \mathfrak{P}. \tag{98}$$

Since $\mathbf{rank}\left[\boldsymbol{E} - \gamma \boldsymbol{P}_1(\tau), \boldsymbol{E} - \gamma \boldsymbol{P}_2(\tau)\right] = 2n-1$ for $\tau = 1$, there exists a $(2n-1) \times (2n-1)$ sub-matrix $\boldsymbol{C}_{\text{sub}}(\tau)$ of $\left[\boldsymbol{E} - \gamma \boldsymbol{P}_1(\tau), \boldsymbol{E} - \gamma \boldsymbol{P}_2(\tau)\right]$ that is invertible for $\tau = 1$. Thus, the function $h(\tau) := \det \boldsymbol{C}_{\text{sub}}(\tau)$ is a non-zero polynomial, which by the fundamental theorem of algebra can only have finitely many roots. Since the set $\mathcal{O}'_{\boldsymbol{P}_0}$ is a neighborhood of $\boldsymbol{P}_0$, it contains an open ball (in any norm[10]) around $\boldsymbol{P}_0$. Furthermore, $\mathfrak{P}$ is convex. Hence, we can always choose a small enough $\varepsilon > 0$ such that $\boldsymbol{P}_1(\varepsilon), \boldsymbol{P}_2(\varepsilon) \in \mathcal{P}' = \mathcal{O}'_{\boldsymbol{P}_0} \cap \mathfrak{P}$ and $h(\varepsilon) \neq 0$. However, for $h(\varepsilon) \neq 0$ the rank condition (97) is satisfied and we have proven the equality (96). □

Remark C.5. Note that in fact we have proven a stronger result than the equality (96) – namely that for large enough $\boldsymbol{b}$ the set of $(\boldsymbol{P}_1, \boldsymbol{P}_2)$ for which

$$\mathsf{IRL}_{\mathcal{F}}^{\boldsymbol{P}_1,\boldsymbol{b}} \circ \mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P}_1,\boldsymbol{b}}(\boldsymbol{r}^{\mathrm{E}}) \cap \mathsf{IRL}_{\mathcal{F}}^{\boldsymbol{P}_2,\boldsymbol{b}} \circ \mathsf{RL}_{\mathcal{F}}^{\boldsymbol{P}_2,\boldsymbol{b}}(\boldsymbol{r}^{\mathrm{E}}) = \boldsymbol{r}^{\mathrm{E}} + \mathbf{span}(\mathbf{1}_{nm}), \tag{99}$$

is dense in $\mathfrak{P} \times \mathfrak{P}$. Cao et al. (2021); Rolland et al. (2022) analyze (96) in the context of identifiability from two experts. They show experimentally that the rank condition (97) is always satisfied when randomly generating two transition laws $(\boldsymbol{P}_1, \boldsymbol{P}_2)$. However, the formal proof that the set where (97) is satisfied is dense in $\mathfrak{P} \times \mathfrak{P}$ is novel.

C.7. Proof of Proposition 4.14

Proposition 4.14 *Let Assumption 3.2, 3.3, 4.1, 4.9 hold with $\boldsymbol{\mu}^E \in \mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}^E)$ for some $\boldsymbol{r}^E \in \mathcal{R}$ and*

$$\mathcal{R} \subseteq \left\{\boldsymbol{r}_{\boldsymbol{w}} = \boldsymbol{\Phi}\boldsymbol{w} : \boldsymbol{\Phi} \in \mathbb{R}^{mn \times d}, \boldsymbol{w} \in \mathbb{R}^d\right\}. \tag{100}$$

Then, if for $\boldsymbol{\Xi} := \left[\boldsymbol{E} - \gamma \boldsymbol{P}, \boldsymbol{\Psi}\right]$ it holds that

$$\mathbf{rank}\left[\boldsymbol{\Phi}, \boldsymbol{\Xi}\right] - (\mathbf{rank}\,\boldsymbol{\Phi} + \mathbf{rank}\,\boldsymbol{\Xi}) = 0, \tag{101}$$

then we have $\mathsf{IRL}_{\mathcal{R},\mathcal{F}}(\boldsymbol{\mu}^E) = \left\{\boldsymbol{r}^E\right\}$.

Proof. First, we note that the rank condition is equivalent to the condition that the subspace spanned by the reward features intersects the Minkowski sum of the subspace of potential shaping transformations and the subspace spanned by the safety constraints, only at zero:

$$\begin{aligned}
&\mathbf{span}\,\boldsymbol{\Phi} \cap (\mathcal{U} + \mathbf{span}\,\boldsymbol{\Psi}) = \mathbf{0} \\
\iff &\dim\left(\mathbf{span}\,\boldsymbol{\Phi} \cap (\mathcal{U} + \mathbf{span}\,\boldsymbol{\Psi})\right) = 0 \\
\iff &\dim\left(\mathbf{span}\,\boldsymbol{\Phi}\right) + \dim\left(\mathbf{span}\,\boldsymbol{\Xi}\right) - \dim\left(\mathbf{span}\left[\boldsymbol{\Phi}, \boldsymbol{\Xi}\right]\right) = 0 \\
\iff &\mathbf{rank}\left[\boldsymbol{\Phi}, \boldsymbol{\Xi}\right] - (\mathbf{rank}\,\boldsymbol{\Phi} + \mathbf{rank}\,\boldsymbol{\Xi}) = 0.
\end{aligned} \tag{102}$$

To ease notation we define $\mathcal{V} := \mathbf{span}\,\boldsymbol{\Psi}$ and $\mathcal{W} := \mathbf{span}\,\boldsymbol{\Phi}$. Observe that $\mathcal{U}, \mathcal{V}, \mathcal{W}$ are linear subspaces of $\mathbb{R}^{nm}$. By Proposition 4.2, we have $\boldsymbol{r}^{\mathrm{E}} \in \mathsf{IRL}_{\mathcal{R},\mathcal{F}}(\boldsymbol{\mu}^{\mathrm{E}})$. Furthermore, due to (101) it holds $(\mathcal{U} + \mathcal{V}) \cap \mathcal{W} = \mathbf{0}$. We therefore get

$$\begin{aligned}
\mathsf{IRL}_{\mathcal{R},\mathcal{F}}(\boldsymbol{\mu}^{\mathrm{E}}) &\overset{(i)}{=} \left(\nabla f(\boldsymbol{\mu}^{\mathrm{E}}) + \mathcal{U} + \mathcal{C}\right) \cap \mathcal{R} \\
&\overset{(ii)}{\subseteq} \left(\nabla f(\boldsymbol{\mu}^{\mathrm{E}}) + \mathcal{U} + \mathcal{C}\right) \cap \mathcal{W} \\
&\overset{(iii)}{\subseteq} \left(\nabla f(\boldsymbol{\mu}^{\mathrm{E}}) + \mathcal{U} + \mathcal{V}\right) \cap \mathcal{W} \\
&\overset{(iv)}{=} \left(\boldsymbol{r}^{\mathrm{E}} + \mathcal{U} + \mathcal{V}\right) \cap \mathcal{W} \\
&\overset{(v)}{=} \boldsymbol{r}^{\mathrm{E}} + (\mathcal{U} + \mathcal{V}) \cap \mathcal{W} \\
&\overset{(vi)}{=} \boldsymbol{r}^{\mathrm{E}}
\end{aligned} \tag{103}$$

Here, we used Theorem 4.5 in (i), the reward class (100) in (ii), and the inclusion (iii) holds since $\mathcal{C} \subset \mathcal{V}$. Furthermore, (iv) and (v) follow from $\boldsymbol{r}^{\mathrm{E}} \in \mathsf{IRL}_{\mathcal{R},\mathcal{F}}(\boldsymbol{\mu}^{\mathrm{E}})$, and (vi) from (101). This concludes the proof. □

[10] Since all norms are equivalent in finite dimensional vector spaces the choice of norm is irrelevant.

D. Proof of Theorem 5.1

Theorem 5.1 *Let Assumption 4.1 and let* $\boldsymbol{\mu}^E \in \mathsf{RL}_{\mathcal{F}}(\boldsymbol{r}^E)$ *for some* $\boldsymbol{r}^E \in \mathcal{R} := \mathcal{R}^{\|\cdot\|_1}$. *Let* $\hat{\boldsymbol{\mu}} \in \mathsf{RL}_{\mathcal{F}} \circ \mathsf{IRL}_{\mathcal{R},\mathcal{F}}(\hat{\boldsymbol{\mu}}^E_{\mathcal{D}})$ *and* $R := \max_{s,a} \|\boldsymbol{\Phi}(s,a)\|_\infty$. *Choosing*

$$N = \left\lceil \frac{32R^2}{\varepsilon^2} \log\left(\frac{2d}{\delta}\right) \right\rceil \text{ and } T = \left\lceil \log\left(\frac{\varepsilon}{8R}\right) / \log(\gamma) \right\rceil, \tag{104}$$

it holds with probability at least $1-\delta$

$$\begin{aligned} &J(\boldsymbol{\mu}^E, \boldsymbol{r}^E) - J(\hat{\boldsymbol{\mu}}, \boldsymbol{r}^E) \leq \varepsilon, \\ &J(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}}) - J(\boldsymbol{\mu}^E, \hat{\boldsymbol{r}}) \leq \varepsilon, \\ &\forall \hat{\boldsymbol{r}} \in \mathsf{IRL}_{\mathcal{R},\mathcal{F}}(\hat{\boldsymbol{\mu}}^E_{\mathcal{D}}), \end{aligned} \tag{105}$$

where $J(\boldsymbol{\mu}, \boldsymbol{r}) := \boldsymbol{r}^\top \boldsymbol{\mu} - f(\boldsymbol{\mu})$. *Moreover, if*

(a) f *is* L*-strongly convex with respect to the norm* $\|\cdot\|$, *it holds with probability at least* $1-\delta$

$$\left\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^E\right\| \leq \sqrt{\frac{2\varepsilon}{L}}. \tag{106}$$

(b) $f(\boldsymbol{\mu}) = -\beta\, \mathbb{E}_{(s,a)\sim\boldsymbol{\mu}}\left[H\left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s)\right)\right]$ *with* $\beta > 0$, *it holds with probability at least* $1-\delta$

$$\mathbb{E}_{(s,a)\sim\boldsymbol{\mu}^E}\left[\left\|\boldsymbol{\pi}^{\hat{\boldsymbol{\mu}}}(\cdot|s) - \boldsymbol{\pi}^E(\cdot|s)\right\|_1\right] \leq \sqrt{\frac{2\varepsilon}{\beta}}. \tag{107}$$

Proof. Consider the idealized and the empirical min-max objective $L(\boldsymbol{\mu}, \boldsymbol{w}) := \boldsymbol{r}_{\boldsymbol{w}}^\top\left(\boldsymbol{\mu} - \boldsymbol{\mu}^{\mathrm{E}}\right) - f(\boldsymbol{\mu})$ and $\hat{L}(\boldsymbol{\mu}, \boldsymbol{w}) := \boldsymbol{r}_{\boldsymbol{w}}^\top\left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}^{\mathrm{E}}_{\mathcal{D}}\right) - f(\boldsymbol{\mu})$, the main idea of the proof it to get a uniform bound on $\left|L - \hat{L}\right|$. The statement in (105) is then following from the saddle point property of $(\boldsymbol{\mu}^{\mathrm{E}}, \boldsymbol{r}^{\mathrm{E}})$ and $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}})$. For (106) (and (107)) we are then using strict concavity (and the soft-suboptimality Lemma) to translate proximity of the optimal value to proximity of the optimal occupancy (and the optimal policy).

Step 1:
First, we define the true expert feature expectations $\boldsymbol{\sigma}^{\mathrm{E}} := \boldsymbol{\Phi}^\top \boldsymbol{\mu}^{\mathrm{E}}$, the empirical expert feature expectation $\hat{\boldsymbol{\sigma}}^{\mathrm{E}}_{\mathcal{D}} := \boldsymbol{\Phi}^\top \hat{\boldsymbol{\mu}}^{\mathrm{E}}_{\mathcal{D}}$, as well as $\boldsymbol{\sigma}^{\mathrm{E}}_T := (1-\gamma)\mathbb{E}\left[\sum_{t=0}^T \gamma^t \boldsymbol{\Phi}(s_t, a_t) \middle| \pi^{\mathrm{E}}\right]$. We can then decompose $\max\left|\hat{L} - L\right|$ as follows:

$$\begin{aligned} &\max_{\boldsymbol{\mu}\in\mathcal{F}, \|\boldsymbol{w}\|_1 \leq 1} \left|L(\boldsymbol{\mu}, \boldsymbol{w}) - \hat{L}(\boldsymbol{\mu}, \boldsymbol{w})\right| = \max_{\|\boldsymbol{w}\|_1 \leq 1} \left|\boldsymbol{w}^\top\left(\hat{\boldsymbol{\sigma}}^{\mathrm{E}}_{\mathcal{D}} - \boldsymbol{\sigma}^{\mathrm{E}}\right)\right| \\ &\overset{(i)}{\leq} \max_{\|\boldsymbol{w}\|_1 \leq 1} \left|\boldsymbol{w}^\top\left(\boldsymbol{\sigma}^{\mathrm{E}} - \boldsymbol{\sigma}^{\mathrm{E}}_T\right)\right| + \max_{\|\boldsymbol{w}\|_1 \leq 1} \left|\boldsymbol{w}^\top\left(\boldsymbol{\sigma}^{\mathrm{E}}_T - \hat{\boldsymbol{\sigma}}^{\mathrm{E}}_{\mathcal{D}}\right)\right| \\ &\overset{(ii)}{\leq} \underbrace{\left\|\boldsymbol{\sigma}^{\mathrm{E}} - \boldsymbol{\sigma}^{\mathrm{E}}_T\right\|_\infty}_{I_1} + \underbrace{\left\|\boldsymbol{\sigma}^{\mathrm{E}}_T - \hat{\boldsymbol{\sigma}}^{\mathrm{E}}_{\mathcal{D}}\right\|_\infty}_{I_2} \end{aligned} \tag{108}$$

Here, (i) follows from the triangle inequality for the supremum norm and (ii) from Hölder's inequality $|\boldsymbol{x}^\top \boldsymbol{y}| \leq \|\boldsymbol{x}\|_1 \|\boldsymbol{y}\|_\infty$. The first term is readily bounded by

$$I_1 = \left\|(1-\gamma)\mathbb{E}_\pi \sum_{t=T+1}^{\infty} \gamma^t \boldsymbol{\Phi}(s_t, a_t)\right\|_\infty \leq \gamma^{T+1} R \leq \gamma^T R, \tag{109}$$

where $R := \max_{s,a} \|\boldsymbol{\Phi}(s,a)\|_\infty = \max_{s,a} \max_{\|\boldsymbol{w}\|_1 \leq 1} |\boldsymbol{r}_{\boldsymbol{w}}(s,a)|$. Thus in order to have $I_1 \leq \varepsilon_1$ it suffices to choose $T = \lceil \log(\varepsilon_1/R) / \log(\gamma) \rceil$.

For the second term I_2, we make use of Hoeffding's inequality

Lemma D.1 (Hoeffding)**.** *Consider iid random variables $X_1, \ldots, X_N$ with $X_i \in [a, b]$ and let $\bar{X}_N := \frac{1}{N}(X_1 + \ldots + X_N)$. Then,*

$$\Pr\left(\left|\bar{X}_N - \mathbb{E}X_i\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2N}{(b-a)^2}\right). \tag{110}$$

Since $\left|\left(\hat{\boldsymbol{\sigma}}^{\mathrm{E}}_{\mathcal{D}}\right)_j\right| \leq R$ and $\mathbb{E}\left(\hat{\boldsymbol{\sigma}}^{\mathrm{E}}_{\mathcal{D}}\right)_j = \left(\boldsymbol{\sigma}^{\mathrm{E}}_{T}\right)_j$ for all $j = 1, \ldots, d$, we get by Hoeffding's inequality

$$\Pr\left(\left|\left(\hat{\boldsymbol{\sigma}}^{\mathrm{E}}_{\mathcal{D}} - \boldsymbol{\sigma}^{\mathrm{E}}_{T}\right)_j\right| \geq \varepsilon_2\right) \leq 2\exp\left(-\frac{\varepsilon_2^2 N}{2R^2}\right), \quad j = 1, \ldots, d. \tag{111}$$

Using the union bound

$$\begin{aligned}\Pr\left(I_2 < \epsilon_2\right) &= 1 - \Pr\left(I_2 \geq \epsilon_2\right) \\ &\geq 1 - \left(\Pr\left(\left|\left(\hat{\boldsymbol{\sigma}}^{\mathrm{E}}_{\mathcal{D}} - \boldsymbol{\sigma}^{\mathrm{E}}_{T}\right)_1\right| \geq \varepsilon_2\right) + \ldots + \Pr\left(\left|\left(\hat{\boldsymbol{\sigma}}^{\mathrm{E}}_{\mathcal{D}} - \boldsymbol{\sigma}^{\mathrm{E}}_{T}\right)_d\right| \geq \varepsilon_2\right)\right),\end{aligned} \tag{112}$$

yields $I_2 \leq \varepsilon_2$ with probability at least $1 - \delta$ when $N \geq \log\left(2d/\delta\right) 2R^2/\varepsilon_2^2$. Thus choosing $N = \left\lceil \log\left(2d/\delta\right) 32R^2/\varepsilon^2 \right\rceil$ and $T = \lceil \log(\varepsilon/(8R))/\log(\gamma) \rceil$ it holds with probability at least $1 - \delta$

$$\max_{\boldsymbol{\mu} \in \mathcal{F}, \|\boldsymbol{w}\|_1 \leq 1} \left| L(\boldsymbol{\mu}, \boldsymbol{w}) - \hat{L}(\boldsymbol{\mu}, \boldsymbol{w}) \right| \leq \varepsilon/2. \tag{113}$$

Step 2:
Next we use the fact that for two real-valued functions g and h we always have $|\max g - \max h| \leq \max|g - h|$ and similarly $|\min g - \min h| \leq \max|g - h|$. Therefore, it holds for all $\hat{\boldsymbol{r}} \in \mathsf{IRL}_{\mathcal{R},\mathcal{F}}(\hat{\boldsymbol{\mu}}^{\mathrm{E}}_{\mathcal{D}})$ with probability at least $1 - \delta$:

$$\begin{aligned}\left|L(\boldsymbol{\mu}^{\mathrm{E}}, \boldsymbol{r}^{\mathrm{E}}) - \hat{L}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}})\right| &= \left|\min_{\|\boldsymbol{w}\|_1 \leq 1} \max_{\boldsymbol{\mu} \in \mathcal{F}} L(\boldsymbol{\mu}, \boldsymbol{w}) - \min_{\|\boldsymbol{w}\|_1 \leq 1} \max_{\boldsymbol{\mu} \in \mathcal{F}} \hat{L}(\boldsymbol{\mu}, \boldsymbol{w})\right| \\ &\leq \max_{\|\boldsymbol{w}\|_1 \leq 1} \left|\max_{\boldsymbol{\mu} \in \mathcal{F}} L(\boldsymbol{\mu}, \boldsymbol{w}) - \max_{\boldsymbol{\mu} \in \mathcal{F}} \hat{L}(\boldsymbol{\mu}, \boldsymbol{w})\right| \\ &\leq \max_{\boldsymbol{\mu} \in \mathcal{F}, \|\boldsymbol{w}\|_1 \leq 1} \left|\hat{L}(\boldsymbol{\mu}, \boldsymbol{w}) - L(\boldsymbol{\mu}, \boldsymbol{w})\right| \leq \varepsilon/2.\end{aligned} \tag{114}$$

Due to Assumption 4.1 there is $\boldsymbol{w}^*$ with $\|\boldsymbol{w}^*\|_1 \leq 1$ such that $\boldsymbol{r}_{\boldsymbol{w}^*} = \boldsymbol{r}^{\mathrm{E}}$. Furthermore, let $\hat{\boldsymbol{w}}$ with $\|\hat{\boldsymbol{w}}\|_1 \leq 1$ be such that $\boldsymbol{r}_{\hat{\boldsymbol{w}}} \in \mathsf{IRL}_{\mathcal{R},\mathcal{F}}(\hat{\boldsymbol{\mu}}^{\mathrm{E}}_{\mathcal{D}})$. Then, L has a saddle-point in $(\boldsymbol{\mu}^{\mathrm{E}}, \boldsymbol{w}^*)$ and $\hat{L}$ in $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{w}})$. Choosing N and T as in (113), it holds with probability at least $1 - \delta$

$$\begin{aligned}J(\boldsymbol{\mu}^{\mathrm{E}}, \boldsymbol{r}^{\mathrm{E}}) - J(\hat{\boldsymbol{\mu}}, \boldsymbol{r}^{\mathrm{E}}) &= L(\boldsymbol{\mu}^{\mathrm{E}}, \boldsymbol{w}^*) - L(\hat{\boldsymbol{\mu}}, \boldsymbol{w}^*) \\ &\overset{(i)}{\leq} L(\boldsymbol{\mu}^{\mathrm{E}}, \boldsymbol{w}^*) - \hat{L}(\hat{\boldsymbol{\mu}}, \boldsymbol{w}^*) + \varepsilon/2 \\ &\overset{(ii)}{\leq} L(\boldsymbol{\mu}^{\mathrm{E}}, \boldsymbol{r}^{\mathrm{E}}) - \hat{L}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{w}}) + \varepsilon/2 \\ &\overset{(iii)}{\leq} \varepsilon,\end{aligned} \tag{115}$$

where (i) is a consequence of (113), inequality (ii) follows since $\hat{L}$ has a saddle point in $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{w}})$, and (iii) from (114). This proves the first result in (105). The second follows analogously from a similar series of inequalities

$$\begin{aligned}J(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}}) - J(\boldsymbol{\mu}^{\mathrm{E}}, \hat{\boldsymbol{r}}) &= L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{w}}) - L(\boldsymbol{\mu}^{\mathrm{E}}, \hat{\boldsymbol{w}}) \\ &\leq \hat{L}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{w}}) - L(\boldsymbol{\mu}^{\mathrm{E}}, \hat{\boldsymbol{w}}) + \varepsilon/2 \\ &\leq \hat{L}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{w}}) - L(\boldsymbol{\mu}^{\mathrm{E}}, \boldsymbol{w}^*) + \varepsilon/2 \\ &\leq \varepsilon.\end{aligned} \tag{116}$$

Step 3:
To prove the inequality (106), we use the fact that since f is L-strongly convex, $J(\boldsymbol{\mu}, \boldsymbol{r}^{\mathrm{E}}) = {\boldsymbol{r}^{\mathrm{E}}}^{\top} \boldsymbol{\mu} - f(\boldsymbol{\mu}_k)$ is L-strongly concave in $\boldsymbol{\mu}$ i.e. it holds

$$J(\boldsymbol{\mu}, \boldsymbol{r}^{\mathrm{E}}) \leq J(\boldsymbol{\mu}^{\mathrm{E}}, \boldsymbol{r}^{\mathrm{E}}) + \nabla_{\boldsymbol{\mu}} J(\boldsymbol{\mu}^{\mathrm{E}}, \boldsymbol{r}^{\mathrm{E}})^{\top} \left(\boldsymbol{\mu} - \boldsymbol{\mu}^{\mathrm{E}}\right) - \frac{L}{2} \left\|\boldsymbol{\mu} - \boldsymbol{\mu}^{\mathrm{E}}\right\|^2. \tag{117}$$

By optimality of $\boldsymbol{\mu}^{\mathrm{E}}$ it holds $\nabla_{\boldsymbol{\mu}} J(\boldsymbol{\mu}^{\mathrm{E}}, \boldsymbol{r}^{\mathrm{E}})^{\top} \left(\boldsymbol{\mu} - \boldsymbol{\mu}^{\mathrm{E}}\right) \leq 0$ for all $\boldsymbol{\mu} \in \mathcal{F}$. Rearranging terms yields and taking the square root yields

$$\left\|\boldsymbol{\mu} - \boldsymbol{\mu}^{\mathrm{E}}\right\| \leq \sqrt{\frac{2}{L} \left(J(\boldsymbol{\mu}^{\mathrm{E}}, \boldsymbol{r}^{\mathrm{E}}) - J(\boldsymbol{\mu}, \boldsymbol{r}^{\mathrm{E}})\right)}, \quad \forall \mu \in \mathcal{F}. \tag{118}$$

Combining this with (105) yields the desired result

$$\left\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^{\mathrm{E}}\right\| \leq \sqrt{\frac{2\varepsilon}{L}}. \tag{119}$$

Although $f(\boldsymbol{\mu}) = -\beta\, \mathbb{E}_{(s,a)\sim\boldsymbol{\mu}} \left[H\left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s)\right)\right]$ is in general not strongly convex (although it is strictly convex), we can make use of the following result for entropy-regularized MDPs.

Lemma D.2 ((Mei et al., 2020)). *For any occupancy measure $\boldsymbol{\mu} \in \mathcal{F}$ it holds*

$$J(\boldsymbol{\mu}^*, \boldsymbol{r}) - J(\boldsymbol{\mu}, \boldsymbol{r}) = \beta \sum_s \boldsymbol{\nu}(s) D_{KL}\left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s) || \boldsymbol{\pi}^{\boldsymbol{\mu}^*}(\cdot|s)\right),$$

where $\boldsymbol{\mu}^* \in \mathsf{RL}_{\mathcal{F}}(\boldsymbol{r})$ and $\boldsymbol{\nu}(s) = \sum_a \boldsymbol{\mu}(s,a)$ is the state occupancy measure.

Making use of the following lower bound on the KL-divergence (Cover, 1999)

$$D_{\mathrm{KL}}(\boldsymbol{q}||\boldsymbol{p}) \leq \frac{1}{2} \|\boldsymbol{q} - \boldsymbol{p}\|_1^2, \tag{120}$$

we arrive at

$$\begin{aligned}
J(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}}) - J(\boldsymbol{\mu}^{\mathrm{E}}, \hat{\boldsymbol{r}}) &\geq \beta \sum_s \boldsymbol{\nu}^{\mathrm{E}}(s) D_{\mathrm{KL}}\left(\boldsymbol{\pi}^{\boldsymbol{\mu}^{\mathrm{E}}}(\cdot|s) || \boldsymbol{\pi}^{\hat{\boldsymbol{\mu}}}(\cdot|s)\right) \\
&\geq \frac{\beta}{2} \sum_s \boldsymbol{\nu}^{\mathrm{E}}(s) \left\|\boldsymbol{\pi}^{\boldsymbol{\mu}^{\mathrm{E}}}(\cdot|s) - \boldsymbol{\pi}^{\hat{\boldsymbol{\mu}}}(\cdot|s)\right\|_1^2 \\
&\geq \frac{\beta}{2} \left(\sum_s \boldsymbol{\nu}^{\mathrm{E}}(s) \left\|\boldsymbol{\pi}^{\boldsymbol{\mu}^{\mathrm{E}}}(\cdot|s) - \boldsymbol{\pi}^{\hat{\boldsymbol{\mu}}}(\cdot|s)\right\|_1\right)^2 \\
&= \frac{\beta}{2} \left(\mathbb{E}_{(s,a)\sim\boldsymbol{\mu}^{\mathrm{E}}} \left[\left\|\boldsymbol{\pi}^{\boldsymbol{\mu}^{\mathrm{E}}}(\cdot|s) - \boldsymbol{\pi}^{\hat{\boldsymbol{\mu}}}(\cdot|s)\right\|_1\right]\right)^2,
\end{aligned} \tag{121}$$

where the last inequality follows from Jensen's inequality. Making use of (105) and rearranging terms yields

$$\mathbb{E}_{(s,a)\sim\boldsymbol{\mu}^{\mathrm{E}}} \left[\left\|\boldsymbol{\pi}^{\boldsymbol{\mu}^{\mathrm{E}}}(\cdot|s) - \boldsymbol{\pi}^{\hat{\boldsymbol{\mu}}}(\cdot|s)\right\|_1\right] \leq \sqrt{\frac{2\varepsilon}{\beta}}. \tag{122}$$

□

E. Algorithm

We present the policy based algorithm for entropy regularization as used in the experiments. To this end, recall the min-max problem (23)

$$\min_{\boldsymbol{r}\in\mathcal{R}} \max_{\boldsymbol{\mu}\in\mathcal{F}} \; \boldsymbol{r}^{\top} \left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\mathcal{D}}^{\mathrm{E}}\right) - f(\boldsymbol{\mu}), \tag{123}$$

and the entropy regularization $f(\boldsymbol{\mu}) = -\beta\, \mathbb{E}_{(s,a)\sim\boldsymbol{\mu}}\left[H\left(\boldsymbol{\pi}^{\boldsymbol{\mu}}(\cdot|s)\right)\right]$. Applying Proposition 3.4 this is equivalent to

$$\min_{\boldsymbol{\xi}\geq\mathbf{0},\boldsymbol{r}\in\mathcal{R}} \max_{\boldsymbol{\mu}\in\mathcal{M}} \boldsymbol{r}^\top\left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\mathcal{D}}^{\mathrm{E}}\right) - f(\boldsymbol{\mu}) + \boldsymbol{\xi}^\top\left(\boldsymbol{b} - \boldsymbol{\Psi}^\top\boldsymbol{\mu}\right). \tag{124}$$

Using the one-to-one mapping between policies and occupancy measures and the linear reward class

$$\mathcal{R} := \left\{\boldsymbol{r}_{\boldsymbol{w}} = \boldsymbol{\Phi}\boldsymbol{w} : \boldsymbol{\Phi}\in\mathbb{R}^{nm\times d}, \|\boldsymbol{w}\|\leq c\right\}, \tag{125}$$

this can be rewritten as

$$\begin{aligned}
&\min_{\boldsymbol{\xi}\geq\mathbf{0},\boldsymbol{r}\in\mathcal{R}} \max_{\boldsymbol{\pi}\in\Pi} \boldsymbol{r}^\top\left(\boldsymbol{\mu}^{\boldsymbol{\pi}} - \hat{\boldsymbol{\mu}}_{\mathcal{D}}^{\mathrm{E}}\right) - f(\boldsymbol{\mu}^{\boldsymbol{\pi}}) + \boldsymbol{\xi}^\top\left(\boldsymbol{b} - \boldsymbol{\Psi}^\top\boldsymbol{\mu}^{\boldsymbol{\pi}}\right) \\
=&\min_{\boldsymbol{\xi}\geq\mathbf{0},\|\boldsymbol{w}\|\leq c} \max_{\boldsymbol{\pi}\in\Pi} \boldsymbol{w}^\top\boldsymbol{\Phi}^\top\left(\boldsymbol{\mu}^{\boldsymbol{\pi}} - \hat{\boldsymbol{\mu}}_{\mathcal{D}}^{\mathrm{E}}\right) - f(\boldsymbol{\mu}^{\boldsymbol{\pi}}) + \boldsymbol{\xi}^\top\left(\boldsymbol{b} - \boldsymbol{\Psi}^\top\boldsymbol{\mu}^{\boldsymbol{\pi}}\right) \\
=&\min_{\boldsymbol{\xi}\geq\mathbf{0},\|\boldsymbol{w}\|\leq c} \max_{\boldsymbol{\pi}\in\Pi} L_{\mathcal{D}}(\boldsymbol{\pi},\boldsymbol{w},\boldsymbol{\xi}),
\end{aligned} \tag{126}$$

with the Lagrangian $L_{\mathcal{D}}(\boldsymbol{\pi},\boldsymbol{w},\boldsymbol{\xi}) := \boldsymbol{w}^\top\boldsymbol{\Phi}^\top\left(\boldsymbol{\mu}^{\boldsymbol{\pi}} - \hat{\boldsymbol{\mu}}_{\mathcal{D}}^{\mathrm{E}}\right) - f(\boldsymbol{\mu}^{\boldsymbol{\pi}}) + \boldsymbol{\xi}^\top\left(\boldsymbol{b} - \boldsymbol{\Psi}^\top\boldsymbol{\mu}^{\boldsymbol{\pi}}\right)$. Motivated by recent advances in min-max optimization (Daskalakis & Panageas, 2018), we suggest to use a gradient descent-ascent method, where policy and reward are updated simultaneously within a single optimization loop.

Algorithm 1 Gradient Descent Ascent for Constrained Entropy-Regularized IRL

Input: Expert data $\mathcal{D}$, learning rate η.
Initialize $\boldsymbol{\pi}\in\Pi, \boldsymbol{w} = \mathbf{0}, \boldsymbol{\xi} = \mathbf{0}$.
for $i = 1$ **to** N_{episodes} **do**
  $\boldsymbol{r} \leftarrow \boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{\Psi}\boldsymbol{\xi}$
  $\boldsymbol{\pi} \leftarrow \mathrm{NPG}(\boldsymbol{\pi}, \boldsymbol{r}, \eta)$
  $\boldsymbol{w} \leftarrow P_{B_c}\left(\boldsymbol{w} - \eta\nabla_{\boldsymbol{w}} L_{\mathcal{D}}(\boldsymbol{\pi},\boldsymbol{w},\boldsymbol{\xi})\right)$
  $\boldsymbol{\xi} \leftarrow P_{[\mathbf{0},\boldsymbol{\infty})}\left(\boldsymbol{\xi} - \eta\nabla_{\boldsymbol{\xi}} L_{\mathcal{D}}(\boldsymbol{\pi},\boldsymbol{w},\boldsymbol{\xi})\right)$
end for
Return: $\boldsymbol{w}, \boldsymbol{\pi}$.

Here, $\mathrm{NPG}(\boldsymbol{\pi}, \boldsymbol{r}, \eta)$ refers to a single entropy-regularized Natural Policy Gradient step with softmax parametrization of the policy (Cen et al., 2022). Note that for the step size $\eta = (1-\gamma)/\beta$ this policy gradient step is completely equivalent to soft policy iteration (Haarnoja et al., 2018). Furthermore, the gradients for $\boldsymbol{w}$ and $\boldsymbol{\xi}$ are given by

$$\begin{aligned}
\nabla_{\boldsymbol{w}} L_{\mathcal{D}}(\boldsymbol{\pi},\boldsymbol{w},\boldsymbol{\xi}) &= \boldsymbol{\Phi}^\top\left(\boldsymbol{\mu}^{\boldsymbol{\pi}} - \hat{\boldsymbol{\mu}}_{\mathcal{D}}^{\mathrm{E}}\right) \\
\nabla_{\boldsymbol{\xi}} L_{\mathcal{D}}(\boldsymbol{\pi},\boldsymbol{w},\boldsymbol{\xi}) &= \left(\boldsymbol{b} - \boldsymbol{\Psi}^\top\boldsymbol{\mu}^{\boldsymbol{\pi}}\right).
\end{aligned} \tag{127}$$

Moreover, P_{B_c} denotes the projection onto the ball $B_c := \{\boldsymbol{w} : \|\boldsymbol{w}\|\leq c\}$ and $P_{[\mathbf{0},\boldsymbol{\infty})}$ the trivial projection onto the non-negative orthant. For the 1-norm projection we use the efficient projection algorithm by Duchi et al. (2008) and its implementation by Ong & Lustig (2019). Algorithm 1 has been shown to provably converge for CMDPs (Ying et al., 2022; Ding et al., 2022) (with known reward) and IRL (Zeng et al., 2022) (without safety constraints).

Finally, we note that in our code we also provide a gradient descent ascent primal dual method that directly optimizes the min-max problem (124) in the occupancy measure space. Since the problem is convex-concave such algorithms enjoy provable convergence guarantees (Daskalakis & Panageas, 2018; Nemirovski, 2004), and work for arbitrary convex regularizations f. As the focus of this paper is not on algorithmic convergence and Algorithm 1 was both – easier to tune and converged more efficiently – we used Algorithm 1 throughout the presented experiments.