Towards the Transferability of Rewards Recovered via Regularized Inverse Reinforcement Learning

Andreas Schlaginhaufen
SYCAMORE, EPFL
andreas.schlaginhaufen@epfl.ch

Maryam Kamgarpour
SYCAMORE, EPFL
maryam.kamgarpour@epfl.ch

Abstract

Inverse reinforcement learning (IRL) aims to infer a reward from expert demonstrations, motivated by the idea that the reward, rather than the policy, is the most succinct and transferable description of a task [Ng et al., 2000]. However, the reward corresponding to an optimal policy is not unique, making it unclear if an IRL-learned reward is transferable to new transition laws in the sense that its optimal policy aligns with the optimal policy corresponding to the expert's true reward. Past work has addressed this problem only under the assumption of full access to the expert's policy, guaranteeing transferability when learning from two experts with the same reward but different transition laws that satisfy a specific rank condition [Rolland et al., 2022]. In this work, we show that the conditions developed under full access to the expert's policy cannot guarantee transferability in the more practical scenario where we have access only to demonstrations of the expert. Instead of a binary rank condition, we propose principal angles as a more refined measure of similarity and dissimilarity between transition laws. Based on this, we then establish two key results: 1) a sufficient condition for transferability to any transition laws when learning from at least two experts with sufficiently different transition laws, and 2) a sufficient condition for transferability to local changes in the transition law when learning from a single expert. Furthermore, we also provide a probably approximately correct (PAC) algorithm and an end-to-end analysis for learning transferable rewards from demonstrations of multiple experts.

1 Introduction

Reinforcement learning (RL) has achieved remarkable success in various domains such as robotics [Hwangbo et al., 2019], autonomous driving [Lu et al., 2023], or fine-tuning of large language models [Stiennon et al., 2020]. Despite these advances, a key challenge lies in designing appropriate reward functions that reflect the desired outcomes and align with human values. Misaligned rewards can lead to suboptimal behaviors [Ngo et al., 2022], undermining the potential benefits of RL in practical scenarios. Inverse reinforcement learning (IRL), also known as inverse optimal control [Kalman, 1964] or structural estimation [Rust, 1994], addresses this problem by inferring a reward from demonstrations of an expert acting optimally in a Markov decision process (MDP).

Compared to behavioral cloning [Pomerleau, 1988], which directly fits a policy to the expert's demonstrations, IRL is believed to provide a more transferable description of the expert's task [Ng et al., 2000], as recovering the expert's underlying reward would enable us to train a policy in a new environment with different dynamics. However, it is also known that the reward corresponding to some optimal policy is not unique [Ng et al., 1999], making it difficult to recover the expert's true underlying reward. This raises the question: *Is a reward recovered via IRL transferable to a new environment in the sense that its optimal policy aligns with the expert's true reward?* For example, in

Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2024).

autonomous driving, could we effectively reuse a reward learned from demonstrations of one car in a given city to train or fine-tune a policy for another car in another city?

Ensuring transferability is challenging, as neither the optimal policy corresponding to a reward nor the reward corresponding to an optimal policy is unique. This leads to trivial solutions to the IRL problem, such as constant rewards that make all policies optimal. Common approaches to address this challenge include characterizing the entire set of rewards for which the expert is optimal [Metelli et al., 2021], or assuming the expert is optimal with respect to an entropy regularized RL problem [Ziebart, 2010], leading to many popular IRL and imitation learning algorithms [Ho and Ermon, 2016, Fu et al., 2017, Garg et al., 2021]. Entropy regularization results in a unique and more uniform optimal policy, serving as a model for the expert's bounded rationality [Ortega et al., 2015].

In the entropy-regularized setting, several recent works study the set of rewards for which a given expert policy is optimal. In particular, Cao et al. [2021], Skalse et al. [2023] show that under entropy regularization, the expert's reward can be identified up to so-called potential shaping transformations [Ng et al., 1999]. The authors of [Schlaginhaufen and Kamgarpour, 2023] extend this result to more general steep regularization. Furthermore, they show that to guarantee transferability to any transition law, the expert's reward needs to be identified up to a constant. The latter can be achieved either by restricting the reward class, e.g., to state-only rewards [Amin et al., 2017], or by learning from multiple experts with the same reward but different transition laws, given that a specific rank condition is satisfied [Cao et al., 2021, Rolland et al., 2022]. However, the above results cannot be applied directly in practice, as they rely on having full access to the experts' policies, whereas in practice, we typically only have a finite set of demonstrations available.

Contributions We consider the framework of regularized IRL [Jeon et al., 2021] and address the transferability of rewards recovered from a finite set of expert demonstrations.

- We define a novel notion of transferability (Definition 3.1), to address the practical limitation of not having perfect access to the experts' policies. Furthermore, we show that when learning from finite data, the conditions developed under full access to the experts' policies are not sufficient to guarantee transferability (Example 3.2).
- Instead of a binary rank condition, we propose to use principal angles to characterize the similarity and dissimilarity between transition laws (Definition 3.7). Based on these principal angles, we then establish two key transferability results: 1) a guarantee for transferability to any transition laws when learning from at least two experts with sufficiently different transition laws (Theorem 3.9), and 2) a guarantee for transferability to local changes in the transition law when learning from a single expert (Theorem 3.10).
- Assuming oracle access to a probably approximately correct (PAC) algorithm for the forward RL problem, we provide a PAC algorithm for the IRL problem, which in $\mathcal{O}(K^2/\hat{\varepsilon}^2)$ steps recovers a reward for which, with high probability, all K experts are $\hat{\varepsilon}$-optimal (Theorem 4.1). Together with our results on transferability, this establishes end-to-end guarantees for learning transferable rewards from a finite set of expert demonstrations.
- We experimentally validate our results in a gridworld environment (Section 5).

2 Background

Notation Given x and y in some Euclidean vector space $\mathcal{V}$, we denote the p-norm by $\|x\|_p$, the orthogonal projection onto a closed convex set $\mathcal{X} \subset \mathcal{V}$ by $\Pi_{\mathcal{X}}(x) = \arg\min_{y \in \mathcal{X}} \|x - y\|_2$, and the standard inner product by $\langle x, y \rangle$. For a linear operator A, we denote its image and rank by $\operatorname{im} A$ and $\operatorname{rank} A$, respectively. Given two sets $\mathcal{X}$ and $\mathcal{Y}$, we denote $\mathcal{X} + \mathcal{Y}$ for their Minkowski sum and $\mathcal{Y}^{\mathcal{X}}$ for the set of all functions mapping from $\mathcal{X}$ to $\mathcal{Y}$. Additionally, we denote $\Delta_{\mathcal{X}}$ for the probability simplex over $\mathcal{X}$ and $\mathbb{1}$ for the indicator function. The interior $\operatorname{int} \mathcal{X}$, the relative interior $\operatorname{relint} \mathcal{X}$, the relative boundary $\operatorname{relbd} \mathcal{X}$, and the convex hull $\operatorname{conv} \mathcal{X}$ of some set $\mathcal{X}$ are defined in Appendix A, along with an overview of all other notations.

Regularized MDPs We consider a regularized MDP [Geist et al., 2019] defined by a tuple $(\mathcal{S}, \mathcal{A}, P, \nu_0, r, \gamma, h)$. Here, $\mathcal{S}$ and $\mathcal{A}$ represent finite state and action spaces with $|\mathcal{S}|, |\mathcal{A}| > 1$, $\nu_0 \in \Delta_{\mathcal{S}}$ the initial state distribution, $P \in \Delta_{\mathcal{S}}^{\mathcal{S} \times \mathcal{A}}$ the transition law, $r \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ the reward, and $\gamma \in (0, 1)$ the discount factor. Furthermore, $h : \mathcal{X} \to \mathbb{R}$ is a strictly convex regularizer that is defined on a closed convex set $\mathcal{X} \subseteq \mathbb{R}^{\mathcal{A}}$ with $\operatorname{relint} \Delta_{\mathcal{A}} \subseteq \operatorname{int} \mathcal{X}$. The goal is to find a Markov policy

$\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ maximizing the regularized objective $\mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t\left[r(s_t, a_t) - h\left(\pi(\cdot|s_t)\right)\right]\right]$. Following the classical linear programming approach to MDPs [Puterman, 2014], this can be cast equivalently as the convex optimization problem

$$\max_{\mu \in \mathcal{M}} J(r, \mu), \quad \text{with} \quad J(r, \mu) := \langle r, \mu \rangle - \bar{h}(\mu), \tag{O-RL}$$

where $\mathcal{M}$ denotes the set of occupancy measures, $\mu^\pi(s, a) := (1-\gamma)\mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}(s_t = s, a_t = a)\right]$, and we have $\bar{h}(\mu) := \mathbb{E}_{(s,a)\sim\mu}\left[h(\pi^\mu(\cdot|s))\right]$, with π^μ being the policy corresponding to μ (see Appendix A). The set of occupancy measures is characterized by the Bellman flow constraints

$$\mathcal{M} = \left\{\mu \in \mathbb{R}_+^{\mathcal{S}\times\mathcal{A}} : (E - \gamma P)^\top \mu = (1-\gamma)\nu_0\right\} \subseteq \Delta_{\mathcal{S}\times\mathcal{A}},$$

where $E : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $P : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ are the linear operators mapping $v \in \mathbb{R}^{\mathcal{S}}$ to $(Ev)(s, a) = v(s)$ and $(Pv)(s, a) = \sum_{s'} P(s'|s, a)v(s')$, respectively.

Due to the strict convexity of h, the regularized MDP problem has a unique optimal policy [Geist et al., 2019], hence guaranteeing the uniqueness of the optimal occupancy measure in (O-RL). In addition, we assume that the gradients of h become unbounded towards the relative boundary of the simplex as detailed in Assumption 2.1 below.

Assumption 2.1 (Steep regularization)**.** Suppose that $h : \mathcal{X} \to \mathbb{R}$ is differentiable in $\operatorname{int}\mathcal{X}$ and that $\lim_{l\to\infty} \|\nabla h(p_l)\|_2 = \infty$ if $(p_l)_{l\in\mathbb{N}}$ is a sequence in $\operatorname{int}\mathcal{X}$ converging to a point $p \in \operatorname{relbd}\Delta_{\mathcal{A}}$.

Assumption 2.1 ensures that the optimal policy is non-vanishing, and together with Assumption 2.2 below, we also have that the optimal occupancy measure is non-vanishing.

Assumption 2.2 (Exploration)**.** Let $\nu(s) := \sum_a \mu(s, a) \geq \nu_{\text{min}} > 0$ for any $s \in \mathcal{S}$ and $\mu \in \mathcal{M}$.

One way to guarantee Assumption 2.2 is to impose a lower bound on the initial state distribution ν_0. In the following, it will be convenient to denote the optimal solution to (O-RL) for the reward r as

$$\mathsf{RL}(r) := \operatorname*{argmax}_{\mu\in\mathcal{M}} J(r, \mu),$$

and the suboptimality of some occupancy measure μ for the reward r as

$$\ell(r, \mu) := \max_{\mu'\in\mathcal{M}} J(r, \mu') - J(r, \mu). \tag{1}$$

That is, $\mu = \mathsf{RL}(r)$ if and only if $\ell(r, \mu) = 0$.

Remark 2.3. As we aim to analyze the transferability of rewards to new transition laws $P \in \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$, it will often be useful to explicitly specify the dependency on P. We do so by adding a subscript – e.g. we write $\mathcal{M}_P$, RL_P, and ℓ_P. However, for better readability, we drop these subscripts whenever there is no potential for confusion.

Inverse reinforcement learning Given a dataset of trajectories sampled from an expert μ^{E} that is optimal for some reward r^{E}, the goal in IRL is to recover a reward $\hat{r} \in \mathcal{R}$, within a predefined reward class $\mathcal{R} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, such that the expert is optimal for $\hat{r}$. That is, ideally, we aim to find a reward in the feasible reward set

$$\mathsf{IRL}(\mu^{\mathsf{E}}) := \left\{r \in \mathcal{R} : \mu^{\mathsf{E}} \in \mathsf{RL}(r)\right\}. \tag{2}$$

However, since we don't have direct access to the expert's policy but only to a finite set of demonstrations, the best we can hope for is an algorithm that with high probability outputs a reward $\hat{r} \in \mathcal{R}$ such that $\ell(\hat{r}, \mu^{\mathsf{E}})$ is small – i.e. an algorithm that is PAC [Syed and Schapire, 2007].

Reward equivalence The reward corresponding to an optimal occupancy is not unique. For example, all rewards in the affine subspace $r + \mathcal{U}$, where $\mathcal{U} := \operatorname{im}(E - \gamma P)$ is the subspace of so-called potential shaping transformations, correspond to the same optimal occupancy measure [Ng et al., 1999]. Hence, it is often convenient to consider these rewards as equivalent [Kim et al., 2021] and to measure distances between rewards in the resulting quotient space. Given a linear subspace $\mathcal{V} \subset \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, the quotient space $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{V}$ is the set of all equivalence classes $[r]_{\mathcal{V}} := \left\{r' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : r' - r \in \mathcal{V}\right\}$. We endow $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{V}$ with the quotient norm $\|[r]_{\mathcal{V}}\|_2 := \min_{v\in\mathcal{V}} \|r + v\|_2 = \|\Pi_{\mathcal{V}^\perp} r\|_2$ and we say that r and r' are close in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{V}$ if $\|[r]_{\mathcal{V}} - [r']_{\mathcal{V}}\|_2$ is small. Moreover, the expert's reward is said to be identifiable up to some equivalence class $[\cdot]_{\mathcal{V}}$ if $\mathsf{IRL}(\mu^{\mathsf{E}}) \subseteq [r^{\mathsf{E}}]_{\mathcal{V}}$. In this paper, we will consider the equivalence relations induced by constant shifts, i.e., $\mathcal{V} = \mathbf{1} := \left\{r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : r(s, a) = c \in \mathbb{R}\right\}$, and by the aforementioned potential shaping transformations, i.e., $\mathcal{V} = \mathcal{U}$. Note that since $\mathbf{1}$ is a subspace of $\mathcal{U}$ and $\mathcal{U}$ is $|\mathcal{S}|$-dimensional, $[r]_{\mathbf{1}}$ is a strict subset of $[r]_{\mathcal{U}}$ whenever $|\mathcal{S}| > 1$.

3 Transferability

In this section, we present our main results on transferability in IRL. To this end, we first introduce the problem of learning ε-transferable rewards from multiple experts acting in different environments.

3.1 Problem formulation

Let $\mathcal{R} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ be a compact reward class, and suppose we are given access to K expert data sets,

$$\mathcal{D}_k^{\mathsf{E}} = \left\{\left(s_0^{k,i}, a_0^{k,i}, \dots, s_{H^{\mathsf{E}}-1}^{k,i}, a_{H^{\mathsf{E}}-1}^{k,i}\right)\right\}_{i=0}^{N^{\mathsf{E}}-1}, \quad k = 0, \dots, K-1,$$

consisting of trajectories sampled independently from the experts $\mu_{P^0}^{\mathsf{E}}, \dots, \mu_{P^{K-1}}^{\mathsf{E}}$. Each expert is optimal for the same unrevealed reward $r^{\mathsf{E}} \in \mathcal{R}$, but under different transition laws, $P^0, \dots, P^{K-1}$. Our goal is to recover a reward $\hat{r} \in \mathcal{R}$ that is transferable across a set of transition laws $\mathcal{P} \subseteq \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$. Specifically, the optimal occupancy measure corresponding to $\hat{r}$ should remain approximately optimal for r^{E} under every transition law in $\mathcal{P}$. This yields the following definition of ε-transferability.

Definition 3.1 (ε-transferability)**.** Fix some $\varepsilon > 0$. We say that $\hat{r}$ is ε-transferable to some set of transition laws $\mathcal{P} \subseteq \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$ if $\ell_P(r^{\mathsf{E}}, \mathsf{RL}_P(\hat{r})) \leq \varepsilon$ for all $P \in \mathcal{P}$. We say that $\hat{r}$ is exactly transferable to $\mathcal{P}$ if it is ε-transferable to $\mathcal{P}$ with $\varepsilon = 0$.

The error margin of ε is crucial, as exact transferability is unrealistic when learning from finite expert data. Moreover, note that Definition 3.1 is a definition of uniform transferability, as it requires $\hat{r}$ to be ε-transferable to any $P \in \mathcal{P}$ with the same fixed ε. In the following, we will analyze the transferability of a reward $\hat{r}$ for which all experts are $\hat{\varepsilon}$-optimal for some $\hat{\varepsilon} > 0$. That is,

$$\ell_{P^k}(\hat{r}, \mu_{P^k}^{\mathsf{E}}) \leq \hat{\varepsilon}, \quad k = 0, \dots, K-1. \tag{3}$$

In particular, we aim to establish appropriate conditions for choosing $\hat{\varepsilon}$ so as to guarantee ε-transferability to some set of transition laws $\mathcal{P}$. In Section 4, we will then provide an IRL algorithm that, with high probability, outputs a reward $\hat{r}$ such that (3) holds.

3.2 Related work

Most previous work has focused on reward identifiability. For a single expert, Cao et al. [2021], Skalse et al. [2023], Schlaginhaufen and Kamgarpour [2023] show that under Assumption 2.1 (steepness) the feasible reward set (2) can be expressed as

$$\mathsf{IRL}(\mu^{\mathsf{E}}) = \left(\nabla\bar{h}(\mu^{\mathsf{E}}) + \mathcal{U}\right) \cap \mathcal{R} = [r^{\mathsf{E}}]_{\mathcal{U}} \cap \mathcal{R}. \tag{4}$$

In other words, steepness ensures that the expert's reward is identifiable up to potential shaping. To identify the reward up to a constant, we can either restrict the reward class, e.g. to state-only rewards as explored by Amin et al. [2017], or learn from multiple experts [Cao et al., 2021, Rolland et al., 2022]. In particular, when we are given access to two experts, $\mu_{P^0}^{\mathsf{E}}$ and $\mu_{P^1}^{\mathsf{E}}$, we can identify the experts' reward up to the intersection

$$\mathsf{IRL}_{P^0}(\mu_{P^0}^{\mathsf{E}}) \cap \mathsf{IRL}_{P^1}(\mu_{P^1}^{\mathsf{E}}) = [r^{\mathsf{E}}]_{\mathcal{U}_{P^0}} \cap [r^{\mathsf{E}}]_{\mathcal{U}_{P^1}} \cap \mathcal{R} = \left(r^{\mathsf{E}} + \mathcal{U}_{P^0} \cap \mathcal{U}_{P^1}\right) \cap \mathcal{R}.$$

That is, for the unrestricted reward class, $\mathcal{R} = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, the reward is identifiable up to a constant if and only if $\mathcal{U}_{P^0} \cap \mathcal{U}_{P^1} = \mathbf{1}$. Or equivalently, if and only if the rank condition

$$\operatorname{rank}\left(\begin{bmatrix} E - \gamma P^0, & E - \gamma P^1 \end{bmatrix}\right) = 2|\mathcal{S}| - 1, \tag{5}$$

is satisfied [Rolland et al., 2022]. Moreover, Schlaginhaufen and Kamgarpour [2023] show that identifying the expert's reward up to a constant is a necessary and sufficient condition for exact transferability to any set $\mathcal{P} \subseteq \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$ whose interior, $\operatorname{int}\mathcal{P}$, is non-empty and open (with respect to the subspace topology on $\Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$ [Bourbaki, 1966]).

Limitations The above results assume perfect access to the expert's policy, which isn't realistic. In practice, we can only learn a reward for which the experts are approximately optimal. In Example 3.2 below, we show that under approximate optimality of the experts, the learned reward can perform very poorly in a new environment, even if the rank condition in Equation (5) is satisfied.

Example 3.2. We consider a two-state, two-action MDP with $\mathcal{S} = \mathcal{A} = \{0, 1\}$, uniform initial state distribution, discount rate $\gamma = 0.9$, and Shannon entropy regularization $h = -\mathcal{H}$ (see Appendix C). Suppose the expert reward is $r^{\mathsf{E}}(s,a) = \mathbb{1}\{s = 1\}$ and consider the transition laws, P^0 and P^1, defined by $P^0(0|s,a) = 1$ and $P^1(0|s,a) = \beta \cdot \mathbb{1}\{s=0, a=0\}$ for some $\beta \in (0,1)$. Also, consider the two experts $\mu^{\mathsf{E}}_{P^0} = \mathsf{RL}_{P^0}(r^{\mathsf{E}})$ and $\mu^{\mathsf{E}}_{P^1} = \mathsf{RL}_{P^1}(r^{\mathsf{E}})$, and suppose we recovered the reward $\hat{r}(s,a) = -r^{\mathsf{E}}$. Then, as detailed in Appendix E, the following holds: 1) We have $\ell_{P^0}(\hat{r}, \mu^{\mathsf{E}}_{P^0}) = 0$ and $\ell_{P^1}(\hat{r}, \mu^{\mathsf{E}}_{P^1}) = \mathcal{O}(\beta)$. That is, for small β, the reward $\hat{r}$ is a good solution to the IRL problem, as both experts are approximately optimal under $\hat{r}$. 2) The rank condition (5) between P^0 and P^1 is satisfied for any $\beta \in (0,1)$. 3) For a new transition law P defined by $P(0|s,a) = \mathbb{1}\{s=1, a=0\}$, we have $\ell_P(r^{\mathsf{E}}, \mathsf{RL}_P(\hat{r})) \approx 4.81$, i.e. $\mathsf{RL}_P(\hat{r})$ performs poorly under the experts' reward.

3.3 Theoretical insights

To establish a sufficient condition for ε-transferability, our goal is to bound the suboptimality of an optimal occupancy measure, $\mathsf{RL}(r)$, for some reward r', in terms of reward distances measured in the quotient space $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{U}$. To this end, we first establish the relationship between the suboptimality in Equation (1) and the Bregman divergence corresponding to the occupancy measure regularization.

Bregman divergences The Bregman divergence [Teboulle, 1992] associated to $\bar{h}$ is defined as

$$D_{\bar{h}}(\mu, \mu') = \bar{h}(\mu) - \bar{h}(\mu') - \langle \nabla \bar{h}(\mu'), \mu - \mu' \rangle.$$

Proposition 3.3. *Under Assumptions 2.1 and 2.2, we have* $\ell(r', \mu) = D_{\bar{h}}(\mu, \mathsf{RL}(r'))$ *for any* $\mu \in \mathcal{M}$.

Proposition 3.3 above demonstrates that the suboptimality of an occupancy measure μ for the reward r' coincides with the Bregman divergence between μ and the optimal occupancy measure under r'. This generalizes [Mei et al., 2020, Lemma 26] from entropy regularization to any steeply regularized MDP. The proof is presented in Appendix D.6.

Reward approximation Next, we show that under strong convexity and local Lipschitz gradients, the Bregman divergence between two optimal occupancy measures is bounded in terms of reward distances in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{U}$.

Assumption 3.4 (Regularity)**.** Suppose the following holds:

a) The regularizer $\bar{h}$ is η-strongly convex over the set of occupancy measures $\mathcal{M}$. That is, we have

$$\bar{h}(\mu') \geq \bar{h}(\mu) + \langle \nabla \bar{h}(\mu), \mu' - \mu \rangle + \frac{\eta}{2} \left\| \mu' - \mu \right\|_2^2, \quad \forall \mu, \mu' \in \mathcal{M}.$$

b) The gradient $\nabla \bar{h}$ is locally Lipschitz continuous over $\operatorname{relint} \mathcal{M}$. That is, for any closed convex subset $\mathcal{K} \subset \operatorname{relint} \mathcal{M}$ there exists $L_{\mathcal{K}} > 0$ such that

$$\left\| \nabla \bar{h}(\mu) - \nabla \bar{h}(\mu') \right\|_2 \leq L_{\mathcal{K}} \left\| \mu - \mu' \right\|_2, \quad \forall \mu, \mu' \in \mathcal{K}.$$

We will show later that Assumption 3.4 is met for Shannon and Tsallis entropy regularization (see Proposition D.8). Under the above assumption, the following lemma establishes the desired upper and lower bound on the Bregman divergence between two optimal occupancy measures with respect to reward distances measured in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{U}$.

Lemma 3.5. *Suppose Assumptions 2.1,2.2, and 3.4 hold, and let* $r, r' \in \mathcal{R}$. *Then, we have*

$$\frac{\sigma_{\mathcal{R}}}{2} \left\| [r]_{\mathcal{U}} - [r']_{\mathcal{U}} \right\|_2^2 \leq \ell(r', \mathsf{RL}(r)) = D_{\bar{h}}\left(\mathsf{RL}(r), \mathsf{RL}(r')\right) \leq \frac{1}{2\eta} \left\| [r]_{\mathcal{U}} - [r']_{\mathcal{U}} \right\|_2^2, \tag{6}$$

for some problem-dependent constant $\sigma_{\mathcal{R}} > 0$.

Remark 3.6. The proof of Lemma 3.5 hinges on the duality between equivalence classes of rewards and optimal occupancy measures (see Appendix B). The main idea is to leverage duality of Bregman divergences, and a dual smoothness and strong convexity result in Proposition D.7. A key challenge arises because, by Assumption 2.1, the regularizer cannot be globally smooth. This results in a problem-dependent dual strong convexity constant $\sigma_{\mathcal{R}}$ [Goebel and Rockafellar, 2008]. In Proposition D.8, we will provide a lower bound on $\sigma_{\mathcal{R}}$ for the specific choices of Shannon and Tsallis entropy regularization. For more details, we refer to the full proof in Appendix D.6.

The above lemma has two key implications: First, the lower bound in (6) implies that if we recover a reward $\hat{r}$ for which all experts are approximately optimal, then the distance between $\hat{r}$ and r^{E} can be bounded in the quotient spaces $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{U}_{P^k}$. Second, the upper bound shows that to control the performance of $\mathsf{RL}_P(\hat{r})$ in a new environment P, we need to tightly bound the distance between $\hat{r}$ and r^{E} in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{U}_P$. As distances in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{U}_P$ are bounded by distances in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathbf{1}$, this can be achieved by bounding the distance between $\hat{r}$ and r^{E} in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathbf{1}$. However, revisiting Example 3.2 in light of Lemma 3.5 shows that even though $\hat{r}$ and r^{E} are close in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{U}_{P^k}$, this does not guarantee their proximity in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathbf{1}$ and $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{U}_P$.

Example 3.2 (continued)**.** Recall the definition $\mathcal{U}_{P^k} = \mathrm{im}(E - \gamma P^k)$. Given that in Example 3.2 we have $\ell_{P^0}(\hat{r}, \mu^{\mathsf{E}}_{P^0}) = 0$ and $\ell_{P^1}(\hat{r}, \mu^{\mathsf{E}}_{P^1}) = \mathcal{O}(\beta)$, Lemma 3.5 ensures that $\hat{r}$ and r^{E} coincide in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{U}_{P^0}$, and for small β, they are close in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{U}_{P^1}$. However, as illustrated in Figure1(a) this doesn't ensure that $\hat{r}$ and r^{E} are close in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathbf{1}$ and $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{U}_P$. In particular, it can be computed that $\left\|[\hat{r}]_{\mathcal{U}_P} - [r^{\mathsf{E}}]_{\mathcal{U}_P}\right\|_2 \approx 1.51$, which by Lemma 3.5 explains the poor transferability to P.

3.4 Sufficient conditions for transferability

With Lemma 3.5 in place, we are set to present our results on ε-transferability. Example 3.2 indicates that a sufficient condition for learning transferable rewards from experts ($K = 2$) should not rely solely on the binary rank condition (5), which only checks if $\mathcal{U}_{P^0} \cap \mathcal{U}_{P^1} = \mathbf{1}$. Instead, we should consider the relative orientation between $\mathcal{U}_{P^0}$ and $\mathcal{U}_{P^1}$. To formalize this, we need to introduce the concept of principal angles between linear subspaces, as outlined in Definition 3.7 below.

Definition 3.7 (Principal angles [Galántai, 2013])**.** Let $\mathcal{V}, \mathcal{W} \subseteq \mathbb{R}^n$ be two subspaces of dimension $m \le n$. The principal angles $0 \le \theta_1(\mathcal{V}, \mathcal{W}) \le \ldots \le \theta_m(\mathcal{V}, \mathcal{W}) =: \theta_{\max}(\mathcal{V}, \mathcal{W}) \le \pi/2$ between $\mathcal{V}$ and $\mathcal{W}$ are defined recursively via

$$\cos(\theta_i(\mathcal{V}, \mathcal{W})) = \max_{v\in\mathcal{V}, w\in\mathcal{W}} \langle v, w\rangle \text{ s.t. } \|v\|_2 = \|w\|_2 = 1,\ \langle v, v_j\rangle = \langle w, w_j\rangle = 0,\ j = 1, \ldots, i-1,$$

where v_j, w_j are the maximizers corresponding to the angle θ_j. For two transition laws P, P', we define $\theta_i(P, P') := \theta_i(\mathcal{U}_P, \mathcal{U}_{P'})$ and refer to $\theta_i(P, P')$ as the i-th principal angles between P and P'.

Principal angles are the natural generalization of angles between two lines or planes to higher dimensional subspaces. For principal angles between transition laws, we have the following proposition.

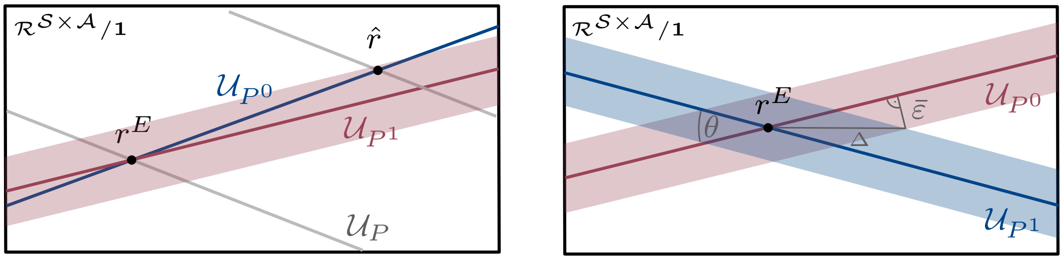
Proposition 3.8. *Let $P, P' \in \Delta^{\mathcal{S}}_{\mathcal{S}\times\mathcal{A}}$ and $H_\gamma = 1/(1-\gamma)$. Then, we have $\theta_1(P, P') = 0$ and $\sin(\theta_{\max}(P, P')) \le \gamma H_\gamma \sqrt{|\mathcal{S}|/|\mathcal{A}|}\, \|P - P'\|$, where $\|\cdot\|$ denotes the spectral norm.*

The proof can be found in Appendix D.7. The above result shows that while the first principal angle between two transition laws is always zero, all principal angles are small if the transition laws are close to one another. In Example 3.2, we have $\|P^0 - P^1\| = \mathcal{O}(\beta)$, indicating that the second and in this case maximal principal angle is small when β is small (see Appendix E). The following result shows that when learning from two experts, the transferability error is directly controlled by the second principal angle between the experts' transition laws.

Theorem 3.9. *Let $K = 2$, $\theta_2(P^0, P^1) > 0$, and suppose that Assumptions 2.1,2.2, and 3.4 hold. If $\ell_{P^k}(\hat{r}, \mu^{\mathsf{E}}_{P^k}) \le \hat{\varepsilon}$ for $k = 0, 1$, then $\hat{r}$ is ε-transferable to $\mathcal{P} = \Delta^{\mathcal{S}\times\mathcal{A}}_{\mathcal{S}}$ with*

$$\varepsilon = \hat{\varepsilon}/\left[\eta\sigma_{\mathcal{R}} \sin\left(\theta_2(P^0, P^1)/2\right)^2\right].$$

Sketch of proof. The main idea of the proof is illustrated in Figure 1(b). First, it follows from Lemma 3.5 that $\hat{r}$ and r^{E} are $\bar{\varepsilon} = \sqrt{2\hat{\varepsilon}/\sigma_{\mathcal{R}}}$-close in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{U}_{P^k}$ for $k = 0, 1$, respectively. From Figure 1(b) we see – using basic trigonometry – that this implies that $\hat{r}$ and r^{E} are at least $\Delta = \bar{\varepsilon}/\sin(\theta/2)$-close in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathbf{1}$. As shown in the full proof in Appendix F, the relevant angle, θ, is the second principal angle $\theta_2(P^0, P^1)$. The result then follows from the upper bound in Lemma 3.5. □



(a) Rewards in Example 3.2.

(b) Proof sketch Theorem 3.9.

Figure 1: For a small β, (a) illustrates the equivalence classes $[\hat{r}]_{\mathcal{U}}$ and $[r^{\mathsf{E}}]_{\mathcal{U}}$, corresponding to the transition laws P^0, P^1, P from Example 3.2, in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathbf{1}$. The blue lines correspond to P^0, the red lines to P^1, and the gray lines to P. Furthermore, the shaded areas illustrate the approximation error around $[r^{\mathsf{E}}]_{\mathcal{U}_{P^k}}$, as guaranteed by Lemma 3.5. (b) illustrates the uncertainty set for the recovered reward when learning from two experts, as discussed in the proof sketch of Theorem 3.9.

Some observations are in order. First, the above theorem shows that the larger the second principal angle between the two experts' transition laws, the better the recovered reward transfers to a new environment. Second, observe that $\theta_2(P^0, P^1) > 0$ is equivalent to the rank condition (5), as the second principal angle between two subspaces is non-zero if and only if their intersection is at most one-dimensional. Therefore, for exact transferability, Theorem 3.9 requires the rank condition (5) to be satisfied and $\hat{\varepsilon} = 0$, recovering the results by Cao et al. [2021], Rolland et al. [2022], Schlaginhaufen and Kamgarpour [2023]. However, in contrast to past results, Theorem 3.9 applies to more realistic scenarios, where $\hat{\varepsilon}$ is merely small, not zero. Finally, we note that Theorem 3.9 can be trivially generalized to $K \geq 2$ experts by replacing $\theta_2(P^0, P^1)$ with the maximum of $\theta_2(P^k, P^l)$ over $0 \leq k \leq l \leq K-1$.

Local transferability When learning a reward $\hat{r}$ from a single expert ($K = 1$), Schlaginhaufen and Kamgarpour [2023] show that, without reducing the dimension of the reward class, $\hat{r}$ cannot be exactly transferable to any neighborhood of the expert's transition law P_0. However, Theorem 3.10 below shows that by allowing for an ε of error, we can guarantee transferability to a neighborhood of P_0.

Theorem 3.10. *Let $K = 1$, $D := \max_{r,r' \in \mathcal{R}} \|r - r'\|_2$, and suppose that Assumptions 2.1,2.2, and 3.4 hold. If $\ell_{P^0}(\hat{r}, \mu^{\mathsf{E}}) \leq \hat{\varepsilon}$, then $\hat{r}$ is ε_P-transferable to $P \in \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$ with*

$$\varepsilon_P = 2\max\left\{2\hat{\varepsilon}/\sigma_{\mathcal{R}}, D^2 \sin\left(\theta_{\max}(P^0, P)\right)^2\right\}/\eta.$$

The above theorem (which is proven in Appendix G) shows that the reward learned from a single expert transfers to transition laws that are sufficiently close to the expert's, where the closeness is measured in terms of the maximal principal angle. In other words, while a large second principal angle between two experts' transition laws, as per Theorem 3.9, ensures that the reward recovered from these two experts is transferable to arbitrary transition laws, a small largest principal angle between two transition laws ensures that a reward recovered in one environment can be successfully transferred to the other environment.

Regularizers To provide more insights about Theorems 3.9 and 3.10, we provide explicit values for the primal and dual strong convexity constants, η and $\sigma_{\mathcal{R}}$, respectively. To this end, we focus on the Shannon entropy regularization $h(p) = -\tau\mathcal{H}(p)$ and the Tsallis-1/2 entropy regularization $h(p) = -\tau\mathcal{H}_{1/2}(p)$ as defined in Appendix C. While the Shannon entropy regularization is commonly used in IRL [Ziebart, 2010, Ho and Ermon, 2016], the Tsallis-1/2 entropy is more often adopted in the multi-armed bandit literature Zimmert and Seldin [2021]. Both regularizations satisfy Assumption 2.1 as well as Assumption 3.4 with the constants detailed in Proposition D.8 in the appendix. In general, the Tsallis entropy leads to a slightly smaller strong convexity constant η, but avoids an exponential dependence on the effective horizon $H_\gamma = 1/(1-\gamma)$ in $\sigma_{\mathcal{R}}$. Below, we summarize the implications of Proposition D.8 for ε-transferability of a reward $\hat{r}$ recovered from two experts.

Corollary 3.11. *Suppose the conditions in Theorem 3.9 hold. Furthermore, let $H_\gamma := 1/(1-\gamma)$, $R := \max_{r \in \mathcal{R}} \|r\|_\infty$, $D = \max_{r,r' \in \mathcal{R}} \|r - r'\|_2$, and $\tau < 2D$. Then, for the Shannon entropy $\hat{r}$ is ε-transferable to $\mathcal{P} = \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$ with*

$$\varepsilon = \frac{H_\gamma^2 D |\mathcal{S}||\mathcal{A}|^{2+H_\gamma} \exp\left(\frac{2RH_\gamma}{\tau}\right)}{\nu_{\mathit{min}}^2 \tau \left(1 - \frac{\tau}{2D}\right) \sin\left(\theta_2(P^0, P^1)/2\right)^2} \hat{\varepsilon},$$

and for the Tsallis entropy with

$$\varepsilon = \frac{2\sqrt{2} H_\gamma^5 D |\mathcal{S}||\mathcal{A}|^2 \left(2R/\tau + 3\sqrt{|\mathcal{A}|}\right)^3}{\nu_{\mathit{min}}^2 \tau \left(1 - \frac{\tau}{2D}\right) \sin\left(\theta_2(P^0, P^1)/2\right)^2} \hat{\varepsilon}.$$

We observe that transferability generally becomes more challenging with decreasing regularization parameter τ, i.e. if the expert's policy becomes more deterministic. Furthermore, we see that it is easier to recover a transferable reward in a Tsallis entropy-regularized MDP. Corollary 3.11 also shows that the constant between ε and $\hat{\varepsilon}$ tends to be large, meaning that we need to recover a reward for which the experts are $\hat{\varepsilon}$-optimal with a very small $\hat{\varepsilon}$ to guarantee ε-transferability for a reasonable ε. However, it's important to note that our results provide sufficient conditions for the worst case, and it remains for future work to determine under what conditions these constants can be improved.

Remark 3.12. Our results in this section, especially Proposition 3.3 and Lemma 3.5, are critically relying on the steepness of the regularization (Assumption 2.1), which is essential to ensure that the expert's reward can be identified up to the equivalence class of potential shaping transformations. Although we can still upper bound the suboptimality $\ell(r^{\mathsf{E}}, \mathsf{RL}(\hat{r}))$ in terms of the distance between $\hat{r}$ to r^{E} in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{U}$ without this assumption (see Proposition D.9), we no longer have a lower bound as in Lemma 3.5, which is essential for establishing closeness of $\hat{r}$ and r^{E} in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{U}$. Hence, we expect it to be difficult to obtain guarantees similar to those in Theorem 3.9 and 3.10 for the unregularized setting, without either reducing the dimension of the reward class [Amin et al., 2017] or making specific assumptions about the feasible reward sets [Metelli et al., 2021, Assumption 4.1].

4 Algorithm

To provide end-to-end guarantees for recovering transferable rewards from a finite set of expert demonstrations, we analyze the convergence and sample complexity (in terms of expert demonstrations) of an algorithm for recovering a reward for which all K experts are approximately optimal. To this end, we focus on the reward class $\mathcal{R} = \left\{r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \|r\|_1 \leq 1\right\}$. Furthermore, we assume oracle access to a (ε, δ)-PAC algorithm for the forward problem (O-RL). That is, a polynomial-time algorithm, $\mathsf{A}^{\varepsilon,\delta}$, that outputs a policy $\pi = \mathsf{A}^{\varepsilon,\delta}(r)$ such that with probability at least $1-\delta$ it holds that $\ell(r, \mu^{\pi}) \leq \varepsilon$ (see e.g. [Lan, 2023] for a specific example). The key idea of our meta-algorithm is to learn a reward minimizing the sum of the suboptimalities of the K experts $\mu^{\mathsf{E}}_{P^0}, \dots, \mu^{\mathsf{E}}_{P^{K-1}}$. This leads us to the following multi-expert IRL problem

$$\min_{r\in\mathcal{R}} \sum_{k=0}^{K-1} \ell_{P^k}(r, \hat{\mu}_{\mathcal{D}^{\mathsf{E}}_k}), \tag{O-IRL}$$

where $\hat{\mu}_{\mathcal{D}^{\mathsf{E}}_k}(s,a) := (1-\gamma)/N^{\mathsf{E}} \sum_{i=0}^{N^{\mathsf{E}}-1} \sum_{t=0}^{H^{\mathsf{E}}-1} \gamma^t \mathbb{1}\{s^{k,i}_t = s, a^{k,i}_t = a\}$ is an empirical expert occupancy measure. To solve Problem (O-IRL), we propose the projected gradient descent scheme as detailed in Algorithm 1 below, where $\texttt{rollout}_{P^k}(\pi, N, H)$ samples N independent trajectories of length H from policy π. Using a stochastic online gradient descent analysis, Theorem 4.1 shows that any PAC algorithm for the forward problem yields a PAC algorithm for the inverse problem.

Algorithm 1: Multi-expert IRL

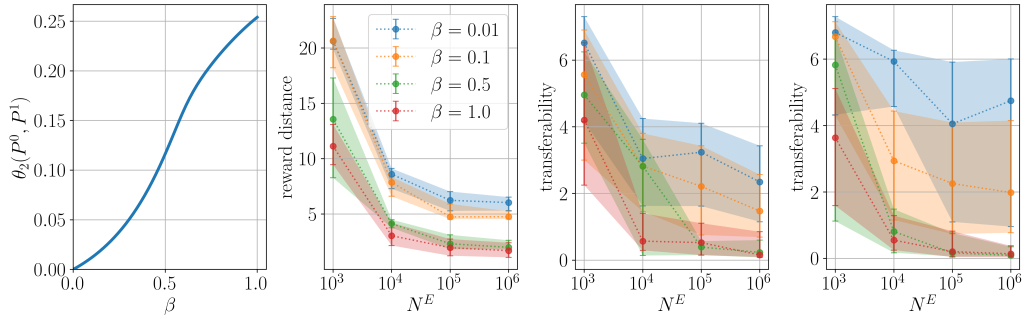
```
Input: $\alpha, T, \{\mathcal{D}^{\mathsf{E}}_k\}_{k=0}^{K-1}, N, H, \varepsilon_{\text{opt}}, \delta_{\text{opt}}$.
Initialize: $\pi \in \Delta^{\mathcal{S}}_{\mathcal{A}}$ and $r \in \mathcal{R}$ arbitrarily.
for $i = 0, \dots, T-2$ do
    for $k = 0, \dots, K-1$ do
        $\pi_{k,t} = \mathsf{A}^{\varepsilon_{\text{opt}},\delta_{\text{opt}}}_{P^k}(r_t)$                    // Forward RL.
        $\mathcal{D}_{k,t} = \texttt{rollout}_{P^k}(\pi_{k,t}, N, H)$
    end
    $g_t = \sum_{k=0}^{K-1} \left(\hat{\mu}_{\mathcal{D}_{k,t}} - \hat{\mu}_{\mathcal{D}^{\mathsf{E}}_k}\right)$
    $r_{t+1} = \Pi_{\mathcal{R}}(r_t - \alpha g_t)$                              // Reward step.
end
Return: $\hat{r} := \frac{1}{T}\sum_{t=0}^{T-1} r_t$.
```

Theorem 4.1. *Suppose that* $N^{E} = \Omega\big(K \log(|\mathcal{S}||\mathcal{A}|/\hat{\delta})/\hat{\varepsilon}^2\big)$ *and* $H^{E} = \Omega\big(\log(K/\hat{\varepsilon})/\log(1/\gamma)\big)$. *Running Algorithm 1 for* $T = \Omega\big(K^2/\hat{\varepsilon}^2\big)$ *iterations with step-size* $\alpha = 1/(K\sqrt{T})$, *where* $\delta_{opt} = \mathcal{O}\big(\hat{\delta}\hat{\varepsilon}^2/K^3\big)$, $\varepsilon_{opt} = \mathcal{O}(\hat{\varepsilon}/K)$, $N = \Omega\big(K\log(K|\mathcal{S}||\mathcal{A}|/(\hat{\delta}\hat{\varepsilon}))/\hat{\varepsilon}^2\big)$, *and* $H = H^{E}$, *it holds with probability at least* $1-\hat{\delta}$ *that* $\ell_{P^k}(\hat{r}, \mu^{E}_{P^k}) \leq \hat{\varepsilon}$, *for* $k = 0, \dots, K-1$.

The above result generalizes [Syed and Schapire, 2007, Theorem 2] by considering multiple experts and by proving convergence in terms of the expert suboptimality. We refer to Appendix H for the proof and the precise constants. Theorem 4.1 shows that with $\Omega(K/\hat{\varepsilon}^2)$ demonstrations of each expert, we recover in $\Omega(K^2/\hat{\varepsilon}^2)$ steps of Algorithm 1 a reward $\hat{r}$ for which all experts are $\hat{\varepsilon}$-optimal. Together with Theorem 3.9 and 3.10, this provides a bound on the sample and time complexity of recovering in ε-transferable rewards in regularized IRL.

5 Experiments

To validate our results experimentally, we adopt a stochastic variant of the `WindyGridworld` environment [Sutton and Barto, 2018]. In this environment, the agent moves to the intended grid cell with a probability of $(1-\beta)$ and is pushed one step further in the direction of the wind with a probability of β. Using Algorithm 1, we recover a reward $\hat{r}$ from demonstrations of two experts, both exposed to the same wind strength β but different wind directions – North and East. The experiments are repeated for a varying number of expert demonstrations $N^{\mathsf{E}} \in \{10^3, 10^4, 10^5, 10^6\}$ and wind strengths $\beta \in \{0.01, 0.1, 0.5, 1.0\}$. We then test the transferability to two different environments: one with South wind, P^{South}, and a zero-wind environment with cyclically shifted actions, P^{Shifted}. Figure 2(a) shows that the second principal angle between the two experts' transition laws P_0 and P_1 increases with increasing wind strength. Moreover, Figure 2(b)-(d) show that both the closeness between $\hat{r}$ and r^{E} in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathbf{1}$ and the transferability to P^{South} and P^{Shifted} improve with a larger second principal angle, as expected from Theorem 3.9. For a more detailed discussion of the experiments we refer to Appendix I.



(*a*) (*b*) (*c*) (*d*)

Figure 2: (*a*) shows the second principal angle between the experts, for varying wind strength β. (*b*) shows the distance between $\hat{r}$ and r^{E} in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathbf{1}$ for a varying number of expert demonstrations N^{E} and wind strength β. (*c*) and (*d*) show the transferability to P^{South} and P^{Shifted} in terms of $\ell_{P^{\text{South}}}(r^{\mathsf{E}}, \mathsf{RL}_{P^{\text{South}}}(\hat{r}))$ and $\ell_{P^{\text{Shifted}}}(r^{\mathsf{E}}, \mathsf{RL}_{P^{\text{Shifted}}}(\hat{r}))$, respectively. The circles indicate the median and the shaded areas the 0.2 and 0.8 quantiles over the 10 independent realizations.

6 Conclusion

Summary In this paper, we investigated the transferability of rewards in regularized IRL. We showed that the conditions established under full access to the experts' policies do not guarantee transferability when learning a reward from a finite set of expert demonstrations. To address this issue, we proposed using principal angles as a more refined measure of the similarity and dissimilarity of transition laws. Assuming a strongly convex and locally smooth regularization, we then showed that if we recover a reward for which at least two experts are nearly optimal, and their environments are sufficiently different in terms of the second principal angle between their transition laws, then the recovered reward is universally transferable. Furthermore, we showed that if two environments are sufficiently similar in terms of the maximal principal angle between their transition laws, rewards learned in one environment can be effectively transferred to the other environment. Additionally, we provided explicit constants for the Shannon and Tsallis entropy, as well as a PAC algorithm for recovering a reward for which all experts are approximately optimal. As a result, we established end-to-end guarantees for learning transferable rewards in regularized IRL. Additionally, we experimentally validated our results through gridworld experiments.

Limitations and future work Our results provide only sufficient conditions for transferability. It would be valuable to investigate necessary conditions to check whether our bounds are tight. Furthermore, extending our analysis to lower-dimensional reward classes could reduce the complexity of learning transferable rewards. Although our paper focuses on discrete state and action spaces, an exciting avenue for future research would be to extend our results to continuous state and action spaces, which are more commonly encountered in practice. Finally, as our work is mainly theoretical, experimental validation on real-world applications could provide valuable insight into the practical aspects and challenges of transferability.

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

References

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

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

A. Beck. *First-order methods in optimization*. SIAM, 2017.

N. Bourbaki. *Elements of mathematics: General topology*. Hermann, 1966.

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

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

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

C. Dann, C.-Y. Wei, and J. Zimmert. Best of both worlds policy optimization. *arXiv preprint arXiv:2302.09408*, 2023.

Z. Drmac. On principal angles between subspaces of euclidean space. *SIAM Journal on Matrix Analysis and Applications*, 22(1):173–194, 2000. doi: 10.1137/S0895479897320824. URL `https://doi.org/10.1137/S0895479897320824`.

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

A. Galántai. *Projectors and projection methods*, volume 6. Springer Science & Business Media, 2013.

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

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

R. Goebel and R. T. Rockafellar. Local strong convexity and local lipschitz continuity of the gradient of convex functions. *Journal of Convex Analysis*, 15(2):263, 2008.

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

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459. URL `http://www.jstor.org/stable/2282952`.

J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26):eaau5872, 2019.

W. Jeon, C.-Y. Su, P. Barde, T. Doan, D. Nowrouzezahrai, and J. Pineau. Regularized inverse reinforcement learning. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=HgLO8yalfwc`.

S. Ji-Guang. Perturbation of angles between linear subspaces. *Journal of Computational Mathematics*, pages 58–61, 1987.

R. E. Kalman. When Is a Linear Control System Optimal? *Journal of Basic Engineering*, 86(1): 51–60, 03 1964. ISSN 0021-9223. doi: 10.1115/1.3653115. URL `https://doi.org/10.1115/1.3653115`.

K. Kim, S. Garg, K. Shiragur, and S. Ermon. Reward identification in inverse reinforcement learning. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5496–5505. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/kim21c.html`.

A. V. Knyazev and M. E. Argentati. Principal angles between subspaces in an a-based scalar product: algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23(6):2008–2040, 2002.

G. Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1):1059–1106, 2023.

Y. Lu, J. Fu, G. Tucker, X. Pan, E. Bronstein, R. Roelofs, B. Sapp, B. White, A. Faust, S. Whiteson, et al. Imitation is not enough: Robustifying imitation with reinforcement learning for challenging driving scenarios. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7553–7560. IEEE, 2023.

J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International conference on machine learning*, pages 6820–6829. PMLR, 2020.

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

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

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

R. Ngo, L. Chan, and S. Mindermann. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

P. A. Ortega, D. A. Braun, J. Dyer, K.-E. Kim, and N. Tishby. Information-theoretic bounded rationality. *arXiv preprint arXiv:1512.06789*, 2015.

R. Ouhamma and M. Kamgarpour. Learning nash equilibria in zero-sum markov games: A single time-scale algorithm under weak reachability. *arXiv preprint arXiv:2312.08008*, 2023.

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

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

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

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970. ISBN 9781400873173. doi: doi:10.1515/9781400873173. URL `https://doi.org/10.1515/9781400873173`.

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

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

J. Rust. Structural estimation of markov decision processes. *Handbook of econometrics*, 4:3081–3143, 1994.

A. Schlaginhaufen and M. Kamgarpour. Identifiability and generalizability in constrained inverse reinforcement learning. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 30224–30251. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/schlaginhaufen23a.html`.

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

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

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

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

M. Teboulle. Entropic proximal mappings with applications to nonlinear programming. *Mathematics of Operations Research*, 17(3):670–690, 1992.

B. D. Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University, 2010.

J. Zimmert and Y. Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *The Journal of Machine Learning Research*, 22(1):1310–1358, 2021.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

Appendix

Table of Contents

A Notations

Overview Here, we provide an overview of some of the most important notations. However, every notation is defined when it is introduced as well.

Table 1: Notations.

$\mathcal{Y}^{\mathcal{X}}$	:=	set of functions $\mathcal{X} \to \mathcal{Y}$
$\Delta_{\mathcal{X}}$	:=	probability simplex over some discrete set $\mathcal{X}$
$\mathcal{M}$	:=	set of feasible occupancy measures
$\mathcal{R}$	:=	$\left\{r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \lVert r\rVert_1 \leq 1\right\}$, reward class
R	:=	$\max_{r\in\mathcal{R}} \lVert r\rVert_\infty$, reward bound
D	:=	$\max_{r,r\in\mathcal{R}} \lVert r - r'\rVert_2$, diameter of the reward class
H_γ	:=	$1/(1-\gamma)$, effective horizon
$\nu(s)$	:=	$\sum_a \mu(s,a)$
$\pi^\mu(a\mid s)$	:=	$\begin{cases} \mu(s,a)/\sum_{a'} \mu(s,a') = \mu(s,a)/\nu(s) & ,\nu(s) > 0 \\ 1/\lvert\mathcal{A}\rvert \quad \text{(arbitrary)} & \text{, otherwise} \end{cases}$
π_s	:=	$\pi(\cdot\mid s)$
$\bar{h}(\mu)$	:=	$\mathbb{E}_{(s,a)\sim\mu}\left[h(\pi^\mu_s)\right]$
$J(r,\mu)$	:=	$\langle r,\mu\rangle - \bar{h}(\mu)$
$\ell(r,\mu')$	:=	$\max_{\mu\in\mathcal{M}} J(r,\mu) - J(r,\mu')$
$\mathsf{RL}(r)$	:=	$\operatorname{argmax}_{\mu\in\mathcal{M}} J(r,\mu)$, optimal occupancy measure for r
$\mathsf{IRL}(\mu^{\mathsf{E}})$	:=	$\left\{r \in \mathcal{R} : \mu^{\mathsf{E}} = \mathsf{RL}(r)\right\}$, feasible reward set for μ^{E}
$\mathsf{A}^{\varepsilon,\delta}(r)$	:=	PAC RL algorithm outputting some ε-optimal policy with probability at least $1-\delta$
$\hat{\mu}_{\mathcal{D}}$	:=	$\frac{1-\gamma}{N}\sum_{i=0}^{N-1}\sum_{t=0}^{T-1}\gamma^t \mathbb{1}\left\{s^i_t = s, a^i_t = a\right\}$, where $\mathcal{D} = \left\{\left(s^i_0, a^i_0, \ldots, s^i_{H-1}, a^i_{H-1}\right)\right\}_{i=0}^{N-1}$
E	:=	linear operator $\mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ defined by $(Ef)(s,a) = f(s)$ for $f \in \mathbb{R}^{\mathcal{S}}$
P	:=	linear operator $\mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ defined by $(Pf)(s,a) = \sum_{s'} P(s'\mid s,a)f(s')$ for $f \in \mathbb{R}^{\mathcal{S}}$
$\mathrm{im}\,(A)$	:=	image of some linear operator A
$\mathcal{U}$	:=	$\mathrm{im}(E-\gamma P) \subset \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, potential shaping subspace
$\mathbf{1}$	:=	$\{f = \text{constant}\} \subset \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, constant subspace
$\mathcal{V}^\perp$	:=	orthogonal complement of some linear subspace $\mathcal{V}$

Additional definitions In the following, we briefly recall some additional definitions. To this end, we denote $\mathcal{B}(x,r) := \{x \in \mathbb{R}^n : \lVert x\rVert_2 < r\}$ for an open ball of radius r with center x.

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

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

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

$$\operatorname{aff}\mathcal{X} := \{\theta_1 x_1 + \ldots + \theta_k x_k : x_1, \ldots, x_k \in \mathcal{X}, \theta_1 + \ldots + \theta_k = 1\}\,.$$

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

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

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

$$\operatorname{relbd}\mathcal{X} := \mathcal{X} \setminus \operatorname{relint}\mathcal{X}.$$

Definition A.5 (Convex hull)**.** The convex hull of a set $\mathcal{X} \subseteq \mathbb{R}^n$ is defined as

$$\operatorname{conv}\mathcal{X} := \{\theta_1 x_1 + \ldots + \theta_k x_k : x_1, \ldots, x_k \in \mathcal{X}, \theta_1 + \ldots + \theta_k = 1, \theta_i \geq 0, i = 1, \ldots, k\}\,.$$

B Conjugate duality in regularized IRL

In this section, we first recall some background from convex analysis and then briefly discuss the duality between reward equivalence classes and optimal occupancy measures.

Definitions We recall a few definitions related to convex functions. In convex analysis it is standard to consider extended real value functions $f : \mathbb{R}^n \to \overline{\mathbb{R}} := [-\infty, \infty]$, where convex functions defined on some subset $\mathcal{X} \subset \mathbb{R}^n$ are extended over the entire space by setting their value to $+\infty$ outside of their domain. The effective domain is defined as $\operatorname{dom} f := \{x : f(x) < \infty\}$, and a convex function f is said to be proper if $f > -\infty$ and $\operatorname{dom} f \neq \emptyset$. Furthermore, f is referred to as closed if its epigraph $\{(x, y) : x \in \operatorname{dom} f, y \geq f(x)\}$ is a closed set.[1] In particular, f is closed if it is continuous on $\operatorname{dom} f$ and $\operatorname{dom} f$ is a closed set [Boyd and Vandenberghe, 2004]. Lastly, we recall two key concepts in convex analysis – the subdifferential and the convex conjugate of some convex function.

Definition B.1 (Subdifferential)**.** A subgradient of $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ at some point $x \in \mathbb{R}^n$ is a vector $g \in \mathbb{R}^n$ such that $f(x') \geq f(x) + g^\top (x - x)$ for all $x' \in \mathcal{X}$. The subdifferential $\partial f(x)$ at $x \in \mathcal{X}$ is the set of all subgradients at x, where $\partial f(x)$ is defined to be empty if $x \notin \operatorname{dom} f$.

Definition B.2 (Convex Conjugate)**.** The convex conjugate of $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is the function $f^* : \mathbb{R}^n \to \overline{\mathbb{R}}$ defined as

$$f^*(y) = \sup_x \langle y, x \rangle - f(x).$$

Key results Next, we list two key results from convex analysis.

Theorem B.3 ([Rockafellar, 1970])**.** *A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is differentiable at some point $x \in \operatorname{dom} f$ if and only if $\partial f(x)$ is singleton. In this case we have $\partial f(x) = \{\nabla f(x)\}$.*

Theorem B.4 ([Rockafellar, 1970])**.** *For any proper convex function $f^* : \mathbb{R}^n \to \overline{\mathbb{R}}$ it holds*

$$f^*(y) = \langle y, x \rangle - f(x) \quad \iff \quad y \in \partial f(x).$$

If additionally f is closed, then

$$f^*(y) = \langle y, x \rangle - f(x) \quad \iff \quad y \in \partial f(x) \quad \iff \quad x \in \partial f^*(y).$$

Duality in IRL Let $f : \mathbb{R}^{\mathcal{S} \times \mathcal{A}} \to \overline{\mathbb{R}}$ be given by $f := \bar{h} + \delta_{\mathcal{M}}$, where $\delta_{\mathcal{M}}$ is a characteristic function defined as $\delta_{\mathcal{M}}(\mu) = 0$ if $\mu \in \mathcal{M}$ and $\delta_{\mathcal{M}}(\mu) = \infty$, otherwise. Since f is closed proper convex, Theorem B.3 and B.4 imply that for a strictly convex $\bar{h}$ we have

$$\mathsf{RL}(r) = \nabla f^*(r) \quad \text{and} \quad \mathsf{IRL}(\mu) = \partial f(\mu) \cap \mathcal{R}.$$

Additionally, under Assumption 2.1 and Slater's condition, which is ensured by Assumption 2.2, we have $\partial f(\mu) = \nabla \bar{h}(\mu) + \mathcal{U}$ [Schlaginhaufen and Kamgarpour, 2023]. Hence, μ is optimal for r if and only if $r \in [\nabla \bar{h}(\mu)]_{\mathcal{U}}$. Therefore, there is a one-to-one mapping between elements of the quotient space $\mathbb{R}^{\mathcal{S} \times \mathcal{A}} / \mathcal{U}$, i.e. reward equivalence classes, and corresponding optimal occupancy measures in $\mathcal{M}$. This mapping is given by

$$\nabla f^* : \mathbb{R}^{\mathcal{S} \times \mathcal{A}} / \mathcal{U} \to \mathcal{M}, [r]_{\mathcal{U}} \mapsto \nabla f^*(r) = \mathsf{RL}(r),$$

and its inverse by

$$\partial f : \mathcal{M} \to \mathbb{R}^{\mathcal{S} \times \mathcal{A}} / \mathcal{U}, \mu \mapsto \partial f(\mu) = \nabla \bar{h}(\mu) + \mathcal{U}.$$

C Regularizers

We dedicate this section to discuss optimal policies in regularized MDPs and to recall their explicit form for Shannon and Tsallis entropy regularization.

[1] A proper convex function is closed if and only if it is lower semi-continuous [Rockafellar, 1970].

Optimal policies Throughout the appendix, it will convenient to use the notation $\pi_s := \pi(\cdot|s)$. Given some proper closed strongly convex policy regularizer h, it can be shown [Geist et al., 2019] that the optimal policy, π^*, satisfies

$$\pi_s^* = \nabla h^*(q_s^*) = \operatorname*{argmax}_{\pi_s \in \Delta_{\mathcal{A}}} \langle \pi_s, q_s^* \rangle - h(\pi_s),\ \forall s \in \mathcal{S},$$

where $q_s^* \in \mathbb{R}^{\mathcal{A}}$ is the optimal q-function defined via

$$q_s^*(a) := q^*(s,a) := \max_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \mathbb{E}_\pi \left[r(s_t,a_t) + \sum_{t \geq 1} \gamma^t \left[r(s_t,a_t) - h(\pi_s) \right] \middle| s_0 = s, a_0 = a \right].$$

Next, we discuss the explicit form of $\pi_s^* = \nabla h^*(q_s^*)$ for the specific cases of Shannon and Tsallis-$1/2$ entropy regularization.

Shannon entropy For some $\tau > 0$, we define the Shannon entropy regularizer as $h := -\tau \mathcal{H}$, where

$$\mathcal{H}(\pi_s) := -\sum_a \pi_s(a) \log \pi_s(a),$$

is the Shannon entropy satisfying $0 \leq \mathcal{H} \leq \log |\mathcal{A}|$. It can be shown that h is τ-strongly convex with respect to $\|\cdot\|_1$ [Cover, 1999], and the optimal policy satisfies [Geist et al., 2019]

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

Tsallis entropy For some parameter $\alpha \in \mathbb{R}$, the Tsallis entropy, $\mathcal{H}_\alpha$, is defined as

$$\mathcal{H}_\alpha(\pi_s) := \frac{1}{\alpha - 1} \left(1 - \sum_a \pi_s(a)^\alpha \right).$$

In the limit $\alpha \to 1$, the Tsallis entropy equals the Shannon entropy defined above. However, in this paper, we use *Tsallis entropy* to refer to the choice $\alpha = 1/2$, which is often adopted as regularization in multi-armed bandit and, more recently, policy optimization algorithms [Zimmert and Seldin, 2021, Dann et al., 2023]. In particular, we consider $h(\pi_s) = -\tau \mathcal{H}_{1/2}(\pi_s) = -2\tau \left(\sum_a \sqrt{\pi_s(a)} - 1 \right)$ for some $\tau > 0$. We have $0 \leq -h \leq 2\tau \left(\sqrt{|\mathcal{A}|} - 1 \right)$ and it can be shown that the optimal policy satisfies

$$\pi^*(a|s) = \left(\frac{\tau}{x_s - q^*(s,a)} \right)^2,$$

where x_s is a normalization parameter such that $\sum_a \pi^*(a|s) = 1$ [Zimmert and Seldin, 2021]. Furthermore, since h has diagonal Hessian $\nabla^2 h(\pi_s)_{a,a} = \tau/(2\pi_s(a)^{3/2})$, it is $\tau/2$-strongly convex with respect to $\|\cdot\|_2$ and hence also $\tau/(2|\mathcal{A}|)$-strongly convex with respect to $\|\cdot\|_1$.

D Technical Lemmas

In this section, we show several new technical results that are required for the proofs of our main theorems.

D.1 Lipschitz continuity from policies to occupancy measures

Proposition D.1. *For any $\mu_1, \mu_2 \in \mathcal{M}$, we have*

$$(1-\gamma) \|\mu_1 - \mu_2\|_1 \leq \max_s \|\pi_s^{\mu_1} - \pi_s^{\mu_2}\|_1 \leq \frac{2}{\nu_{min}} \|\mu_1 - \mu_2\|_1.$$

Proof. To show the first inequality, we decompose

$$\begin{aligned} \|\mu_1 - \mu_2\|_1 &\leq \sum_{s,a} |\nu_1(s)(\pi^{\mu_1}(a|s) - \pi^{\mu_2}(a|s))| + \sum_{s,a} |(\nu_1(s) - \nu_2(s))\pi^{\mu_2}(a|s)| \\ &\leq \max_s \|\pi_s^{\mu_1} - \pi_s^{\mu_2}\|_1 + \|\nu_1 - \nu_2\|_1, \end{aligned}$$

where we used the triangle and Hölder's inequality. From the Bellman flow constraints,

$$\nu(s) = \gamma \sum_{s',a'} P(s|s',a')\mu(s',a') + (1-\gamma)\nu_0(s),$$

it follows that

$$\begin{aligned}
\|\nu_1 - \nu_2\|_1 &= \gamma \sum_s \left| \sum_{s',a'} P(s|s',a')(\mu_1(s',a') - \mu_2(s',a')) \right| \\
&\leq \gamma \sum_{s',a'} \underbrace{\sum_s P(s|s',a')}_{=1} |\mu_1(s',a') - \mu_2(s',a')| \\
&= \gamma \|\mu_1 - \mu_2\|_1 ,
\end{aligned}$$

where we again used the triangle inequality. Hence, we have

$$\max_s \|\pi_s^{\mu_1} - \pi_s^{\mu_2}\|_1 \geq \|\mu_1 - \mu_2\|_1 - \|\nu_1 - \nu_2\|_1 \geq (1-\gamma) \|\mu_1 - \mu_2\|_1 .$$

To show the second inequality, we use the reverse triangle inequality

$$\begin{aligned}
\|\mu_1 - \mu_2\|_1 &\geq \sum_{s,a} |\nu_1(s)(\pi^{\mu_1}(a|s) - \pi^{\mu_2}(a|s))| - \sum_{s,a} |(\nu_1(s) - \nu_2(s))\pi^{\mu_2}(a|s)| \\
&= \sum_s \nu_1(s) \|\pi_s^{\mu_1} - \pi_s^{\mu_2}\|_1 - \|\nu_1 - \nu_2\|_1 , \\
&\geq \nu_{\text{min}} \max_s \|\pi_s^{\mu_1} - \pi_s^{\mu_2}\|_1 - \gamma \|\mu_1 - \mu_2\|_1 ,
\end{aligned}$$

where in the last step we used

$$\begin{aligned}
\|\nu_1 - \nu_2\|_1 &= \gamma \sum_s \left| \sum_{s',a'} P(s|s',a') (\mu_1(s',a') - \mu_2(s',a')) \right| \\
&\leq \gamma \sum_{s',a'} \underbrace{\sum_s P(s|s',a')}_{=1} |(\mu_1(s',a') - \mu_2(s',a'))| \\
&= \gamma \|\mu_1 - \mu_2\|_1 .
\end{aligned}$$

By rearranging terms we arrive at the desired inequality

$$\max_s \|\pi_s^{\mu_1} - \pi_s^{\mu_2}\|_1 \leq \frac{1+\gamma}{\nu_{\text{min}}} \|\mu_1 - \mu_2\|_1 \leq \frac{2}{\nu_{\text{min}}} \|\mu_1 - \mu_2\|_1 .$$

□

D.2 Strong convexity

The strong convexity part of Proposition D.8 follows from the following more general result.

Proposition D.2 (Strong convexity). *Let Assumption 2.2 hold. Suppose that h is η_h-strongly convex with respect to the $\|\cdot\|_1$ norm. Then, $\bar{h}$ is $\eta_h \nu_{min}/H_\gamma^2$-strongly convex with respect to $\|\cdot\|_1$.*

Proof. We need to show that for $\alpha \in (0,1)$ and $\bar{\alpha} = 1 - \alpha$ and any two $\mu_1, \mu_2 \in \mathcal{M}$ with $\mu_1 \neq \mu_2$ it holds that

$$\bar{h}(\alpha\mu_1 + \bar{\alpha}\mu_2) \leq \alpha\bar{h}(\mu_1) + \bar{\alpha}\bar{h}(\mu_2) - \frac{\alpha\bar{\alpha}\eta}{2} \|\mu_1 - \mu_2\|_1^2 ,$$

for $\eta = \eta_h \nu_{\text{min}}/H_\gamma^2$. To this end, we start similarly as in the proof of strict convexity by Schlaginhaufen and Kamgarpour [2023], but use $\nu(s) \geq \nu_{\text{min}}$ and strong convexity of h.

$$
\begin{aligned}
&\bar{h}(\alpha\mu_1 + \bar{\alpha}\mu_2) \qquad\qquad (7)\\
&= \sum_s \left(\alpha\nu_1(s) + \bar{\alpha}\nu_2(s)\right) h\left(\frac{\alpha\mu_1(s,\cdot) + \bar{\alpha}\mu_2(s,\cdot)}{\alpha\nu_1(s) + \bar{\alpha}\nu_2(s)}\right)\\
&= \sum_s \left(\alpha\nu_1(s) + \bar{\alpha}\nu_2(s)\right) h\left(\frac{\alpha\mu_1(s,\cdot)}{\alpha\nu_1(s) + \bar{\alpha}\nu_2(s)}\frac{\nu_1(s)}{\nu_1(s)} + \frac{\bar{\alpha}\mu_2(s,\cdot)}{\alpha\nu_1(s) + \bar{\alpha}\nu_2(s)}\frac{\nu_2(s)}{\nu_2(s)}\right)\\
&= \sum_s \left(\alpha\nu_1(s) + \bar{\alpha}\nu_2(s)\right) h\left(\underbrace{\frac{\alpha\nu_1(s)}{\alpha\nu_1(s) + \bar{\alpha}\nu_2(s)}}_{\beta_s}\pi_s^{\mu_1} + \underbrace{\frac{\bar{\alpha}\nu_2(s)}{\alpha\nu_1(s) + \bar{\alpha}\nu_2(s)}}_{1-\beta_s}\pi_s^{\mu_2}\right)\\
&\leq \sum_s \left(\alpha\nu_1(s) + \bar{\alpha}\nu_2(s)\right)\left(\beta_s h\left(\pi_s^{\mu_1}\right) + (1-\beta_s) h\left(\pi_s^{\mu_2}\right) - \frac{\beta_s(1-\beta_s)\eta_h}{2}\left\|\pi_s^{\mu_1} - \pi_s^{\mu_2}\right\|_1^2\right)\\
&= \sum_s \left(\alpha\nu_1(s) h\left(\pi_s^{\mu_1}\right) + \bar{\alpha}\nu_2(s) h\left(\pi_s^{\mu_2}\right) - \frac{\alpha\bar{\alpha}\eta_h}{2}\frac{\nu_1(s)\nu_2(s)}{\alpha\nu_1(s) + \bar{\alpha}\nu_2(s)}\left\|\pi_s^{\mu_1} - \pi_s^{\mu_2}\right\|_1^2\right).
\end{aligned}
$$

From here on, we use that

$$
\begin{aligned}
&\sum_s \frac{\nu_1(s)\nu_2(s)}{\alpha\nu_1(s) + \bar{\alpha}\nu_2(s)}\left\|\pi_s^{\mu_1} - \pi_s^{\mu_2}\right\|_1^2\\
&= \sum_s \underbrace{\frac{\max\{\nu_1(s), \nu_2(s)\}}{\alpha\nu_1(s) + \bar{\alpha}\nu_2(s)}}_{\geq 1}\underbrace{\min\{\nu_1(s), \nu_2(s)\}}_{\geq \nu_{\text{min}}}\left\|\pi_s^{\mu_1} - \pi_s^{\mu_2}\right\|_1^2\\
&\geq \sum_s \nu_{\text{min}}\left\|\pi_s^{\mu_1} - \pi_s^{\mu_2}\right\|_1^2\\
&\geq \nu_{\text{min}} \max_s \left\|\pi_s^{\mu_1} - \pi_s^{\mu_2}\right\|_1^2\\
&\geq \nu_{\text{min}}/H_\gamma^2 \left\|\mu_1 - \mu_2\right\|_1^2. \qquad\qquad (8)
\end{aligned}
$$

where we used Proposition D.1 in the last step. Plugging the inequality (8) back into (7) concludes the proof. □

D.3 Lipschitz gradients

In this section, we show how we can get bounds on the Lipschitz constant $L_\mathcal{K}$ from Assumption 3.4. To this end, we first need to lower bound the optimal policies.

Policy lower bounds The following proposition establishes a lower bound for optimal policies with Shannon and Tsallis entropy regularization.

Proposition D.3. *Let $H_\gamma = 1/(1-\gamma)$ and $r_{\max} := \|r\|_\infty$. Then, we have the following lower bounds:*

a) If $h = -\tau\mathcal{H}$, then $\pi^*(a|s) \geq \exp\left(-2r_{\max}H_\gamma/\tau\right)/|\mathcal{A}|^{1+H_\gamma}$.

b) If $h = -\tau\mathcal{H}_{1/2}$, then $\pi^*(a|s) \geq \left(\left(2r_{\max}/\tau + 3\sqrt{|\mathcal{A}|}\right)H_\gamma\right)^{-2}$.

Proof. Recall the formula for the optimal policies in Appendix C.

Part a): Since, $-r_{\max}H_\gamma \leq q^*(s,a) \leq \left(r_{\max} + \tau\log|\mathcal{A}|\right)H_\gamma$, it holds that

$$
\begin{aligned}
\pi^*(a|s) &\geq \frac{\exp\left(-r_{\max}H_\gamma/\tau\right)}{|\mathcal{A}|\exp\left(\left(r_{\max} + \tau\log|\mathcal{A}|\right)H_\gamma/\tau\right)}\\
&= \frac{\exp\left(-r_{\max}H_\gamma/\tau\right)}{|\mathcal{A}|^{1+H_\gamma}\exp\left(r_{\max}H_\gamma/\tau\right)} = \frac{\exp\left(-2r_{\max}H_\gamma/\tau\right)}{|\mathcal{A}|^{1+H_\gamma}}.
\end{aligned}
$$

Part b): The proof of b) is similar to [Ouhamma and Kamgarpour, 2023, Lemma 8]. However, our settings are slightly different. Recall that

$$\pi^*(a|s) = \left(\frac{\tau}{x_s - q^*(s,a)}\right)^2.$$

By Ouhamma and Kamgarpour [2023, Lemma 10] we have $\tau \leq x_s - \|q_s^*\|_\infty \leq \tau\sqrt{|\mathcal{A}|}$. Furthermore, it holds that $-r_{\max}H_\gamma \leq q^*(s,a) \leq \left(r_{\max} + 2\tau\sqrt{|\mathcal{A}|}\right) H_\gamma$. Hence, we have

$$0 < x_s - q^*(s,a) \leq \tau\sqrt{|\mathcal{A}|} + \left(r_{\max} + 2\tau\sqrt{|\mathcal{A}|}\right) H_\gamma + r_{\max}H_\gamma \leq \left(2r_{\max} + 3\tau\sqrt{|\mathcal{A}|}\right) H_\gamma,$$

which yields the desired lower bound. □

We also highlight the following result, which shows that if the policy π^μ is lower bounded on some set $\mathcal{K} \subset \mathcal{M}$, then it is also lower bounded on its convex hull $\operatorname{conv}\mathcal{K}$.

Proposition D.4. *Suppose $\mu = \alpha\mu_1 + (1-\alpha)\mu_2$ with $\alpha \in (0,1)$ and $\mu_1, \mu_2 \in \mathcal{M}$. Then,*

$$\pi_s^\mu = \underbrace{\frac{\alpha\nu_1(s)}{\alpha\nu_1(s) + \bar{\alpha}\nu_2(s)}}_{\beta_s} \pi_s^{\mu_1} + \underbrace{\frac{\bar{\alpha}\nu_2(s)}{\alpha\nu_1(s) + \bar{\alpha}\nu_2(s)}}_{1-\beta_s} \pi_s^{\mu_2},$$

where $\beta_s \in (0,1)$.

Proof. The proof follows immediately from the proof of Proposition D.2. □

Hessian upper bounds In the following, we establish upper bounds for the Hessians of the occupancy measure regularizations, $\bar{h}$, resulting from Shannon and Tsallis entropy regularization of the policy. In particular, we aim to upper bound the maximum norm of the Hessian in terms of the smallest entry of the policy.

Proposition D.5. *Consider $\mu \in \mathcal{M}$ and let $\pi_{min} = \min_{s,a} \pi^\mu(a|s)$. Then, the Hessian of $\bar{h}$ is upper bounded as follows:*

a) If $h = -\tau\mathcal{H}$, then $\left\|\nabla^2\bar{h}(\mu)\right\|_\infty \leq \frac{\tau}{\nu_{min}\pi_{min}}$.

b) If $h = -\tau\mathcal{H}_{1/2}$, then $\left\|\nabla^2\bar{h}(\mu)\right\|_\infty \leq \frac{\tau}{\nu_{min}\pi_{min}^{3/2}}$.

Here, $\|\cdot\|_\infty$ denotes the maximum norm $\|A\|_\infty = \max_{ij}|A_{ij}|$.

Proof. As shown by Schlaginhaufen and Kamgarpour [2023, Proposition B.2], the gradient of $\bar{h}$ satisfies

$$\nabla\bar{h}(\mu)(s,a) = h(\pi_s^\mu) + \nabla h(\pi_s^\mu)(a) - \langle\nabla h(\pi_s^\mu), \pi_s^\mu\rangle. \tag{9}$$

Moreover, we have

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

Using the above two formulas, we can calculate the Hessians explicitly.

Part a): For the Shannon entropy it holds by (9) that $\nabla\bar{h}(\mu)(s,a) = \tau\log\pi^\mu(a|s)$. Hence,

$$\frac{\partial^2\bar{h}(\mu)}{\partial\mu(s',a')\partial\mu(s,a)} = \tau \cdot \frac{1}{\pi^\mu(a|s)} \cdot \delta_{s,s'} \cdot \frac{\delta_{a,a'} - \pi^\mu(a|s)}{\nu(s)},$$

and

$$\left|\nabla^2\bar{h}(\mu)_{(s',a'),(s,a)}\right| = \left|\frac{\partial^2\bar{h}(\mu)}{\partial\mu(s',a')\partial\mu(s,a)}\right| \leq \frac{\tau}{\nu_{\min}\pi_{\min}}.$$

Part b): For the Tsallis entropy it holds by (9) that

$$\nabla\bar{h}(\mu)(s,a) = -\tau\left(\sum_{a''}\sqrt{\pi^\mu(a''|s)} + \frac{1}{\sqrt{\pi^\mu(a|s)}} - 2\right).$$

Therefore, the second derivative is bounded as follows

$$\begin{aligned}\left|\frac{\partial^2 \bar{h}(\mu)}{\partial\mu(s',a')\partial\mu(s,a)}\right| &= \Bigg| -\tau\cdot\delta_{s,s'}\cdot\Bigg(\sum_{a''}\frac{1}{2\sqrt{\pi^\mu(a''|s)}}\frac{\delta_{a',a''}-\pi^\mu(a''|s)}{\nu(s)} \\ &\qquad\qquad -\frac{1}{2\pi^\mu(a|s)^{3/2}}\frac{\delta_{a,a'}-\pi^\mu(a|s)}{\nu(s)}\Bigg)\Bigg| \\ &= \frac{\tau\cdot\delta_{s,s'}}{2\nu(s)}\left|\frac{1}{\sqrt{\pi^\mu(a'|s)}}-\sum_{a''}\sqrt{\pi^\mu(a''|s)}-\frac{\delta_{a,a'}}{\pi^\mu(a|s)^{3/2}}+\frac{1}{\sqrt{\pi^\mu(a|s)}}\right| \\ &\leq \frac{\tau}{2\nu_{\min}}\left(\underbrace{\left|\frac{1}{\sqrt{\pi^\mu(a'|s)}}-\sum_{a''}\sqrt{\pi^\mu(a''|s)}\right|}_{(A)}+\underbrace{\left|\frac{\delta_{a,a'}}{\pi^\mu(a|s)^{3/2}}-\frac{1}{\sqrt{\pi^\mu(a|s)}}\right|}_{(B)}\right).\end{aligned}$$

Now, since $\sum_a \sqrt{\pi^\mu(a|s)} \leq \sqrt{A} \leq 1/\sqrt{\pi_{\min}}$, we have $(A)+(B) \leq 1/\sqrt{\pi_{\min}} + 1/\pi_{\min}^{3/2} \leq 2/\pi_{\min}^{3/2}$, which yields the desired result

$$\left|\nabla^2 \bar{h}(\mu)_{(s',a'),(s,a)}\right| \leq \tau/\left(\nu_{\min}\pi_{\min}^{3/2}\right).$$

□

Lipschitz gradients Next, we provide explicit Lipschitz constants $L_\mathcal{K}$ for $\nabla\bar{h}$ corresponding to Shannon and Tsallis entropy regularization.

Proposition D.6 (Lipschitz gradients)**.** *Consider some closed convex set $\mathcal{K} \subset \mathcal{M}$ and suppose $\pi_{min} = \min_{\mu\in\mathcal{K}}\min_{s,a}\pi^\mu(a|s) > 0$. Then, the gradient of $\bar{h}$ is Lipschitz continuous over $\mathcal{K}$ i.e.*

$$\left\|\nabla\bar{h}(\mu_1)-\nabla\bar{h}(\mu_2)\right\|_2 \leq L_\mathcal{K}\left\|\mu_1-\mu_2\right\|_2, \quad \forall\mu_1,\mu_2\in\mathcal{K},$$

where the respective Lipschitz constants are as follows:

a) If $h = -\tau\mathcal{H}$, then $L_\mathcal{K} = \tau|\mathcal{S}||\mathcal{A}|/(\nu_{min}\pi_{min})$.

b) If $h = -\tau\mathcal{H}_{1/2}$, then $L_\mathcal{K} = \tau|\mathcal{S}||\mathcal{A}|/(\nu_{min}\pi_{min}^{3/2})$.

Proof. Defining $h = \mu_2 - \mu_1$, Lipschitz continuity follows from

$$\begin{aligned}\left\|\nabla\bar{h}(\mu_1)-\nabla\bar{h}(\mu_2)\right\|_2 &\overset{(i)}{\leq} \sqrt{|\mathcal{S}||\mathcal{A}|}\left\|\nabla\bar{h}(\mu_1)-\nabla\bar{h}(\mu_2)\right\|_\infty \\ &\overset{(ii)}{=} \sqrt{|\mathcal{S}||\mathcal{A}|}\left\|\int_0^1 \nabla^2\bar{h}(\mu_1+th)hdt\right\|_\infty \\ &\overset{(iii)}{\leq} \sqrt{|\mathcal{S}||\mathcal{A}|}\int_0^1 \left\|\nabla^2\bar{h}(\mu_1+th)h\right\|_\infty dt \\ &\overset{(iv)}{\leq} \sqrt{|\mathcal{S}||\mathcal{A}|}\int_0^1 \left\|\nabla^2\bar{h}(\mu_1+th)\right\|_\infty \left\|h\right\|_1 dt \\ &\overset{(v)}{\leq} |\mathcal{S}||\mathcal{A}|\max_{0\leq t\leq 1}\left\|\nabla^2\bar{h}(\mu_1+th)\right\|_\infty \left\|\mu_1-\mu_2\right\|_2.\end{aligned}$$

Here, we used in (i) and (v) that $\|x\|_1 \leq \sqrt{n}\,\|x\|_2 \leq n\,\|x\|_\infty$ for $x\in\mathbb{R}^n$, and in (ii) we applied the fundamental theorem of calculus. Moreover, (iii) follows from $\left|\int f\right| \leq \int |f|$, and (iv) from Hölder's inequality. Now, by convexity $\mu_1,\mu_2\in\mathcal{K}$ implies that $\mu_1+th\in\mathcal{K}$ for $t\in[0,1]$. Hence, plugging in the upper bounds from Proposition D.5 concludes the proof. □

D.4 Dual smoothness and strong convexity

Next, we show that the convex conjugate, f^*, of the extended real value function $f := \bar{h} + \delta_{\mathcal{M}}$ (see Appendix B) is – if understood as a mapping from $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{U}$ to $\mathbb{R}$ – both smooth and strongly convex on $\mathcal{R}$ with respect to the quotient norm. While it is well-known that global smoothness and strong convexity are dual properties [Rockafellar and Wets, 2009, Proposition 12.60], the key challenge for proving dual strong convexity is that $\bar{h}$ has only locally Lipschitz gradients (see Proposition D.6). Proposition D.7 below shows that η-strong convexity of $\bar{h}$ implies dual $1/\eta$-smoothness and locally Lipschitz gradients imply dual $\sigma_{\mathcal{R}}$-strong convexity on $\mathcal{R}$ for some $\sigma_{\mathcal{R}} > 0$. Moreover, we provide explicit lower bounds on $\sigma_{\mathcal{R}}$ for Shannon and Tsallis entropy entropy regularization.

Proposition D.7. *Let f^* be the convex conjugate of $f := \bar{h} + \delta_{\mathcal{M}}$. Then, the following holds:*

a) Suppose that $\bar{h}$ is η-strongly convex over $\mathcal{M}$, that is for all $\mu, \mu' \in \mathcal{M}$ it holds that

$$\bar{h}(\mu') \geq \bar{h}(\mu) + \langle \nabla \bar{h}(\mu), \mu' - \mu \rangle + \frac{\eta}{2} \left\| \mu - \mu' \right\|_2^2,$$

then we have for all $r, r' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ that

$$f^*(r') \leq f(r) + \langle \nabla f^*(r), r' - r \rangle + \frac{1}{2\eta} \left\| [r']_{\mathcal{U}} - [r]_{\mathcal{U}} \right\|_2^2. \tag{10}$$

b) Suppose that for any closed convex subset $\mathcal{K} \subset \operatorname{relint} \mathcal{M}$, there is some $L_{\mathcal{K}} > 0$ such that for all $\mu, \mu' \in \mathcal{K}$ it holds that

$$\left\| \nabla \bar{h}(\mu) - \nabla \bar{h}(\mu') \right\|_2 \leq L_{\mathcal{K}} \left\| \mu - \mu' \right\|_2,$$

then we have for all $r, r' \in \mathcal{R}$ that

$$f^*(r') \geq f^*(r) + \langle \nabla f^*(r), r' - r \rangle + \frac{\sigma_{\mathcal{R}}}{2} \left\| [r']_{\mathcal{U}} - [r]_{\mathcal{U}} \right\|_2^2, \tag{11}$$

for some $\sigma_{\mathcal{R}} > 0$.

c) Let $H_\gamma := 1/(1-\gamma)$, $R := \max_{r\in\mathcal{R}} \left\| r \right\|_\infty$, $D = \max_{r,r'\in\mathcal{R}} \left\| r - r' \right\|_2$, and suppose that $\tau < 2D$, then for the Shannon entropy the inequality (11) holds with

$$\sigma_{\mathcal{R}} = \frac{\nu_{min} \left(1 - \frac{\tau}{2D}\right) \exp\left(\frac{-2RH_\gamma}{\tau}\right)}{D|\mathcal{S}||\mathcal{A}|^{2+H_\gamma}},$$

and for the Tsallis entropy with

$$\sigma_{\mathcal{R}} = \frac{\nu_{min} \left(1 - \frac{\tau}{2D}\right)}{\sqrt{2}\left(\left(2R/\tau + 3\sqrt{|\mathcal{A}|}\right) H_\gamma\right)^3 D|\mathcal{S}||\mathcal{A}|}.$$

Proof. Part a): Since f is η-strongly convex with respect to $\left\|\cdot\right\|_2$, the convex conjugate f^* is $1/\eta$-smooth with respect to the dual norm in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{U}$ [Rockafellar and Wets, 2009, Proposition 12.60], which is equivalent to (10).

Part b): The show b), we closely follow [Goebel and Rockafellar, 2008, Theorem 4.1], but we need to account for the quotient spaces. We define the sets $\mathcal{K} = \nabla f^*(\mathcal{R}) = \mathsf{RL}(\mathcal{R})$ and $\mathcal{K}_\epsilon = \operatorname{conv}(\mathcal{K}) + \epsilon(\mathcal{B} \cap \operatorname{aff}(\mathcal{M}))$, where $\mathcal{B} \subset \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ denotes the closed unit ball with respect to $\left\|\cdot\right\|_2$ and $\epsilon > 0$ is chosen such that $\mathcal{K}_\epsilon \subset \operatorname{relint} \mathcal{M}$. Moreover, we let L be the Lipschitz constant of $\nabla \bar{h}$ over

$\mathcal{K}_\epsilon$. Now, consider $r \in \mathcal{R}$ and $\mu = \nabla f^*(r)$. Then, for any $r' \in \mathcal{R}$, we have

$$\begin{aligned}
f^*(r') &= \sup_{\bar{\mu}} \left[\langle r', \bar{\mu}\rangle - f(\bar{\mu})\right] \\
&\overset{(i)}{\geq} \sup_{\bar{\mu}\in\mathcal{K}_\epsilon} \left[\langle r', \bar{\mu}\rangle - f(\bar{\mu})\right] \\
&\overset{(ii)}{=} \sup_{\bar{\mu}\in\mathcal{K}_\epsilon} \left[\langle r', \bar{\mu}\rangle - f(\mu) - \langle r, \bar{\mu} - \mu\rangle - \frac{L}{2}\left\|\bar{\mu} - \mu\right\|_2^2\right] \\
&\overset{(iii)}{=} \langle r, \mu\rangle - f(\mu) + \sup_{\bar{\mu}\in\mathcal{K}_\epsilon} \left[\langle r' - r, \bar{\mu}\rangle - \frac{L}{2}\left\|\bar{\mu} - \mu\right\|_2^2\right] \\
&\overset{(iv)}{=} f^*(r) + \sup_{\bar{\mu}\in\mathcal{K}_\epsilon} \left[\langle r' - r, \bar{\mu}\rangle - \frac{L}{2}\left\|\bar{\mu} - \mu\right\|_2^2\right] \\
&\overset{(v)}{=} f^*(r) + \langle r' - r, \mu\rangle + \sup_{\bar{\mu}\in\mathcal{K}_\epsilon} \left[\langle r' - r, \bar{\mu} - \mu\rangle - \frac{L}{2}\left\|\bar{\mu} - \mu\right\|_2^2\right].
\end{aligned}$$

Here, (i) follows from $\mathcal{K}_\epsilon \subset \mathcal{M}$, (ii) from the fact that Lipschitz gradients imply that

$$f(\bar{\mu}) \geq f(\mu) + \langle g, \bar{\mu} - \mu\rangle + \frac{L}{2}\left\|\bar{\mu} - \mu\right\|_2^2,$$

for any $g \in \partial f(\mu)$ [Beck, 2017, Lemma 5.7] and $r \in \partial f(\mu)$ (see Theorem B.4). Moreover, (iii) and (v) follow from rearranging terms, and (iv) from $f^*(r) = \langle r, \mu\rangle - f(\mu)$. Now, consider any $\alpha > 0$ such that $\sigma_{\mathcal{R}} = 2(\alpha - L\alpha^2/2) > 0$ and $\alpha D \leq \epsilon$. Setting $\bar{\mu} \in \mathcal{K}_\epsilon$ in the above supremum to $(\bar{\mu} - \mu) = \alpha \Pi_{\mathcal{U}^\perp}(r' - r)$ yields the desired result

$$\begin{aligned}
f^*(r') &\geq f^*(r) + \langle r' - r, \nabla f^*(r)\rangle + \alpha\langle r' - r, \Pi_{\mathcal{U}^\perp}(r' - r)\rangle - \frac{L\alpha^2}{2}\left\|\Pi_{\mathcal{U}^\perp}(r' - r)\right\|_2^2 \\
&= f^*(r) + \langle r' - r, \nabla f^*(r)\rangle + \frac{\sigma_{\mathcal{R}}}{2}\left\|[r']_{\mathcal{U}} - [r]_{\mathcal{U}}\right\|_2^2.
\end{aligned}$$

Note that we indeed have $\bar{\mu} \in \mathcal{K}_\epsilon$ as $\mu \in \mathcal{K}$ and $\left\|\mu - \bar{\mu}\right\|_2 \leq \alpha\left\|r - r'\right\|_2 \leq \alpha D \leq \epsilon$.

Part c): To get an explicit constant for $\sigma_{\mathcal{R}}$, we need to appropriately choose ϵ and calculate the corresponding Lipschitz constant. To this end, we first recall that according to Proposition D.3 and D.4 policies corresponding to occupancy measures in $\operatorname{conv}\mathcal{K}$ are, for Shannon and Tsallis entropy, lower bounded by

$$\pi_{\text{min, Sh}} = \frac{\exp\left(-2RH_\gamma/\tau\right)}{|\mathcal{A}|^{1+H_\gamma}} \quad \text{and} \quad \pi_{\text{min, Ts}} = \left(\left(2R/\tau + 3\sqrt{|\mathcal{A}|}\right)H_\gamma\right)^{-2},$$

respectively. Furthermore, for any $\mu \in \mathcal{K}_\epsilon$, we have by Proposition D.1 the lower bound

$$\pi^\mu(a|s) \geq \pi_{\text{min}} - \frac{2\sqrt{|\mathcal{S}||\mathcal{A}|}}{\nu_{\text{min}}}\epsilon = \pi'_{\text{min}}.$$

Hence, by setting $\epsilon = \nu_{\text{min}}\pi_{\text{min}}/(4\sqrt{|\mathcal{S}||\mathcal{A}|})$, we have $\pi^\mu(a|s) \geq \pi'_{\text{min}} = \pi_{\text{min}}/2$ for any $\mu \in \mathcal{K}_\epsilon$. As for the Lipschitz constant over $\mathcal{K}_\epsilon$, we have by Proposition D.6

$$L_{\text{sh}} = \frac{\tau|\mathcal{S}||\mathcal{A}|}{\nu_{\text{min}}\pi'_{\text{min, Sh}}} \quad \text{and} \quad L_{\text{ts}} = \frac{\tau|\mathcal{S}||\mathcal{A}|}{\nu_{\text{min}}{\pi'_{\text{min, Ts}}}^{3/2}}, \tag{13}$$

for the Shannon and Tsallis entropy, respectively. Now, we need to ensure that $\alpha > 0$ such that $\sigma_{\mathcal{R}} = 2(\alpha - L\alpha^2/2) > 0$ and $\alpha \leq \epsilon/D$. To that end, we set for both regularizations $\alpha = \tau/(LD)$, which in light of (13) ensures that

$$\alpha = \frac{\tau}{LD} \leq \frac{\nu_{\text{min}}\pi'_{\text{min}}}{D|\mathcal{S}||\mathcal{A}|} \leq \frac{\nu_{\text{min}}\pi'_{\text{min}}}{2D\sqrt{|\mathcal{S}||\mathcal{A}|}} = \frac{\epsilon}{D},$$

for $|\mathcal{S}|, |\mathcal{A}| \geq 2$. Moreover, we get the dual strong convexity constant

$$\sigma_{\mathcal{R}} = 2\left(\frac{\tau}{LD} - \frac{\tau^2}{2LD^2}\right) = \frac{2\tau}{LD}\left(1 - \frac{\tau}{2D}\right),$$

which is larger than zero as long as $\tau < 2D$. Plugging in the Lipschitz constants for the two regularizations yields

$$\sigma_{\mathcal{R}} = \frac{2\pi'_{\text{min, Sh}}\nu_{\min}}{D|\mathcal{S}||\mathcal{A}|}\left(1 - \frac{\tau}{2D}\right) = \frac{\pi_{\text{min, Sh}}\nu_{\min}}{D|\mathcal{S}||\mathcal{A}|}\left(1 - \frac{\tau}{2D}\right) = \frac{\exp\left(-2RH_\gamma/\tau\right)\nu_{\min}}{D|\mathcal{S}||\mathcal{A}|^{2+H_\gamma}}\left(1 - \frac{\tau}{2D}\right),$$

for the Shannon entropy, and

$$\begin{aligned}\sigma_{\mathcal{R}} &= \frac{2{\pi'_{\text{min, Ts}}}^{3/2}\nu_{\min}}{D|\mathcal{S}||\mathcal{A}|}\left(1 - \frac{\tau}{2D}\right) = \frac{\pi_{\text{min, Ts}}^{3/2}\nu_{\min}}{\sqrt{2}D|\mathcal{S}||\mathcal{A}|}\left(1 - \frac{\tau}{2D}\right) \\ &= \frac{\nu_{\min}}{\sqrt{2}\left(\left(2R/\tau + 3\sqrt{|\mathcal{A}|}\right)H_\gamma\right)^3 D|\mathcal{S}||\mathcal{A}|}\left(1 - \frac{\tau}{2D}\right),\end{aligned}$$

for the Tsallis entropy. □

D.5 Regularity constants

In the following Proposition, we summarize the regularity constants for Shannon and Tsallis entropy regularization. We highlight that these constants are lower bounds for η and $\sigma_{\mathcal{R}}$.

Proposition D.8. *Let $H_\gamma := 1/(1-\gamma)$, $R := \max_{r\in\mathcal{R}} \|r\|_\infty$, and $D = \max_{r,r'\in\mathcal{R}} \|r - r'\|_2$. Suppose that $\tau < 2D$, then for the Shannon entropy, Assumption 3.4 holds with*

$$\eta = \tau\nu_{\mathit{min}}/H_\gamma^2 \quad \textit{and} \quad \sigma_{\mathcal{R}} = \frac{\nu_{\mathit{min}}\left(1 - \frac{\tau}{2D}\right)\exp\left(\frac{-2RH_\gamma}{\tau}\right)}{D|\mathcal{S}||\mathcal{A}|^{2+H_\gamma}},$$

and for the Tsallis entropy with

$$\eta = \tau\nu_{\mathit{min}}/(2H_\gamma^2|\mathcal{A}|) \quad \textit{and} \quad \sigma_{\mathcal{R}} = \frac{\nu_{\mathit{min}}\left(1 - \frac{\tau}{2D}\right)}{\sqrt{2}\left(\left(2R/\tau + 3\sqrt{|\mathcal{A}|}\right)H_\gamma\right)^3 D|\mathcal{S}||\mathcal{A}|}.$$

Proof. The derivation for η is given in Proposition D.2 and for $\sigma_{\mathcal{R}}$ in Proposition D.7 above. □

D.6 Suboptimality bounds

Proposition 3.3. *Under Assumptions 2.1 and 2.2, we have $\ell(r',\mu) = D_{\bar{h}}(\mu, \mathsf{RL}(r'))$ for any $\mu \in \mathcal{M}$.*

Proof. Let $\mu = \mathsf{RL}(r)$. We have

$$\begin{aligned}\ell(r,\mu') &= \langle r, \mu - \mu'\rangle - \bar{h}(\mu) + \bar{h}(\mu') \\ &= \langle \nabla\bar{h}(\mu), \mu - \mu'\rangle - \bar{h}(\mu) + \bar{h}(\mu') \\ &= D_{\bar{h}}(\mu', \mu),\end{aligned}$$

where the second equality holds, as by (4) we have $r - \nabla\bar{h}(\mu) \in \mathcal{U}$, and $\mu - \mu' \in \mathcal{U}^\perp$. □

Lemma 3.5. *Suppose Assumptions 2.1,2.2, and 3.4 hold, and let $r, r' \in \mathcal{R}$. Then, we have*

$$\frac{\sigma_{\mathcal{R}}}{2}\left\|[r]_{\mathcal{U}} - [r']_{\mathcal{U}}\right\|_2^2 \le \ell(r', \mathsf{RL}(r)) = D_{\bar{h}}\left(\mathsf{RL}(r), \mathsf{RL}(r')\right) \le \frac{1}{2\eta}\left\|[r]_{\mathcal{U}} - [r']_{\mathcal{U}}\right\|_2^2, \tag{6}$$

for some problem-dependent constant $\sigma_{\mathcal{R}} > 0$.

Proof. Let $f := \bar{h} + \delta_{\mathcal{M}}$ and $\mu = \mathsf{RL}(r), \mu' = \mathsf{RL}(r')$. We then have

$$\begin{aligned}D_{\bar{h}}(\mu,\mu') &\overset{(i)}{=} f(\mu) - f(\mu') - \langle r', \mu - \mu'\rangle \\ &\overset{(ii)}{=} f^*(r') - \langle r', \mu'\rangle - f^*(r) + \langle r, \mu\rangle - \langle r', \mu - \mu'\rangle \\ &\overset{(iii)}{=} f^*(r') - f^*(r) - \langle r' - r, \nabla f^*(r)\rangle = D_{f^*}(r', r).\end{aligned}$$

Here, (i) follows from the definition of f and $r' \in [\nabla\bar{h}(\mu')]_{\mathcal{U}}$, in (ii) we use that $f(\mu) = \langle r, \mu\rangle - f^*(r)$ and $f(\mu') = \langle r', \mu'\rangle - f^*(r')$, and (iii) follows from rearranging terms and $\mu = \nabla f^*(r)$. The result then follows from dual strong convexity and smoothness as established in Proposition D.7. □

Note that without steep regularization it is impossible to lower bound the suboptimality in terms of reward distances in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathcal{U}$ (Proposition 3.3 doesn't hold). However, we still have the following upper bound.

Proposition D.9. *Consider an arbitrary regularization and let $\mu \in \mathsf{RL}(r), \mu' \in \mathsf{RL}(r')$. Then,*

$$\ell(r,\mu') \le 2\left\|[r]_{\mathcal{U}} - [r']_{\mathcal{U}}\right\|_\infty \le 2\left\|[r]_{\mathcal{U}} - [r']_{\mathcal{U}}\right\|_2 .$$

Proof. Let $r'' := \operatorname{argmin}_{\tilde{r}\in[r']_{\mathcal{U}}} \|\tilde{r} - r\|_\infty$, then the following holds

$$\begin{aligned}
\ell(r,\mu') &= \max_{\mu\in\mathcal{M}} J(r,\mu) - J(r,\mu') \\
&\overset{(i)}{\le} \left|\max_{\mu\in\mathcal{M}} J(r,\mu) - \max_{\mu\in\mathcal{M}} J(r'',\mu)\right| + |J(r'',\mu') - J(r,\mu')| \\
&\overset{(ii)}{\le} \max_{\mu\in\mathcal{M}} |\langle r - r'', \mu\rangle| + |\langle r - r'', \mu'\rangle| \\
&\overset{(iii)}{\le} 2\|r - r''\|_\infty \overset{(iv)}{=} 2\left\|[r]_{\mathcal{U}} - [r']_{\mathcal{U}}\right\|_\infty \le 2\left\|[r]_{\mathcal{U}} - [r']_{\mathcal{U}}\right\|_2 .
\end{aligned}$$

Here, (i) follows from the triangle inequality and optimality of μ', (ii) from $|\max f - \max g| \le \max|f-g|$ and simplifying, (iii) from Hölder's inequality, and (iv) from the definition of r'' and the quotient norm. □

D.7 Perturbation bounds

Next, we provide a bound quantifying the change in the quotient norm when changing the generating subspace.

Proposition D.10. *Consider $x, y \in \mathbb{R}^n$ and two subspaces $\mathcal{V}, \mathcal{W} \subset \mathbb{R}^n$ of dimension $m < n$. Then,*

$$\left\|[x]_{\mathcal{W}} - [y]_{\mathcal{W}}\right\|_2 \le \|\Pi_{\mathcal{W}} - \Pi_{\mathcal{V}}\| \cdot \|x - y\|_2 + \left\|[x]_{\mathcal{V}} - [y]_{\mathcal{V}}\right\|_2 ,$$

where $\|\Pi_{\mathcal{W}} - \Pi_{\mathcal{V}}\| = \sin(\theta_{\max}(\mathcal{V},\mathcal{W}))$.

Proof. The result follows from the triangle inequality and the definition of the spectral norm:

$$\begin{aligned}
\left\|[x]_{\mathcal{W}} - [y]_{\mathcal{W}}\right\|_2 &= \|\Pi_{\mathcal{W}^\perp}(x-y)\|_2 \\
&= \|(\Pi_{\mathcal{W}^\perp} - \Pi_{\mathcal{V}^\perp})(x-y) + \Pi_{\mathcal{V}^\perp}(x-y)\|_2 \\
&\le \|(\Pi_{\mathcal{W}^\perp} - \Pi_{\mathcal{V}^\perp})(x-y)\|_2 + \|\Pi_{\mathcal{V}^\perp}(x-y)\|_2 \\
&= \|(\Pi_{\mathcal{W}} - \Pi_{\mathcal{V}})(x-y)\|_2 + \left\|[x]_{\mathcal{V}} - [y]_{\mathcal{V}}\right\|_2 \\
&\le \|\Pi_{\mathcal{W}} - \Pi_{\mathcal{V}}\| \cdot \|x-y\|_2 + \left\|[x]_{\mathcal{V}} - [y]_{\mathcal{V}}\right\|_2 .
\end{aligned}$$

Furthermore, for a proof of $\|\Pi_{\mathcal{W}} - \Pi_{\mathcal{V}}\| = \sin(\theta_{\max}(\mathcal{V},\mathcal{W}))$ we refer to [Drmac, 2000]. □

The following proposition shows that the maximal principal angle between two transition laws can be upper bounded by the spectral norm difference of the transition laws.

Proposition 3.8. *Let $P, P' \in \Delta_{\mathcal{S}\times\mathcal{A}}^{\mathcal{S}}$ and $H_\gamma = 1/(1-\gamma)$. Then, we have $\theta_1(P,P') = 0$ and $\sin(\theta_{\max}(P,P')) \le \gamma H_\gamma \sqrt{|\mathcal{S}|/|\mathcal{A}|}\,\|P - P'\|$, where $\|\cdot\|$ denotes the spectral norm.*

Before proceeding with the proof of Proposition 3.8, we need the following technical result.

Proposition D.11. *For any $P \in \Delta_{\mathcal{S}\times\mathcal{A}}^{\mathcal{S}}$ the smallest singular value of $E - \gamma P$ satisfies*

$$\sigma_{\min}(E - \gamma P) \ge \sqrt{|\mathcal{A}|/|\mathcal{S}|}(1-\gamma).$$

Proof. The main idea of the proof is to use that $\sigma_{\min}(A) = \min_{x\neq 0} \|Ax\|_2 / \|x\|_2$ for any matrix $A \in \mathbb{R}^{n\times m}$. We first lower bound $\sigma_{\min}(I - \gamma P_a) = \sigma_{\min}\left(I - \gamma (P_a)^\top\right)$, where P_a denotes the state transition matrix under action a. Let $x \in \mathbb{R}^S$, then we have

$$\begin{aligned}\left\|\left(I-\gamma(P_a)^\top\right)x\right\|_2 &\geq 1/\sqrt{|\mathcal{S}|}\left\|\left(I-\gamma(P_a)^\top\right)x\right\|_1\\ &\geq 1/\sqrt{|\mathcal{S}|}\left(\|x\|_1-\gamma\left\|(P_a)^\top x\right\|_1\right)\\ &\geq (1-\gamma)/\sqrt{|\mathcal{S}|}\,\|x\|_1 \geq (1-\gamma)/\sqrt{|\mathcal{S}|}\,\|x\|_2\,,\end{aligned}$$

where the third inequality follows from

$$\left\|(P_a)^\top x\right\|_1 = \sum_s \left|\sum_{s'} P_a(s|s')x(s')\right| \leq \sum_s\sum_{s'} P_a(s|s')\,|x(s')| = \|x\|_1\,.$$

Hence, $\sigma_{\min}(I-\gamma P_a) \geq (1-\gamma)/\sqrt{|\mathcal{S}|}$. Therefore, we have for any $x \in \mathbb{R}^{\mathcal{S}}$ that

$$\|(E-\gamma P)x\|_2^2 = \sum_a \|(I-\gamma P_a)x\|_2^2 \geq |\mathcal{A}|/|\mathcal{S}|(1-\gamma)^2\,\|x\|_2^2\,,$$

which yields the desired result. □

We are now ready to prove Proposition 3.8.

Proof of Proposition 3.8. The first principal angle $\theta_1(P,P')=0$ is zero as we always have $\mathbf{1}\subseteq\mathcal{U}$. The bound on the maximal angle follows from a well-known perturbation result for orthogonal projections. Namely, if $A,B\in\mathbb{R}^{n\times m}$ are matrices of the same rank and Π_A,Π_B denote the orthogonal projections onto their column span, then we have [Ji-Guang, 1987]

$$\|\Pi_A-\Pi_B\| \leq \min\left\{\left\|A^\dagger\right\|,\left\|B^\dagger\right\|\right\}\|A-B\|\,,$$

where $A^\dagger$ denotes the Moore-Penrose inverse [Penrose, 1955]. Recall that $\mathcal{U}_P=\operatorname{im}(E-\gamma P)$, $\sin\left(\theta_{\max}(P,P')\right)=\left\|\Pi_{\mathcal{U}_P}-\Pi_{\mathcal{U}_{P'}}\right\|$, and by Proposition D.11

$$\left\|(E-\gamma P)^\dagger\right\| = \left(\sigma_{\min}\left(E-\gamma P\right)\right)^{-1} \leq \sqrt{|\mathcal{S}|/|\mathcal{A}|}H_\gamma.$$

Therefore, we get

$$\sin\left(\theta_{\max}(P,P')\right) \leq \sqrt{|\mathcal{S}|/|\mathcal{A}|}\cdot\gamma\cdot H_\gamma\cdot\|P-P'\|\,.$$

□

E Proof of claim in Example 3.2

We recall Example 3.2 from the main paper.

Example E.1. We consider a two-state, two-action MDP with $\mathcal{S}=\mathcal{A}=\{0,1\}$, uniform initial state distribution, discount rate $\gamma=0.9$, and Shannon entropy regularization $h=-\mathcal{H}$ (see Appendix C). Suppose the expert reward is $r^{\mathsf{E}}(s,a)=\mathbb{1}\{s=1\}$ and consider the transition laws, P^0 and P^1, defined by $P^0(0|s,a)=1$ and $P^1(0|s,a)=\beta\cdot\mathbb{1}\{s=0,a=0\}$ for some $\beta\in(0,1)$. Also, consider the two experts $\mu^{\mathsf{E}}_{P^0}=\mathsf{RL}_{P^0}(r^{\mathsf{E}})$ and $\mu^{\mathsf{E}}_{P^1}=\mathsf{RL}_{P^1}(r^{\mathsf{E}})$, and suppose we recovered the reward $\hat{r}(s,a)=-r^{\mathsf{E}}$. Then, the following holds: 1) We have $\ell_{P^0}(\hat{r},\mu^{\mathsf{E}}_{P^0})=0$ and $\ell_{P^1}(\hat{r},\mu^{\mathsf{E}}_{P^1})=\mathcal{O}(\beta)$. That is, for small β, the reward $\hat{r}$ is a good solution to the IRL problem, as both experts are approximately optimal under $\hat{r}$. 2) The rank condition (5) between P^0 and P^1 is satisfied for any $\beta\in(0,1)$. 3) For a new transition law P defined by $P(0|s,a)=\mathbb{1}\{s=1,a=0\}$, we have $\ell_P(r^{\mathsf{E}},\mathsf{RL}_P(\hat{r}))\approx 4.81$, i.e. $\mathsf{RL}_P(\hat{r})$ performs poorly under the experts' reward.

In the following we prove the claims 1. and 2., while 3. is computed via regularized dynamic programming [Geist et al., 2019].[2]

[2] The code is provided as a .zip file in the supplementary material.

1. Consider the transition law P' defined by $P'(0|s,a) = 0$. Observe that while $\left\|P^0 - P'\right\|$ is large, the potential shaping spaces $\mathcal{U}_{P^0}$ and $\mathcal{U}_{P'}$ coincide. Moreover, we have

$$\left\|P^1 - P'\right\| \leq \sqrt{\sum_{s,s',a} (P^1(s'|s,a) - P'(s'|s,a))^2} = \sqrt{2}\beta.$$

   In light of Propositions D.9, D.10, and 3.8, we have

$$\begin{aligned} \ell_{P^1}(\hat{r}, \mathsf{RL}_{P^1}(r^{\mathsf{E}})) &\leq 2\left\|[\hat{r}]_{\mathcal{U}_{P^1}} - [r^{\mathsf{E}}]_{\mathcal{U}_{P^1}}\right\|_2 \\ &\leq 2\left\|\Pi_{\mathcal{U}_{P^1}} - \Pi_{\mathcal{U}_{P'}}\right\| \left\|\hat{r} - r^{\mathsf{E}}\right\|_2 \\ &\leq 2\sqrt{2}\gamma \cdot H_\gamma \cdot \|P - P'\| \leq 4\gamma \cdot H_\gamma \cdot \beta. \end{aligned}$$

2. We need to show that P^0 and P^1 are satisfying the rank condition

$$\operatorname{rank}\left(\begin{bmatrix} E - \gamma P^0, & E - \gamma P^1 \end{bmatrix}\right) = 2|\mathcal{S}| - 1.$$

   To this end, we choose the matrix representation

$$E = \begin{bmatrix} I \\ I \end{bmatrix} \quad \text{and} \quad P = \begin{bmatrix} P^0 \\ P^1 \end{bmatrix},$$

   where $I \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ is the identity matrix and $P^0, P^1 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ are the state transition matrices corresponding to the actions $0, 1$, respectively. Let $C = \begin{bmatrix} E - \gamma P^0, & E - \gamma P^1 \end{bmatrix}$. We have

$$P^0 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad P^1 = \begin{bmatrix} \beta & 1-\beta \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix},$$

   and

$$C = \begin{bmatrix} 1-\gamma & 0 & 1-\beta\gamma & -\gamma+\beta\gamma \\ -\gamma & 1 & 0 & 1-\gamma \\ 1-\gamma & 0 & 1 & -\gamma \\ -\gamma & 1 & 0 & 1-\gamma \end{bmatrix}.$$

   It's straightforward to see that the vector $\begin{bmatrix} 1 & 1 & -1 & -1 \end{bmatrix}^\top$ lies in the kernel of C, but there is a 3×3 submatrix with non-zero determinant:

$$\det\left(\begin{bmatrix} 1-\gamma & 0 & 1-\beta\gamma \\ -\gamma & 1 & 0 \\ 1-\gamma & 0 & 1 \end{bmatrix}\right) = 1 \cdot [(1-\gamma) - (1-\gamma)(1-\beta\gamma)] = \beta\gamma(1-\gamma) > 0.$$

   In other words, we have $\operatorname{rank} C = 3$ for any $\beta > 0$.

F Proof of Theorem 3.9

Theorem 3.9. *Let $K = 2$, $\theta_2(P^0, P^1) > 0$, and suppose that Assumptions 2.1,2.2, and 3.4 hold. If $\ell_{P^k}(\hat{r}, \mu^{\mathsf{E}}_{P^k}) \leq \hat{\varepsilon}$ for $k = 0, 1$, then $\hat{r}$ is ε-transferable to $\mathcal{P} = \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$ with*

$$\varepsilon = \hat{\varepsilon} / \left[\eta\sigma_{\mathcal{R}} \sin\left(\theta_2(P^0, P^1)/2\right)^2\right].$$

The proof of Theorem 3.9 hinges on Lemma 3.5 and the following reward approximation result.

Lemma F.1. *Let $\left\|[r^E]_{\mathcal{U}_{P^k}} - [\hat{r}]_{\mathcal{U}_{P^k}}\right\|_2 \leq \bar{\varepsilon}$ for $k = 0, 1$. Then, if $\theta_2(P^0, P^1) > 0$, it holds that*

$$\left\|[r^E]_{\mathbf{1}} - [\hat{r}]_{\mathbf{1}}\right\|_2 \leq \frac{\bar{\varepsilon}}{\sin\left(\theta_2(P^0, P^1)/2\right)}.$$

Proof of Lemma F.1. Throughout this proof, we will use the short-hand notation $\mathcal{U}_k := \mathcal{U}_{P^k}$ for $k = 0, 1$. Recall that since $\mathbf{1} \subseteq \mathcal{U}_0 \cap \mathcal{U}_1$, we have $\theta_1(\mathcal{U}_0, \mathcal{U}_1) = 0$ and by assumption we also have $\theta_2(\mathcal{U}_0, \mathcal{U}_1) > 0$, which implies that $\mathcal{U}_0 \cap \mathcal{U}_1 = \mathbf{1}$. Furthermore, since for $k = 0, 1$ we can rewrite $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ as the orthogonal sum $\mathbb{R}^{\mathcal{S}\times\mathcal{A}} = \mathcal{U}_k \cap \mathbf{1}^\perp \oplus \mathcal{U}_k^\perp \oplus \mathbf{1}$, we can uniquely decompose $r^{\mathsf{E}} - \hat{r}$ into

$r^{\mathsf{E}} - \hat{r} = x_k + y_k + z$, where $x_k \in \mathcal{U}_k \cap \mathbf{1}^\perp$, $y_k \in \mathcal{U}_k^\perp$, $z \in \mathbf{1}$, for $k = 0, 1$. Then, it holds that $x_0 + y_0 = x_1 + y_1$. Since $\left\|[r^{\mathsf{E}}]_{P^k} - [\hat{r}]_{P^k}\right\|_{P^k,2} = \|y_k\|_2$, the Assumption of Lemma F.1 implies that $\|y_k\|_2 \leq \bar{\varepsilon}$. For the 2-distance between the equivalence classes $[r^{\mathsf{E}}]_{\mathbf{1}}$ and $[\hat{r}]_{\mathbf{1}}$ the Pythagorean theorem implies that

$$\left\|[r^{\mathsf{E}}]_{\mathbf{1}} - [\hat{r}]_{\mathbf{1}}\right\|_{\mathbf{1},2}^2 = \|x_0\|_2^2 + \|y_0\|^2 = \|x_1\|_2^2 + \|y_1\|_2^2 \leq \max_{\substack{u_k \in \mathcal{U}_k \cap \mathbf{1}^\perp, v_k \in \mathcal{U}_k^\perp, \\ \|u_k\|_2 = \|v_k\|_2 = 1, \\ \alpha_k \in \mathbb{R}_+, \beta_k \in [0, \bar{\varepsilon}], k=0,1, \\ \alpha_0 u_0 + \beta_0 v_0 = \alpha_1 u_1 + \beta_1 v_1}} \alpha_0^2 + \beta_0^2, \tag{26}$$

where the upper bound follows from $x_0 + y_0 = x_1 + y_1$ and $\|y_k\|_2 \leq \bar{\varepsilon}$. Next, we want to show that the maximum on the right-hand side of (26) is achieved for $\beta_0 = \beta_1 = \bar{\varepsilon}$. To see this, note that taking inner products between u_0 and u_1, respectively, and the equation $\alpha_0 u_0 + \beta_0 v_0 = \alpha_1 u_1 + \beta_1 v_1$, we arrive at

$$\alpha_0 = \alpha_1 \langle u_0, u_1 \rangle + \beta_1 \langle u_0, v_1 \rangle, \ \alpha_1 = \alpha_0 \langle u_0, u_1 \rangle + \beta_0 \langle u_1, v_0 \rangle,$$

which is for any choice of $\beta_k, u_k, v_k, k = 0, 1$ an invertible linear system of equations for α_0, α_1 with the solutions

$$\alpha_0 = \frac{\beta_0 \langle u_0, u_1 \rangle \langle u_1, v_0 \rangle + \beta_1 \langle u_0, v_1 \rangle}{1 - \langle u_0, u_1 \rangle^2}, \ \alpha_1 = \frac{\beta_1 \langle u_1, u_0 \rangle \langle u_0, v_1 \rangle + \beta_0 \langle u_1, v_0 \rangle}{1 - \langle u_1, u_0 \rangle^2}$$

where $\langle u_0, u_1 \rangle < 1$, due to $\mathcal{U}_0 \cap \mathcal{U}_1 \cap \mathbf{1}^\perp = 0$. As the sign of $\langle u_0, u_1 \rangle \langle u_1, v_0 \rangle$ and $\langle u_0, v_1 \rangle$ can be chosen arbitrarily by an appropriate choice of v_0, v_1, the objective in the right-hand-side of (26) is increasing in β_0, β_1 and hence the maximum is achieved for $\beta_0 = \beta_1 = \bar{\varepsilon}$ and $\alpha := \alpha_0 = \alpha_1 = \frac{\bar{\varepsilon}\langle u_0, v_1 \rangle}{1 - \langle u_0, u_1 \rangle}$. Therefore, it holds that

$$\begin{aligned}
\left\|[r^{\mathsf{E}}]_{\mathbf{1}} - [\hat{r}]_{\mathbf{1}}\right\|_{\mathbf{1},2}^2 &\leq \max_{\substack{u_k \in \mathcal{U}_k \cap \mathbf{1}^\perp, v_1 \in \mathcal{U}_1^\perp, \\ \|u_k\|_2 = \|v_1\|_2 = 1, k=0,1}} \bar{\varepsilon}^2 \left[1 + \left(\frac{\langle u_0, v_1 \rangle}{1 - \langle u_0, u_1 \rangle}\right)^2\right] \\
&\overset{(i)}{=} \max_{\substack{u_0 \in \mathcal{U}_0 \cap \mathbf{1}^\perp, \\ \|u_0\|_2 = 1}} \bar{\varepsilon}^2 \left[1 + \left(\frac{\max_{v_1 \in \mathcal{U}_1^\perp, \|v_1\|_2 = 1} \langle u_0, v_1 \rangle}{1 - \max_{u_1 \in \mathcal{U}_1 \cap \mathbf{1}^\perp, \|u_1\|_2 = 1} \langle u_0, u_1 \rangle}\right)^2\right] \\
&\overset{(ii)}{=} \max_{\substack{u_0 \in \mathcal{U}_0 \cap \mathbf{1}^\perp, \\ \|u_0\|_2 = 1}} \bar{\varepsilon}^2 \left[1 + \left(\frac{\left\|\Pi_{\mathcal{U}_1^\perp} u_0\right\|_2}{1 - \|\Pi_{\mathcal{U}_1 \cap \mathbf{1}^\perp} u_0\|_2}\right)^2\right] \\
&\overset{(iii)}{=} \max_{\substack{u_0 \in \mathcal{U}_0 \cap \mathbf{1}^\perp, \\ \|u_0\|_2 = 1}} \bar{\varepsilon}^2 \left[1 + \left(\frac{\sqrt{1 - \|\Pi_{\mathcal{U}_1} u_0\|_2^2}}{1 - \|\Pi_{\mathcal{U}_1 \cap \mathbf{1}^\perp} u_0\|_2}\right)^2\right] \\
&\overset{(iv)}{=} \max_{\substack{u_0 \in \mathcal{U}_0 \cap \mathbf{1}^\perp, \\ \|u_0\|_2 = 1}} \bar{\varepsilon}^2 \left[1 + \left(\frac{\sqrt{1 - \|\Pi_{\mathcal{U}_1 \cap \mathbf{1}^\perp} u_0\|_2^2}}{1 - \|\Pi_{\mathcal{U}_1 \cap \mathbf{1}^\perp} u_0\|_2}\right)^2\right] \\
&\overset{(v)}{=} \max_{\substack{u_0 \in \mathcal{U}_0 \cap \mathbf{1}^\perp, \\ \|u_0\|_2 = 1}} \bar{\varepsilon}^2 \left[1 + \frac{1 + \|\Pi_{\mathcal{U}_1 \cap \mathbf{1}^\perp} u_0\|_2}{1 - \|\Pi_{\mathcal{U}_1 \cap \mathbf{1}^\perp} u_0\|_2}\right] \\
&\overset{(vi)}{=} \bar{\varepsilon}^2 \left[1 + \frac{1 + \cos\left(\theta_2(\mathcal{U}_0, \mathcal{U}_1)\right)}{1 - \cos\left(\theta_2(\mathcal{U}_0, \mathcal{U}_1)\right)}\right] \\
&\overset{(vii)}{=} \bar{\varepsilon}^2 \frac{2}{1 - \cos\left(\theta_2(\mathcal{U}_0, \mathcal{U}_1)\right)} \\
&\overset{(viii)}{=} \frac{\bar{\varepsilon}^2}{\sin\left(\theta_2(\mathcal{U}_0, \mathcal{U}_1)/2\right)^2}.
\end{aligned}$$

Here, we took the maximum over u_1, v_1 in (i), we used that $\max_{v \in \mathcal{V}, \|v\|_2 = 1} \langle v, u \rangle = \|\Pi_{\mathcal{V}} u\|_2$ in (ii), and (iii) follows from the Pythagorean theorem. Furthermore, (iv) follows from $u_0 \in \mathbf{1}^\perp$ and

(v) from simplifying. In (vi) we then again use $\max_{v \in \mathcal{V}, \|v\|_2 = 1} \langle v, u \rangle = \|\Pi_{\mathcal{V}} u\|_2$, the definition of the second principal angle (Definition 3.7), and the fact that the first principal vectors lie in $\mathbf{1}$. Lastly, (vii) follows from simplifying and $(viii)$ from $\sin(x/2)^2 = (1 - \cos x)/2$. □

Proof of Theorem 3.9. As mentioned in the proof sketch in the main paper, it follows from the lower bound in Lemma 3.5 that $\left\|[r^{\mathsf{E}}]_{\mathcal{U}_{P^k}} - [\hat{r}]_{\mathcal{U}_{P^k}}\right\|_2 \leq \sqrt{2\hat{\varepsilon}/\sigma_{\mathcal{R}}}$. In light of Lemma F.1, this implies that for any $P \in \Delta^{\mathcal{S}}_{\mathcal{S}\times\mathcal{A}}$ we have

$$\left\|[r^{\mathsf{E}}]_{\mathcal{U}_P} - [\hat{r}]_{\mathcal{U}_P}\right\|_2 \leq \left\|[r^{\mathsf{E}}]_{\mathbf{1}} - [\hat{r}]_{\mathbf{1}}\right\|_2 \leq \frac{\sqrt{2\hat{\varepsilon}/\sigma_{\mathcal{R}}}}{\sin\left(\theta_2(P^0, P^1)/2\right)}.$$

Hence, applying the upper bound in Lemma 3.5 yields

$$\ell_P(r^{\mathsf{E}}, \mathsf{RL}_P(\hat{r})) \leq \frac{1}{2\eta}\left\|[r^{\mathsf{E}}]_{\mathcal{U}_P} - [\hat{r}]_{\mathcal{U}_P}\right\|_2^2 \leq \frac{\hat{\varepsilon}}{\eta\sigma_{\mathcal{R}}\sin\left(\theta_2(P^0, P^1)/2\right)^2}.$$

□

G Proof of Theorem 3.10

Theorem 3.10. *Let $K = 1$, $D := \max_{r, r' \in \mathcal{R}} \|r - r'\|_2$, and suppose that Assumptions 2.1,2.2, and 3.4 hold. If $\ell_{P^0}(\hat{r}, \mu^{\mathsf{E}}) \leq \hat{\varepsilon}$, then $\hat{r}$ is ε_P-transferable to $P \in \Delta^{\mathcal{S}\times\mathcal{A}}_{\mathcal{S}}$ with*

$$\varepsilon_P = 2\max\left\{2\hat{\varepsilon}/\sigma_{\mathcal{R}}, D^2 \sin\left(\theta_{\max}(P^0, P)\right)^2\right\}/\eta.$$

Proof. Similar to Theorem 3.9, it follows from Lemma 3.5 that $\left\|[r^{\mathsf{E}}]_{\mathcal{U}_{P^0}} - [\hat{r}]_{\mathcal{U}_{P^0}}\right\|_2 \leq \sqrt{2\hat{\varepsilon}/\sigma_{\mathcal{R}}}$. By Proposition D.10, we then have that

$$\begin{aligned}\left\|[r^{\mathsf{E}}]_{\mathcal{U}_P} - [\hat{r}]_{\mathcal{U}_P}\right\|_2 &\leq \sin\left(\theta_{\max}(P, P^0)\right)\left\|r^{\mathsf{E}} - \hat{r}\right\|_2 + \left\|[r^{\mathsf{E}}]_{\mathcal{U}_{P^0}} - [\hat{r}]_{\mathcal{U}_{P^0}}\right\|_2 \\ &\leq \sin\left(\theta_{\max}(P, P^0)\right) D + \sqrt{2\hat{\varepsilon}/\sigma_{\mathcal{R}}}.\end{aligned}$$

Hence, applying Lemma 3.5 again yields

$$\begin{aligned}\ell_P\left(r^{\mathsf{E}}, \mathsf{RL}_P(\hat{r})\right) &\leq \frac{1}{2\eta}\left\|[r^{\mathsf{E}}]_{\mathcal{U}_P} - [\hat{r}]_{\mathcal{U}_P}\right\|_2^2 \\ &\leq \frac{\left(D\sin\left(\theta_{\max}(P, P^0)\right) + \sqrt{2\hat{\varepsilon}/\sigma_{\mathcal{R}}}\right)^2}{2\eta} \\ &\leq \frac{2\max\left\{D^2\sin\left(\theta_{\max}(P, P^0)\right)^2, 2\hat{\varepsilon}/\sigma_{\mathcal{R}}\right\}}{\eta}.\end{aligned}$$

Furthermore, the bound on the maximal principal angle follows from Proposition 3.8. □

H Proof of Theorem 4.1

Theorem 4.1. *Suppose that $N^{\mathsf{E}} = \Omega\big(K\log(|\mathcal{S}||\mathcal{A}|/\hat{\delta})/\hat{\varepsilon}^2\big)$ and $H^{\mathsf{E}} = \Omega\big(\log(K/\hat{\varepsilon})/\log(1/\gamma)\big)$. Running Algorithm 1 for $T = \Omega\big(K^2/\hat{\varepsilon}^2\big)$ iterations with step-size $\alpha = 1/(K\sqrt{T})$, where $\delta_{opt} = \mathcal{O}\big(\hat{\delta}\hat{\varepsilon}^2/K^3\big)$, $\varepsilon_{opt} = \mathcal{O}(\hat{\varepsilon}/K)$, $N = \Omega\big(K\log(K|\mathcal{S}||\mathcal{A}|/(\hat{\delta}\hat{\varepsilon}))/\hat{\varepsilon}^2\big)$, and $H = H^{\mathsf{E}}$, it holds with probability at least $1 - \hat{\delta}$ that $\ell_{P^k}(\hat{r}, \mu^{\mathsf{E}}_{P^k}) \leq \hat{\varepsilon}$, for $k = 0, \ldots, K-1$.*

The proof of Theorem 4.1 is inspired by [Syed and Schapire, 2007, Theorem 2]. However, in contrast to Syed and Schapire [2007], we consider the regularized problem with multiple experts, we use the suboptimality as the convergence metric, and we use a projected gradient descent update (instead of multiplicative weights). The proof hinges on Hoeffding's inequality and a regret bound for online gradient descent, which are provided in Theorem H.1 and H.2 below.

Theorem H.1 (Hoeffding's inequality [Hoeffding, 1963])**.** *Let $X_0, \ldots, X_{M-1}$ be independent random variables with $X_l \in [a,b]$ and let $S_M := X_0 + \ldots + X_{M-1}$. Then,*

$$\Pr\left(|S_M - \mathbb{E}S_M| \geq c\right) \leq 2\exp\left(-\frac{2c^2}{M(b-a)^2}\right).$$

Theorem H.2 (Online gradient descent [Zinkevich, 2003])**.** *Consider some bounded closed convex set $\mathcal{X} \subset \mathbb{R}^n$ with $D := \max_{x,x' \in \mathcal{X}} \|x - x'\|_2$. Moreover, let $\Pi_{\mathcal{X}} : \mathbb{R}^n \to \mathcal{X}$ be the orthogonal projection onto $\mathcal{X}$. For any sequence of convex differentiable functions $f_0, \ldots, f_{T-1} : \mathcal{X} \to \mathbb{R}$ satisfying $\max_{x \in \mathcal{X}} \|\nabla f_t(x)\|_2 \leq G$, the online projected gradient descent update*

$$x_{t+1} \leftarrow \Pi_{\mathcal{X}}\left(x_t - \alpha \nabla f_t(x_t)\right),$$

with step-size $\alpha = D/(G\sqrt{T})$ satisfies

$$\sum_{t=0}^{T-1} f_t(x_t) - \min_{x^* \in \mathcal{X}} \sum_{t=0}^{T-1} f_t(x^*) \leq DG\sqrt{T}.$$

Proof of Theorem 4.1. The proof is in three steps. First, we use Hoeffding's inequality to prove concentration of the empirical occupancy measures around the true occupancy measures. Then, we use the union bound to upper bound the probability that any of our bounds fails to hold. Finally, we prove the convergence rate of Algorithm 1 using the regret bound in Theorem H.2.

Step 1: Let $\mathcal{D} = \{(s_0, a_0, \ldots, s_{H-1}, a_{H-1})\}_{i=0}^{N-1}$ be sampled from some policy π^μ and recall that the corresponding empirical occupancy measure is defined as

$$\hat{\mu}_{\mathcal{D}}(s,a) = \frac{1-\gamma}{N} \sum_{i=0}^{N-1} \sum_{t=0}^{H-1} \gamma^t \mathbb{1}\{s_t^i = s, a_t^i = a\}.$$

It will be convenient to define the truncated occupancy measure

$$\mu_H(s,a) = (1-\gamma) \sum_{t=0}^{H-1} \gamma^t \mathbb{P}_{\nu_0}^{\pi^\mu}\{s_t^i = s, a_t^i = a\}.$$

For K data sets $\mathcal{D}_1, \ldots, \mathcal{D}_K$ sampled from π^{μ_k} we then have

$$\begin{aligned}\max_{r \in \mathcal{R}} \sum_{k=0}^{K-1} \langle r, \mu_k - \hat{\mu}_{\mathcal{D}_k} \rangle &\overset{(i)}{\leq} \max_{r \in \mathcal{R}} \|r\|_1 \left\| \sum_{k=0}^{K-1} (\mu_k - \hat{\mu}_{\mathcal{D}_k}) \right\|_\infty \overset{(ii)}{\leq} \left\| \sum_{k=0}^{K-1} (\mu_k - \hat{\mu}_{\mathcal{D}_k}) \right\|_\infty \\ &\overset{(iii)}{\leq} \underbrace{\left\| \sum_{k=0}^{K-1} (\mu_k - \mu_{H,k}) \right\|_\infty}_{I_1} + \underbrace{\left\| \sum_{k=0}^{K-1} (\mu_{H,k} - \hat{\mu}_{\mathcal{D}_k}) \right\|_\infty}_{I_2},\end{aligned}$$

where (i) follows from Hölder's inequality, (ii) from our definition of $\mathcal{R}$ as the 1-norm ball, and (iii) from the triangle inequality. Since $\|\mu - \mu_H\|_\infty \leq \gamma^H$, we have $I_1 \leq \gamma^H K$. Moreover, applying Hoeffding's inequality to the $M = KN$ independent random variables

$$X_{kN+i} = \frac{1-\gamma}{N} \sum_{t=0}^{H-1} \gamma^t \mathbb{1}\{s_t^{k,i} = s, a_t^{k,i} = a\},\ i \in [N], k \in [K],$$

with $X_i \in [0, 1/N]$, we arrive at

$$\Pr\left(|S_M - \mathbb{E}S_M| \geq \varepsilon_{\text{stat}}/2\right) = \Pr\left(\left|\sum_{k=0}^{K-1} \hat{\mu}_{\mathcal{D}_k}(s,a) - \mu_{K,k}(s,a)\right| \geq \varepsilon_{\text{stat}}/2\right) \leq 2\exp\left(-\frac{\varepsilon_{\text{stat}}^2 N}{2K}\right).$$

Hence, applying the union bound over all $|\mathcal{S}||\mathcal{A}|$ components of the occupancy measure yields

$$\Pr(I_2 < \varepsilon_{\text{stat}}/2) = 1 - \Pr(I_2 \geq \varepsilon_{\text{stat}}/2) \geq 1 - 2|\mathcal{S}||\mathcal{A}| \exp\left(-\frac{\varepsilon_{\text{stat}}^2 N}{2K}\right).$$

Therefore, to ensure that with probability at least $1-\delta_{\text{stat}}$ it holds that

$$\max_{r\in\mathcal{R}} \sum_{k=0}^{K-1} \langle r, \mu_k - \hat{\mu}_{\mathcal{D}_k}\rangle \leq \varepsilon_{\text{stat}},$$

it suffices to choose

$$N \geq \frac{2K\log\left(2|\mathcal{S}||\mathcal{A}|/\delta_{\text{stat}}\right)}{\varepsilon_{\text{stat}}^2} \quad \text{and} \quad H \geq \frac{\log\left(2K/\varepsilon_{\text{stat}}\right)}{\log(1/\gamma)}.$$

This concentration result applies to both empirical occupancy measures generated from the expert data sets $\mathcal{D}^{\mathsf{E}}_k$, as well as the data sets $\mathcal{D}_{k,t}$ generated by Algorithm 1.

Step 2: When analyzing Algorithm 1 there are three sources of stochasticity. The first two are due to the randomness in the data sets $\mathcal{D}^{\mathsf{E}}_k$ and $\mathcal{D}_{k,t}$, and the third is due to the randomness in the forward RL algorithm, $\mathsf{A}^{\varepsilon_{\text{opt}},\delta_{\text{opt}}}_{P^k}$, that upon a query with the reward r_t outputs a policy $\pi_{k,t}$ such that with probability at least $1-\delta_{\text{opt}}$ it holds $\ell_{P^k}(r_t, \mu^{\pi_{k,t}}) \leq \varepsilon_{\text{opt}}$. Let's denote the event that $\max_{r\in\mathcal{R}}\sum_{k=0}^{K-1}\langle r, \mu^{\mathsf{E}}_{P^k} - \hat{\mu}_{\mathcal{D}^{\mathsf{E}}_k}\rangle > \varepsilon_{\text{stat},\mathsf{E}}$ by $\mathcal{E}_{\text{stat},\mathsf{E}}$, the event that $\max_{r\in\mathcal{R}}\sum_{k=0}^{K-1}\langle r, \mu^{\pi_{k,t}} - \hat{\mu}_{\mathcal{D}_{k,t}}\rangle > \varepsilon_{\text{stat}}$ by $\mathcal{E}_{\text{stat},t}$, and the event that $\ell_{P^k}(r_t, \mu^{\pi_{k,t}}) > \varepsilon_{\text{opt}}$ by $\mathcal{E}_{\text{opt},k,t}$. Moreover, let us assume that $\mathcal{E}_{\text{stat},\mathsf{E}}$ happens with probability at most $\delta_{\text{stat},\mathsf{E}}$, $\mathcal{E}_{\text{stat},t}$ happens with probability at most δ_{stat}, and $\mathcal{E}_{\text{opt},k,t}$ happens with probability at most δ_{opt}. By union bound, the probability of the event

$$\mathcal{F} := \neg\mathcal{E}_{\text{stat},\mathsf{E}} \wedge \bigwedge_{t=0}^{T-1} \neg\mathcal{E}_{\text{stat},t} \wedge \bigwedge_{t=0}^{T-1}\bigwedge_{k=0}^{K-1} \neg\mathcal{E}_{\text{opt},k,t},$$

that none of the above events happens is lower bounded by

$$\begin{aligned}
\Pr(\mathcal{F}) &= 1 - \Pr\left(\mathcal{E}_{\text{stat},\mathsf{E}} \vee \bigvee_{t=0}^{T-1} \mathcal{E}_{\text{stat},t} \vee \bigvee_{t=0}^{T-1}\bigvee_{k=0}^{K-1} \mathcal{E}_{\text{opt},k,t}\right) \\
&\geq 1 - \left(\Pr(\mathcal{E}_{\text{stat},\mathsf{E}}) + \sum_{t=0}^{T-1}\Pr(\mathcal{E}_{\text{stat},t}) + \sum_{t=0}^{T-1}\sum_{k=0}^{K-1}\Pr(\mathcal{E}_{\text{opt},k,t})\right) \\
&\geq 1 - \left(\delta_{\text{stat},\mathsf{E}} + T\delta_{\text{stat}} + KT\delta_{\text{opt}}\right).
\end{aligned}$$

Hence, to ensure that $\mathcal{F}$ happens with probability at least $1-\hat{\delta}$, it suffices to choose

$$\begin{aligned}
N &\geq \frac{2K\log\left(6|\mathcal{S}||\mathcal{A}|/\hat{\delta}\right)}{\varepsilon_{\text{stat},\mathsf{E}}^2} \quad \text{and} \quad H \geq \frac{\log\left(2K/\varepsilon_{\text{stat},\mathsf{E}}\right)}{\log(1/\gamma)}, \\
N_t &\geq \frac{2K\log\left(6T|\mathcal{S}||\mathcal{A}|/\hat{\delta}\right)}{\varepsilon_{\text{stat}}^2} \quad \text{and} \quad \delta_{\text{opt}} = \frac{\hat{\delta}}{3KT}.
\end{aligned}$$

Step 3: Note that we can bound $\|g_t\|_2 \leq \|g_t\|_1 \leq \sum_{k=0}^{K-1}\left\|\hat{\mu}_{\mathcal{D}^{\mathsf{E}}_k}\right\|_1 + \|\hat{\mu}_{k,t}\|_1 \leq 2K =: G$ and the diameter of $\mathcal{R}$ is $D=2$. Hence, given that event $\mathcal{F}$ happens, we can bound the suboptimalities of the

K experts under the reward, $\hat{r}$, recovered by Algorithm 1 with stepsize $\alpha = D/(G\sqrt{T})$ as follows

$$
\begin{aligned}
&\sum_{k=0}^{K-1} \ell_{P^k}(\hat{r}, \mu^{\mathsf{E}}_{P^k}) \\
&= \sum_{k=0}^{K-1} \left[\max_{\mu \in \mathcal{M}_{P^k}} \langle \hat{r}, \mu - \mu^{\mathsf{E}}_{P^k} \rangle - \bar{h}(\mu) + \bar{h}(\mu^{\mathsf{E}}_{P^k}) \right] \\
&\overset{(i)}{\leq} \varepsilon_{\text{stat},\mathsf{E}} + \sum_{k=0}^{K-1} \left[\max_{\mu \in \mathcal{M}_{P^k}} \langle \hat{r}, \mu - \hat{\mu}_{\mathcal{D}^{\mathsf{E}}_k} \rangle - \bar{h}(\mu) + \bar{h}(\mu^{\mathsf{E}}_{P^k}) \right] \\
&\overset{(ii)}{\leq} \varepsilon_{\text{stat},\mathsf{E}} + \sum_{k=0}^{K-1} \frac{1}{T} \sum_{t=0}^{T-1} \left[\max_{\mu \in \mathcal{M}_{P^k}} \langle r_t, \mu - \hat{\mu}_{\mathcal{D}^{\mathsf{E}}_k} \rangle - \bar{h}(\mu) + \bar{h}(\mu^{\mathsf{E}}_{P^k}) \right] \\
&= \varepsilon_{\text{stat},\mathsf{E}} + \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k=0}^{K-1} \left[\max_{\mu \in \mathcal{M}_{P^k}} \langle r_t, \mu - \hat{\mu}_{\mathcal{D}^{\mathsf{E}}_k} \rangle - \bar{h}(\mu) + \bar{h}(\mu^{\mathsf{E}}_{P^k}) \right] \\
&\overset{(iii)}{\leq} \varepsilon_{\text{stat},\mathsf{E}} + K\varepsilon_{\text{opt}} + \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k=0}^{K-1} \left[\langle r_t, \mu_{k,t} - \hat{\mu}_{\mathcal{D}^{\mathsf{E}}_k} \rangle - \bar{h}(\mu_{k,t}) + \bar{h}(\mu^{\mathsf{E}}_{P^k}) \right] \\
&\overset{(iv)}{\leq} \varepsilon_{\text{stat},\mathsf{E}} + K\varepsilon_{\text{opt}} + \varepsilon_{\text{stat}} + \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k=0}^{K-1} \left[\langle r_t, \hat{\mu}_{\mathcal{D}_{k,t}} - \hat{\mu}_{\mathcal{D}^{\mathsf{E}}_k} \rangle - \bar{h}(\mu_{k,t}) + \bar{h}(\mu^{\mathsf{E}}_{P^k}) \right] \\
&\overset{(v)}{\leq} \varepsilon_{\text{stat},\mathsf{E}} + K\varepsilon_{\text{opt}} + \varepsilon_{\text{stat}} + \frac{DG}{\sqrt{T}} + \min_{r \in \mathcal{R}} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k=0}^{K-1} \left[\langle r, \hat{\mu}_{\mathcal{D}_{k,t}} - \hat{\mu}_{\mathcal{D}^{\mathsf{E}}_k} \rangle - \bar{h}(\mu_{k,t}) + \bar{h}(\mu^{\mathsf{E}}_{P^k}) \right] \\
&\overset{(vi)}{\leq} 2\varepsilon_{\text{stat},\mathsf{E}} + K\varepsilon_{\text{opt}} + 2\varepsilon_{\text{stat}} + \frac{DG}{\sqrt{T}} + \min_{r \in \mathcal{R}} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k=0}^{K-1} [\langle r, \mu_{k,t} - \mu^{\mathsf{E}}_{P^k} \rangle - \bar{h}(\mu_{k,t}) + \bar{h}(\mu^{\mathsf{E}}_{P^k})] \\
&\overset{(vii)}{\leq} 2\varepsilon_{\text{stat},\mathsf{E}} + K\varepsilon_{\text{opt}} + 2\varepsilon_{\text{stat}} + \frac{DG}{\sqrt{T}} \\
&\qquad\qquad + \underbrace{\min_{r \in \mathcal{R}} \sum_{k=0}^{K-1} [\langle r, \bar{\mu}_k - \mu^{\mathsf{E}}_{P^k} \rangle - \bar{h}(\bar{\mu}_k) + \bar{h}(\mu^{\mathsf{E}}_{P^k})]}_{\leq 0}, \quad \text{with } \bar{\mu}_k := \frac{1}{T} \sum_{t=0}^{T-1} \mu_{k,t}, \\
&\overset{(viii)}{\leq} 2\varepsilon_{\text{stat},\mathsf{E}} + K\varepsilon_{\text{opt}} + 2\varepsilon_{\text{stat}} + \frac{DG}{\sqrt{T}} = 2\varepsilon_{\text{stat},\mathsf{E}} + K\varepsilon_{\text{opt}} + 2\varepsilon_{\text{stat}} + \frac{4K}{\sqrt{T}}.
\end{aligned}
$$

Here, the inequalities (i), (iv), and (vi) follow from the concentration bound established in step 1. Moreover, inequality (ii) holds since $\hat{r} \mapsto \max_{\mu \in \mathcal{M}_{P^k}} \langle \hat{r}, \mu - \mu^{\mathsf{E}}_{P^k} \rangle - \bar{h}(\mu) + \bar{h}(\mu^{\mathsf{E}}_{P^k})$ is the pointwise maximum of affine functions and therefore convex. Furthermore, (iii) follows from ε_{opt}-optimality of $\mu_{k,t}$, (v) from Theorem H.2, and (vii) from concavity of the mapping $\mu_{k,t} \mapsto \langle r, \mu_{k,t} - \mu^{\mathsf{E}}_{P^k} \rangle - \bar{h}(\mu_{k,t})$. Finally, $(viii)$ holds because all experts are optimal for the reward r^{E}. In conclusion, to ensure that with probability at least $1 - \hat{\delta}$ it holds that $\ell_{P^k}(\hat{r}, \mu^{\mathsf{E}}_{P^k}) \leq \sum_{k=0}^{K-1} \ell_{P^k}(\hat{r}, \mu^{\mathsf{E}}_{P^k}) \leq \hat{\varepsilon}$ it suffices to choose $T = \frac{256K^2}{\hat{\varepsilon}^2}$, $\alpha = \frac{\hat{\varepsilon}}{16K^2}$, $N = \frac{128K \log\left(6|\mathcal{S}||\mathcal{A}|/\hat{\delta}\right)}{\hat{\varepsilon}^2}$, $H = H_t = \frac{\log(16K/\hat{\varepsilon})}{\log(1/\gamma)}$, $N_t = \frac{128K \log\left(1536K^2|\mathcal{S}||\mathcal{A}|/(\hat{\delta}\hat{\varepsilon}^2)\right)}{\hat{\varepsilon}^2}$, $\delta_{\text{opt}} = \frac{\hat{\delta}\hat{\varepsilon}^2}{768K^3}$, $\varepsilon_{\text{opt}} = \frac{\hat{\varepsilon}}{4K}$. □

I Experimental details

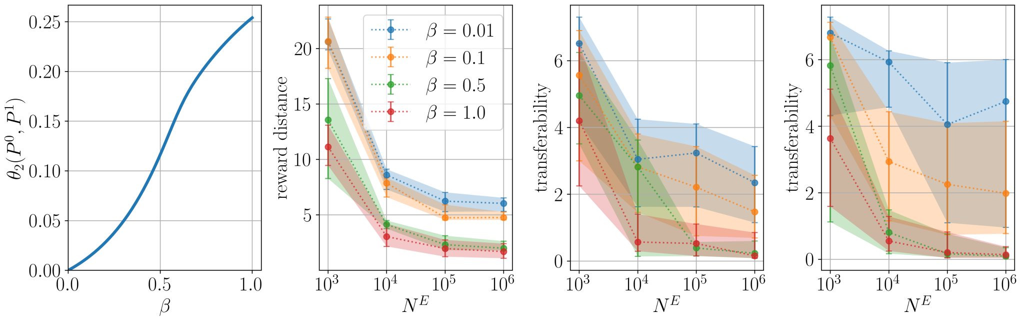
Setup To validate our results experimentally, we are using a stochastic adaption of the `WindyGridworld` environment [Sutton and Barto, 2018].[3] In particular, we consider a 6x6 grid with

[3] All our experiments were carried out – within a day – on a MacBook Pro with an Apple M1 Pro chip and 32 GB of RAM.

4 actions (Up, Down, Left, Right), a wind direction (North, East, South, West), and a wind strength $\beta \in [0, 1]$. When the agent takes an action, with probability $(1 - \beta)$, it moves to the intended grid cell, and with probability β, the wind pushes the agent one step further in the direction of the wind. This means that the transition law is a convex combination of two laws: $(1 - \beta)P^{\text{Gridworld}} + \beta P^{\text{Wind}}$, where $P^{\text{Gridworld}}$ and P^{Wind} represent the transition laws for a deterministic `Gridworld` and a deterministic `WindyGridworld`. For our experiments, we then consider the pairs of expert transition laws $P^0_\beta = (1-\beta)P^{\text{Gridworld}}+\beta P^{\text{North}}$ and $P^1_\beta = (1-\beta)P^{\text{Gridworld}}+\beta P^{\text{East}}$ with β in $\{0.01, 0.1, 0.5, 1.0\}$. As shown in Figure 3(a), the second principal angle between P^0_β and P^1_β, calculated using a singular value decomposition [Knyazev and Argentati, 2002], increases as the wind strength β increases.

Inverse reinforcement learning We observed that under a small second principal angle, the recovered reward heavily depends on both the expert reward and the reward initialization. Hence, we sample 10 independent expert rewards, each generated by first sampling a random set of 10 state-action pairs and then randomly assigning a reward of ± 1. Using Shannon entropy regularization with $\tau = 0.3$, we then use soft policy iteration to get expert policies for each combination of expert reward and wind strength β. For each of these expert policies, we then generate expert data sets with $N^{\mathsf{E}} \in \{10^3, 10^4, 10^5, 10^6\}$ trajectories of length $H = 100$. Next, we run Algorithm 1, with soft policy iteration as a subroutine, for $30'000$ iterations, where rewards are initialized by sampling from a standard normal distribution. As a reward class, we choose the $\|\cdot\|_1$-ball with radius 10^3 (essentially unbounded), as a stepsize $\alpha = 0.05$ for the first $15'000$ iterations and $\alpha = 0.005$ for the second half. Moreover, we sample $N = 100$ new trajectories of horizon $H = 100$ at each gradient step. Figure 3(b) illustrates the distances between the recovered $\hat{r}$ and the experts' reward r^{E}, measured in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathbf{1}$. It is evident that the recovered reward gets closer to the experts' reward as the number of expert demonstrations increases. Moreover, we observe that the recovered reward is closer to the experts' reward when the second principal angle between the experts is larger, as expected from Lemma F.1.

Transferability We evaluate the transferability of the obtained reward by considering two new environments. First, a south wind setting P^{South} with wind strength $\beta = 1$, and second, a deterministic gridworld P^{Shifted}, with cyclically shifted actions, i.e., Right→ Down, Up→ Right, Left→ Up, Down→ Left. In Figure 3(c) and (d), we illustrate the transferability in terms of $\ell_{P^{\text{South}}}(r^{\mathsf{E}}, \mathsf{RL}_{P^{\text{South}}}(\hat{r}))$ and $\ell_{P^{\text{Shifted}}}(r^{\mathsf{E}}, \mathsf{RL}_{P^{\text{Shifted}}}(\hat{r}))$, respectively. We observe that for both environments the transferability improves with a larger second principal angle, thus confirming our theoretical result in Theorem 3.9. The effect is even more pronounced for the shifted environment. While confirming our results, the experiments also reveal a high sample complexity in terms of expert demonstrations. This is to be expected, as IRL aims to match the expert's empirical occupancy measure, leading to overfitting when there are not enough demonstrations [Ho and Ermon, 2016]. This issue can be mitigated by reducing the dimension of the reward class (see e.g. [Abbeel and Ng, 2004]).



(a) (b) (c) (d)

Figure 3: (a) shows the second principal angle between P^0_β and P^1_β for varying wind strength β. Furthermore, (b) shows the distance between $\hat{r}$ and r^{E} in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}/\mathbf{1}$ for a varying number of expert demonstrations N^{E} and wind strength β. Moreover, (c) and (d) show the transferability to P^{South} and P^{Shifted} in terms of $\ell_{P^{\text{South}}}(r^{\mathsf{E}}, \mathsf{RL}_{P^{\text{South}}}(\hat{r}))$ and $\ell_{P^{\text{Shifted}}}(r^{\mathsf{E}}, \mathsf{RL}_{P^{\text{Shifted}}}(\hat{r}))$, respectively. The dots indicate the median and the shaded areas the 0.2 and 0.8 quantiles over the 10 independent realizations.