On Feasible Rewards in Multi-Agent Inverse Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multi-agent inverse reinforcement learning (MAIRL) aims to recover agent reward functions from expert demonstrations. We characterize the feasible reward set in Markov games, identifying all reward functions that rationalize a given equilibrium. However, equilibrium-based observations are often ambiguous: a single Nash equilibrium can correspond to many reward structures, potentially changing the game's nature in multi-agent systems. We address this by introducing entropy-regularized Markov games, which yield a unique equilibrium while preserving strategic incentives. For this setting, we provide a sample complexity analysis detailing how errors affect learned policy performance. Our work establishes theoretical foundations and practical insights for MAIRL.

1 Introduction

Multi-agent Reinforcement Learning (MARL) has garnered substantial attention in recent years due to its capacity to model scenarios involving interacting agents. Notable successes have been achieved across diverse domains, including autonomous driving [Shalev-Shwartz et al., 2016, Zhou et al., 2020], internet marketing [Jin et al., 2018], multi-robot control [Dawood et al., 2023], traffic control [Wang et al., 2019], and multi-player games [Baker et al., 2019, Samvelyan et al., 2019]. A critical prerequisite for these applications is the careful design of reward functions, a task that proves challenging even in single-agent settings [Amodei et al., 2016, Hadfield-Menell et al., 2017] and becomes significantly more complex in multi-agent environments where each agent's reward function must be tailored to their specific, potentially conflicting, objectives.

In numerous real-world scenarios, expert demonstrations of optimal behavior may be observable, while the underlying reward function driving these actions remains unknown. This is precisely the domain of Inverse Reinforcement Learning (IRL) [Ng and Russell, 2000]. The objective of IRL is to recover plausible reward functions that can rationalize the observed behavior as optimal. However, early research in IRL highlighted a fundamental challenge: the problem is inherently ill-posed, as multiple reward functions can potentially explain the same observed behavior. Subsequent research has therefore focused on reformulating the IRL problem to enhance its practicality and applicability in real-world contexts [Abbeel and Ng, 2004, Ziebart et al., 2008, Ramachandran and Amir, 2007, Ratliff et al., 2006a].

The extension of IRL to the multi-agent setting introduces novel complexities, particularly concerning the definition of optimality and the multiplicity of Nash equilibria, given that each agent's optimal strategy is dependent on the strategies of all other agents. This necessitates the adoption of game-theoretic solution concepts, with the Nash equilibrium being the most prevalent [Goktas et al., 2024, Song et al., 2018, Ramponi et al., 2023, Fu et al., 2021]. In contrast to the substantial progress in understanding the theoretical underpinnings of single-agent IRL, the theoretical foundations of Multi-Agent Inverse Reinforcement Learning remain comparatively underexplored. In single-

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

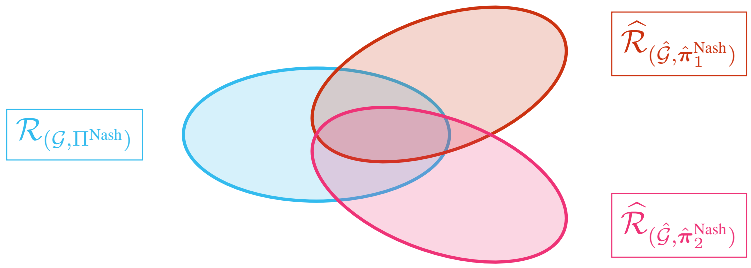

Figure 1: Feasible Reward Sets of true set of Nash equilibria and the recovered feasible reward sets for two different observed Nash equilibria.

agent IRL Metelli et al. [2021] established explicit conditions for feasible reward functions and developed efficient algorithms for unknown transition models and expert policies, assuming access to a generative model. This work has been extended to settings without a generative model [Lindner et al., 2022], stricter optimality metrics [Zhao et al., 2024, Metelli et al., 2023], and offline settings Lazzati et al. [2024b], Zhao et al. [2024]. However, these studies are confined to single-agent scenarios and evaluate performance based on criteria that are not directly transferable to general-sum Markov games. In Appendix B we provide an extensive discussion on related works.

This paper aims to bridge the existing gap between the theoretical understanding of IRL in single-agent systems and its application to multi-agent systems. Specifically, we first address the research question:

(Q1) *What constitutes a rigorous definition of Multi-Agent Inverse Reinforcement Learning?*

To address this question independently of a specific MAIRL algorithm, we derive properties of the feasible reward set, drawing inspiration from the initial work in single-agent settings by Metelli et al. [2021]. First, we define a straight-forward extension from the single-agent feasible reward set to the multi-agent setting as all the rewards under which a *single* Nash equilibrium expert is optimal, meaning it is indeed a Nash equilibrium. Then, we demonstrate that a single observed equilibrium is insufficient for identifying expressive reward sets, as distinct observed equilibria can induce different feasible reward sets (see Fig. 1 for an illustration). This can result in a Nash Gap of order $(1-\gamma)^{-1}$ due to the multiplicity of the Nash equilibria. To mitigate the equilibrium selection problem, we introduce entropy-regularized Multi-agent IRL. Then, we formally characterize the inherent increase in complexity associated with the multi-agent setting. Within this framework, we characterize feasible rewards and establish sample complexity bounds that account for errors in transition dynamics and policy estimation, assuming access to a generative model.

In single-agent settings, the introduction of entropy-regularized experts has facilitated the derivation of conditions under which the reward function is identifiable [Cao et al., 2021, Rolland et al., 2022]. This motivates our second research question:

(Q2) *Is reward identifiability achievable in Multi-agent settings?*

We provide a partial positive answer to this question. We show that in general-sum Markov Games without additional structural assumptions, reward identifiability is only possible in the average reward sense. However, we prove that if the underlying reward structure is linearly separable, meaning that the reward can be decomposed into a reward for player 1 and player 2, $R(s,a,b) = R_A(s,a) + R_B(s,b)$, then reward identification (up to additive constants) is possible.

2 Preliminaries

We present the essential background and notation used throughout this paper, also summarized in Appendix A.

Mathematical background. Let $\mathcal{X}$ be a finite set, then we denote by $\mathbb{R}^{\mathcal{X}}$ all functions mapping from $\mathcal{X}$ to $\mathbb{R}$. Additionally, we denote by $\Delta^{\mathcal{X}}$ the set of probability measures over $\mathcal{X}$. For $n \in \mathbb{N}$ we use $[n] := \{1, \dots, n\}$. We introduce for a (pre)metric space $(\mathcal{X}, d)$ with

$\mathcal{Y}, \mathcal{Y}' \subseteq \mathcal{X}$ two non-empty sets the *Hausdorff (pre)metric* $\mathcal{H}_d : 2^{\mathcal{X}} \times 2^{\mathcal{X}} \to [0, +\infty)$ as $\mathcal{H}_d(\mathcal{Y}, \mathcal{Y}') := \max \left\{ \sup_{y \in \mathcal{Y}} \inf_{\sup_{y' \in \mathcal{Y}'}} d(y, y'), \sup_{y' \in \mathcal{Y}'} \inf_{\sup_{y \in \mathcal{Y}}} d(y, y') \right\}$.

Markov Games. An infinite time, discounted n-person general-sum Markov Game [Shapley, 1953, Takahashi, 1964, Fink, 1964] without reward function (MG $\setminus R$) is characterized by a tuple $\mathcal{G} = (n, \mathcal{S}, \mathcal{A}, P, \gamma, \rho)$, where $n \in \mathbb{N}$ denotes the finite number of players; $\mathcal{S}$ the finite state space; $\mathcal{A} := \mathcal{A}^1 \times \ldots \times \mathcal{A}^n$ the joint action space of the individual action spaces $\mathcal{A}^i$; $P : \mathcal{S} \times \mathcal{A} \to \Delta^{\mathcal{S}}$ the transition model; γ is the discount factor and ρ is the initial state distribution. We will make use of the words persons, agents and players interchangeably. The strategy of a single agent, also called the policy, we denote by $\pi^i : S \to \Delta^{\mathcal{A}^i}$. A joint strategy is given by $\boldsymbol{\pi} = (\pi^1, \ldots, \pi^n) = (\pi^i, \pi^{-i})$, where π^{-i} refers to the (joint-) policy of all players except player i. A joint action is denoted by $\mathbf{a} = (a^1, \ldots, a^n) \in \mathcal{A}$. Therefore, the probability of a joint strategy is given by $\boldsymbol{\pi}(\mathbf{a} \mid s) := \prod_{j=1}^n \pi^j(a^j \mid s)$. Π^i denotes the set of all policies for agent i. The discounted probability of visiting a state-(joint-)action pair, given that the starting state is drawn from ρ, is defined as $\overline{w}^{\boldsymbol{\pi}, \rho}_{s,\mathbf{a}} = \sum_{t=0}^\infty \gamma^t \mathbb{P}(s, \mathbf{a}, \rho)$, where $\mathbb{P}$ denotes the probability of visiting the joint state action pair when drawing the initial state from ρ and following the joint policy $\boldsymbol{\pi}$. If the starting distribution is deterministic for a state s, we will omit the dependence on ρ and simply write $\overline{w}^{\boldsymbol{\pi}}_{s,\mathbf{a}}$.

Reward function. The reward function for an agent, $R^i : \mathcal{S} \times \mathcal{A} \to [-R^i_{\max}, R^i_{\max}]$, takes a state and a joint action as inputs, mapping them to a bounded real number. The joint reward is represented as $R = (R^1, \ldots, R^n)$. The uniform reward bound across all agents is defined by $R_{\max} := \max_{i \in [n]} R^i_{\max}$. A Markov game without reward $\mathcal{G}$ combined with a joint reward results in a standard Markov game denoted as $\mathcal{G} \cup R$.

Value functions and equilibrium concepts. For a Markov game $\mathcal{G} \cup R$ with a policy $\boldsymbol{\pi}$ we define the *Q-function* and the *value-function* of an agent i for a given state and action as $Q^{i,\boldsymbol{\pi}}_{\mathcal{G} \cup R}(s, \mathbf{a}) = \mathbb{E}^{\boldsymbol{\pi}}[\sum_{t=0}^\infty \gamma^t R^i(s, \mathbf{a}) \mid s, \mathbf{a}]$ and $V^{i,\boldsymbol{\pi}}_{\mathcal{G} \cup R}(s) = \sum_{\mathbf{a}} \boldsymbol{\pi}(\mathbf{a} \mid s) Q^{i,\boldsymbol{\pi}}_{\mathcal{G} \cup R}(s, \mathbf{a})$. If it is clear from the context what the underlying Markov Game is, we will omit the subscript.

Nash Equilibrium. Various types of equilibrium solutions have been proposed to model optimal strategies in Markov Games. In this work, we focus on the NE similarly to previous works on MAIRL [Goktas et al., 2024, Song et al., 2018]. A Nash strategy is one where no agent can improve their outcome by independently deviating from their strategy, assuming the strategies of the other agents remain unchanged. Formally, a policy $\boldsymbol{\pi}^{\text{Nash}}$ is a (perfect) Nash equilibrium strategy, if for every state s and every agent i $V^{i,\boldsymbol{\pi}^{\text{Nash}}}_{\mathcal{G} \cup R}(s) \geq V^{i,(\pi^i, \pi^{-i,\text{Nash}})}_{\mathcal{G} \cup R}(s) \quad \forall \pi^i \in \Pi^i$. To simplify the notation used in the remainder of this work, we will denote this as $V^i(\boldsymbol{\pi}^{\text{Nash}}) \geq V^i(\pi^i, \pi^{-i,\text{Nash}})$. The set of Nash equilibria, depending on the underlying reward, we will denote as $\Pi^{\text{Nash}}(R)$.

Entropy Regularized Markov Games. For better readability, let us consider the case $\boldsymbol{\pi} = (\mu, \nu)$ and $\mathcal{A}_2 := \mathcal{B}$. Then, the value function of player 1 in a λ entropy regularized Markov Game is defined as $V^{1,(\mu,\nu)}_\lambda(s) = \mathbb{E}^{(\mu,\nu)}[\sum_{t=0}^\infty \gamma^t R^1(s, a, b) - \lambda \log(\mu(a \mid s))]$. The value function for player 2 is defined analogously. Additionally, the Q-function is defined as $Q^{1,(\mu,\nu)}_\lambda(s, a, b) = R^1(s, a, b) + \sum_{s'} P(s' \mid s, a, b) V^{1,(\mu,\nu)}_\lambda(s')$.

Multi-Agent Inverse Reinforcement Learning. Given a Markov Game without a reward function $\mathcal{G}$ and a Nash equilibrium expert, the MAIRL problem is defined as the tuple $(\mathcal{G}, \boldsymbol{\pi}^{\text{Nash}})$. If we only have access to an estimated version of $(\mathcal{G}, \boldsymbol{\pi}^{\text{Nash}})$, we call it recovered MAIRL problem and denote it as $(\hat{\mathcal{G}}, \hat{\boldsymbol{\pi}}^{\text{Nash}})$ and analogously for entropy equilibria by replacing $\hat{\boldsymbol{\pi}}^{\text{Nash}}$.

3 Feasible Reward Set in Multi-Agent systems

This section addresses research question (Q1). First, we will define MAIRL for a single NE observation. Then, we will show, that this can be ill-suited for multi-agent systems due to the multiplicity of equilibria. This motivates to considers entropy regularized Markov Games to guarantee a unique equilibrium.

3.1 Nash equilibrium observations

In this section, we begin by revisiting single-agent Inverse Reinforcement Learning. The feasible reward set in single-agent IRL, first defined by Metelli et al. [2021] and later refined for different settings as e.g. the offline case in [Zhao et al., 2024, Metelli et al., 2023, Lazzati et al., 2024b], lacks a multi-agent counterpart for observed expert equilibria. We thus translate the single-agent definition, formalizing feasible rewards as in prior works [Lin et al., 2014, 2018].

Definition 3.1. Let a MAIRL problem $(\mathcal{G}, \boldsymbol{\pi}^{\mathrm{Nash}})$ with a single (observed) Nash equilirbium policy be given. Then, the feasible reward set for general-sum Markov Games is given by

$$\mathcal{R}_{(\mathcal{G},\boldsymbol{\pi}^{\mathrm{Nash}})} = \left\{ R \in \mathcal{R} \mid \forall i \in [n], \forall s \in \mathcal{S}, \forall \pi_i \in \Pi_i : V^{i,(\pi_i^{\mathrm{Nash}},\pi_{-i}^{\mathrm{Nash}})}_{\mathcal{G}\cup R}(s) > V^{i,(\pi_i,\pi_{-i}^{\mathrm{Nash}})}_{\mathcal{G}\cup R}(s) \right\}.$$

Here, $\boldsymbol{\pi}^{\mathrm{Nash}} \in \Pi^{\mathrm{Nash}}(R)$ is any Nash equilibrium, analogous to an optimal policy in single-agent IRL. A key difference in MAIRL is that different NEs can yield varying values, and we impose no restriction on the observed NE (pure or mixed) from $\Pi^{\mathrm{Nash}}(R)$. This is the first fundamental difference between IRL and MAIRL.

Since varying NE values make single-agent value-based gap objectives [Metelli et al., 2021, 2023, Zhao et al., 2024, Lazzati et al., 2024b] unsuitable for MAIRL, we adapt the Nash Gap, recently used in multi-agent imitation learning [Ramponi et al., 2023, Tang et al., 2024], as our objective.

Definition 3.2 (Nash Imitation Gap for MAIRL)**.** Let $\mathcal{G} \cup R$ be the underlying n-person general-sum Markov Game. Furthermore, let $\hat{\boldsymbol{\pi}}$ be the policy recovered from the corresponding MAIRL problem. Then we define the Nash Imitation Gap of $\hat{\boldsymbol{\pi}}$ as

$$\mathcal{E}(\hat{\boldsymbol{\pi}}) := \max_{i \in [n]} \max_{\pi^i \in \Pi^i} V^i_{\mathcal{G}\cup R}(\pi^i, \hat{\pi}^{-i}) - V^i_{\mathcal{G}\cup R}(\hat{\boldsymbol{\pi}}).$$

The definition possesses the desirable property that it equals 0 if $\hat{\boldsymbol{\pi}}$ is an NE, and, it is > 0 if $\hat{\boldsymbol{\pi}}$ is not an NE in the underlying Markov Game.

Normally, we cannot assume to know the expert equilibrium nor the transition function. Therefore, to analyze how estimation errors in the transition probability and expert policy affect the recovered feasible reward set, we relate them to our proposed optimality criterion.

Definition 3.3 (Optimality Criterion)**.** Let $\mathcal{R} := \mathcal{R}_{(\mathcal{G},\boldsymbol{\pi}^{\mathrm{Nash}})}$ be the exact feasible set and $\hat{\mathcal{R}} := \mathcal{R}_{(\hat{\mathcal{G}},\hat{\boldsymbol{\pi}})}$ the recovered feasible set after observing $N \geq 0$ samples from the underlying MAIRL problem $(\mathcal{G}, \boldsymbol{\pi}^{\mathrm{Nash}})$. We consider an algorithm to be (ϵ, δ, N)-correct after observing N samples if with a probability of at least $1 - \delta$ it holds:

$$\sup_{R \in \mathcal{R}} \inf_{\hat{R} \in \hat{\mathcal{R}}} \sup_{\hat{\boldsymbol{\pi}} \in \Pi^{\mathrm{Nash}}(\hat{R})} \mathcal{E}(\hat{\boldsymbol{\pi}}) \leq \epsilon\,, \quad \sup_{\hat{R} \in \hat{\mathcal{R}}} \inf_{R \in \mathcal{R}} \sup_{\hat{\boldsymbol{\pi}} \in \Pi^{\mathrm{Nash}}(\hat{R})} \mathcal{E}(\hat{\boldsymbol{\pi}}) \leq \epsilon,$$

where $\Pi^{\mathrm{Nash}}(\hat{R}) := \{\boldsymbol{\pi} \mid V^{\boldsymbol{\pi}}_{\hat{\mathcal{G}}\cup\hat{R}}(s) > V^{\tilde{\pi}^i,\pi_{-i}}_{\hat{\mathcal{G}}\cup\hat{R}}(s)\, \forall \tilde{\pi}^i \in \Pi^i\,, \forall s \in \mathcal{S}, \forall i \in [n]\}$.

The optimality criterion is in the Hausdorff metric style, see 2. The first condition ensures that the recovered feasible set captures a reward function that makes sure that the recovered policy is at most an ϵ-NE in the true Markov Game. However, this would support choosing a set that captures all possible reward functions $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. Consequently, the second condition ensures that this is not possible by requiring every recovered reward to also have a true reward function that captures the desired behavior. Additionally, note that the optimality criterion depends on $\hat{R}$ as $\hat{\boldsymbol{\pi}}$ is the Nash equilibrium of the recovered Markov Game $\hat{\mathcal{G}} \cup \hat{R}$.

The shortcomings of observing only a single equilibrium are discussed next. For completeness, this framework is further analyzed in Appendix D.

Feasible Reward Set and Equilibrium Ambiguity. In this section, we show that observing only a single equilibrium solution leads to a non-expressive feasible reward set. Fu et al. [2021] noted that relying on a single Nash equilibrium for inverse learning introduces inherent limitations. Here, we extend this finding to the feasible reward set, rather than focusing solely on specific reward functions.

Tang et al. [2024] showed that minimizing the regret gap is hard in Imitation Learning problems. However, in this work we do not consider the setting of mimicking the expert policy instead we examine the feasible reward set. In general, (MA)IRL can be more powerful as it allows to transfer the reward function to new environments. In Appendix H we provide an experiment in a simple

Grid World game that emphasizes this. Unfortunately, the next theorem shows, that learning under a recovered reward function from the feasible set stemming from a MAIRL problem with a single equilibrium observation can lead to a NE that has a Nash Gap of the order $(1-\gamma)^{-1}$ in the original Markov Game.

Proposition 3.4. *Let us consider any MAIRL algorithm* $\mathrm{Alg}_{\mathrm{MAIRL}}$ *that chooses* $\hat{R} \in \mathcal{R}_{(\hat{\mathcal{G}},\hat{\boldsymbol{\pi}}^{\mathrm{Nash}})}$ *that is not a constant reward, i.e.* $\hat{R} \neq C$ *for* $C \in [-R_{\max}, R_{\max}]$*. Furthermore, consider a MARL algorithm* $\mathrm{Alg}_{\mathrm{MARL}}$ *that guarantees learning a policy* $\tilde{\boldsymbol{\pi}} \in \Pi^{\mathrm{Nash}}(\hat{R})$*. Then, there exists a Markov Game, such that even if* $\hat{\boldsymbol{\pi}} \in \Pi_{\mathrm{Nash}}$ *and* $\hat{R} \in \mathcal{R}_{(\mathcal{G},\boldsymbol{\pi}^{\mathrm{Nash}})}$ *it holds true that* $\mathcal{E}(\tilde{\boldsymbol{\pi}})$ *is of order* $(1-\gamma)^{-1}$*.*

The construction of the underlying general-sum Markov Game can be found in Fig. 2. The idea of the proof is that Definition 3.1 only ensures that the recovered expert $\hat{\boldsymbol{\pi}}^{\mathrm{Nash}}$ is a NE under $\hat{R}$, an MAIRL algorithm cannot capture a meaningful reward for other equilibria that potentially have different values. Therefore, the constraints on the rewards inside the feasible reward set only gives a relation for a fixed strategy of the opponent. Consider a simple example with two players where player 2 plays action a_2 with probability one. Then, for player one the constraints only tells something about $R^1(s, a^{\mathrm{Nash}}, b) \geq R^1(s, a, b)\, \forall a \in \mathcal{A}_1$, but nothing on rewards $R^1(s, a, b')\, \forall (a, b') \in \mathcal{A}_1 \times \mathcal{A}_2$. This allows that the recovered reward functions allow for new equilibria, i.e. $\Pi^{\mathrm{Nash}}(\hat{R}) \supset \Pi^{\mathrm{Nash}}(R)$ that can be exploited in the original Markov Game. We empirically investigate this for the considered Game in Proposition 3.4. Summarized, the reward set captures too many reward functions and this flexibility in reward specification allows for undesirable scenarios, such as **changing the nature of the game**. This means the game can transform the set of all Nash equilibria e.g. from a coordination game into an anti-coordination variant. For further intuition we provided additional examples in Example E.1.

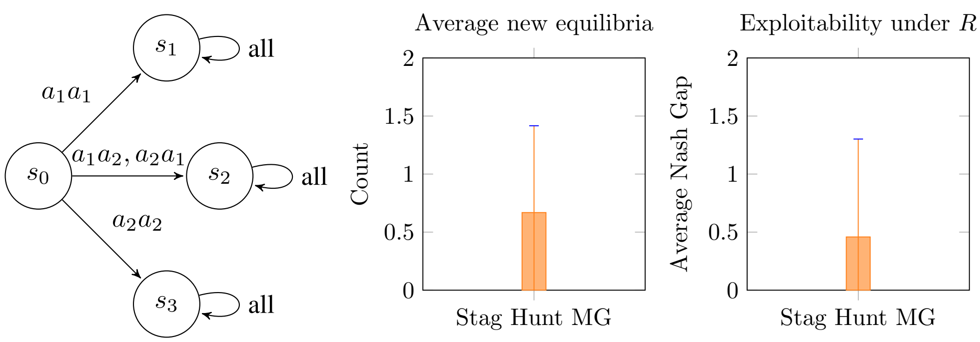

Figure 2: Failure of single equilibrium observation.

To address this, we propose the following definition of the feasible reward set.

Definition 3.5. Let $\Pi^{\mathrm{Nash}}(R)$ denote the set of Nash equilibrium in the underlying Markov Game $\mathcal{G} \cup R$. Then, the feasible reward set for general-sum Markov Games is defined as: $\mathcal{R}_{(\mathcal{G},\Pi^{\mathrm{Nash}})} = \left\{ R \in \mathcal{R} \mid \forall \boldsymbol{\pi}^{\mathrm{Nash}} \in \Pi^{\mathrm{Nash}},\, i \in [n],\, s \in \mathcal{S},\, \pi_i \in \Pi_i : V^{i,(\pi_i^{\mathrm{Nash}}, \pi_{-i}^{\mathrm{Nash}})}_{\mathcal{G}\cup R}(s) > V^{i,(\pi_i, \pi_{-i}^{\mathrm{Nash}})}_{\mathcal{G}\cup R}(s) \right\}$.

Note that, the subscript is now on the set of equilibria instead of a single one. This definition ensures that the feasible reward set aligns with all equilibrium solutions of the original game, rather than a single observed equilibrium. While this improves the interpretability of the feasible reward set, calculating even one Nash equilibrium is computationally intractable in general-sum Markov Games. Hence, this definition does not fully resolve the tractability issue in MAIRL.

The multiplicity of equilibria and the absence of a unique value pose a significant challenge in Inverse Reinforcement Learning, in particular which equilibria should be chosen, commonly referred to as the *Equilibrium Selection problem*. This ambiguity complicates reward inference, as different equilibria can lead to inconsistent or unreliable outcomes.

Leonardos et al. [2021] address this by proposing game structure modifications, such as *regularized Markov Games*. Techniques like entropy regularization refine the equilibrium set to a unique one, eliminating the selection problem. We will transfer this concept to MAIRL and investigate its implications in subsequent chapters.

3.2 Feasible Rewards for Entropy Regularized Games

This section examines *Entropy-Regularized Markov Games* and their unique *Quantal Response Equilibrium (QRE)*, which reflects bounded rationality. We show that QREs help avoid problematic reward configurations (cf. Fig. 2) and yield recovered rewards more similar to the true game rewards. We then formally define the feasible reward set for QRE experts and provide a sample complexity analysis for MAIRL to achieve the entropy version of Definition 3.3. Technically, this is not the Nash Gap anymore, instead it only measures the exploitability of a policy in the entropy regularized game.

Avoiding the ambiguity with QRE expert. We begin by demonstrating that entropy regularization, yielding a unique fully mixed QRE, provides sufficient structure for reward recovery to avoid the degenerate cases highlighted in Fig. 2. Recalling our empirical validations (Fig. 2), single, potentially pure, equilibrium observations permitted exploitable new equilibria in the original game. We now present analogous experiments using QRE observations (see Fig. 3).

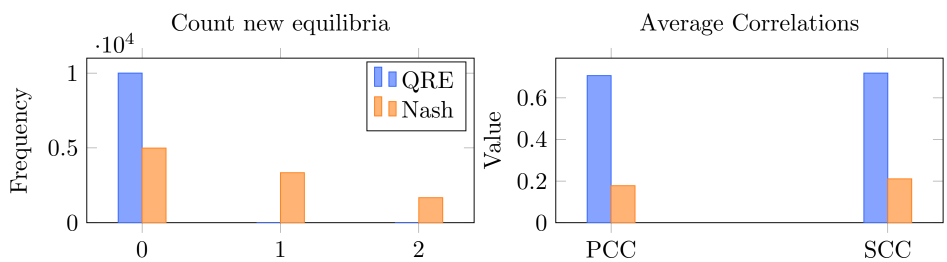

Figure 3: Recovered rewards under QRE equilibrium observations.

We can observe that in simple games the QRE ensures that no new pure equilibria arise and additionally the correlation of the recovered reward function and the true reward function measured by the Pearson Correlation Coefficient (PCC) and Spearman Correlation Coefficient (SCC) are significantly higher. The reason for this is that the QRE enforces the equilibrium observation to explore the environment better and gets rid of the equilibrium selection problem. This can be further motivated by assumptions needed also in the Multi-Agent Imitation Learning setting, where Tang et al. [2024] showed that under coverage assumption equilibria can successfully be recovered. This is automatically fulfilled if the observed expert is a QRE expert.

Characterization of feasible rewards. In the single-agent IRL setting, an explicit characterization of the reward function in entropy-regularized Markov Decision Processes (MDPs) was first derived by Cao et al. [2021]. The authors introduced entropy regularization to tackle the ill-posedness of the IRL problem. In particular, the authors established conditions under which the reward function is identifiable up to a constant. These conditions were subsequently simplified by Rolland et al. [2022], providing further insights into the structure of the reward function. This naturally motivates our second research question (Q2), which we will answer in Section 4. In this section we will first focus on the feasible reward set.

We begin our analysis in a manner similar to the single-agent case. For better readability we from now on assume to only have two players, with policies μ, ν. We can give an explicit characterization of the optimal policy for the agents. Next, we give this definition for player 1, it analogously holds also for player 2

$$\mu^*(a \mid s) = \frac{\exp\left(\frac{1}{\lambda}\sum_{b'\in\mathcal{B}}\nu^*(b' \mid s)Q_\lambda^{*,1}(s,a,b')\right)}{\sum_{a'\in\mathcal{A}}\exp\left(\frac{1}{\lambda}\sum_{b'\in\mathcal{B}}\nu^*(b' \mid s)Q_\lambda^{*,1}(s,a',b')\right)}, \tag{1}$$

where μ^* denotes the optimal policy, i.e. the quantal response equilibrium policy of player 1 and $Q^{*,1}$ the corresponding Q-function for player 1. Several remarks are in order to clarify this equation. First, the optimal strategy depends not only on the Q-function but also on the strategy of the opposing agent. This can be interpreted as fixing one agent and considering the induced MDP (see, for example, Definition 4.1 in Fu et al. [2021]).

From now on we will state everything from the perspective of player one. The results for player 2 follow analogously. Therefore, we will also omit the index of the reward and value functions. Next, using the definition of the Q-function and the value function, we can rewrite (1) to get an explicit formulation of the *average* reward that agent 1 receives for playing a specific action $a \in \mathcal{A}$:

$$\sum_{b' \in \mathcal{B}} \nu^*(b' \mid s) R(s,a,b') = \lambda \log\left(\mu^*(a \mid s)\right) + V_\lambda^*(s) - \gamma \sum_{s'} \sum_{b' \in \mathcal{B}} \nu^*(b' \mid s) P(s' \mid s,a,b') V_\lambda^*(s'). \tag{2}$$

Regarding the feasible reward set, our focus is to find an explicit reward formulation to characterize the reward functions in the feasible reward set. Rewriting equation (2) in terms of a specific reward, we get the following characterization.

Lemma 3.6. *Let (μ^*, ν^*) be two equilibrium policies for the 2 Person λ Entropy-Regularized Markov Game $\mathcal{G}$. Then for the MAIRL problem $(\mathcal{G}, (\mu^*, \nu^*))$ a reward R is feasible if and only if there exists a function $V \in \mathbb{R}^{\mathcal{S}}$ and $|\mathcal{B}| - 1$ functions $R : \mathbb{R}^{\mathcal{S}\times\mathcal{A}\times\mathcal{B}}$, such that for all $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$*

$$R(s,a,b) = \frac{1}{\nu^*(b|s)} \left(\lambda \log(\mu^*(a \mid s)) + V(s) - \gamma \sum_{s'} \sum_{b' \in \mathcal{B}} \nu^*(b' \mid s) P(s' \mid s,a,b') V(s') - \sum_{b' \neq b} \nu^*(b' \mid s) R(s,a,b') \right).$$

As in practice, we do not have access to P, ν^* and μ^*, our goal is to analyze the impact of estimating the transition probability and the expert's policy, and how this affects the existence of a recovered feasible reward. Therefore, we now want to analyze how a recovered MAIRL problem $(\hat{\mathcal{G}}, (\hat{\mu^*}, \hat{\nu^*}))$ translates to the original MAIRL problem. Care is required for this analysis, as we must consider the estimation of the expert itself, the induced transition model, which incorporates the estimation of the other expert's policy and their deviations for alternative actions.

Theorem 3.7 (Error propagation)**.** *Let the MAIRL problem be given by $(\mathcal{G}, (\mu^*, \nu^*))$ and another MAIRL problem by $(\hat{\mathcal{G}}, (\hat{\mu}^*, \hat{\nu}^*))$. Then, we have that*

$$|R(s,a,b) - \hat{R}(s,a,b)| \leq \frac{1}{\nu^*(b \mid s)\hat{\nu}^*(b \mid s)} \Bigg(\lambda |\log \mu^*(a \mid s) - \log \hat{\mu}^*(a \mid s)| + \gamma \max_b |\sum_{s'} V(s') P(s' \mid s,a,b) - \hat{P}(s' \mid s,a,b)| + R_{\max} \mathrm{TV}(\nu, \hat{\nu}) \Bigg).$$

The introduced theorem highlights the additional complexity of the problem. Specifically, it reveals two key challenges that arise in error propagation for multi-agent IRL. First, the maximum of the joint transition probabilities plays a critical role, amplifying the sensitivity of the system to inaccuracies in transition estimation. Second, any deviation in estimating the other expert's policy, quantified by the total variation distance, directly contributes to errors in the recovered reward.

Recovering Feasible Rewards. The previous section revealed, that also in the Multi-agent case, the explicit feasible reward can be decomposed into parts that depend on the policy of both agents and the transition model. Therefore, we will now analyze the amount of samples required to obtain a meaningful reward function. Let us first introduce the assumption, also common in single-agent IRL, that the lowest probability of an action taken from the experts is bounded away from zero by some constant (see e.g. Assumption D.1. in Metelli et al. [2023]).

Assumption 3.8. Let μ^*, ν^* be the QRE equilibrium expert policies. Then we assume that

$$\min_{a \in \mathcal{A}, b \in \mathcal{B}} (\mu^*(a \mid s), \nu^*(b \mid s)) \geq \Delta_{\min} \, \forall s \in \mathcal{S}.$$

For both estimation tasks, the expert policies and the transition probability, we employ empirical estimators. For each iteration $k \in [K]$, let $n_k(s,a,b,s') = \sum_{t=1}^k \mathbf{1}_{(s_t,a_t,b_t,s'_t)=(s,a,b,s')}$ denote the count of visits to the triplet $(s,a,b,s') \in \mathcal{S} \times (\mathcal{A} \times \mathcal{B}) \times \mathcal{S}$, and let $n_k(s,a,b) = \sum_{s' \in \mathcal{S}} n_k(s,a,b,s')$ denote the count of visits to the state-action pair (s,a). Additionally, we introduce $n_k(s,a) = \sum_{t=1}^k \mathbf{1}_{(s_t,a_t)=(s,a)}$ and $n_k(s,b) = \sum_{t=1}^k \mathbf{1}_{(s_t,b_t)=(s,b)}$ as the count of times action a and respectively b was sampled in state $s \in \mathcal{S}$ for each agent i, and $n_k(s) = \sum_{a \in \mathcal{A}} n_k(s,a)$ as the count of visits to state s for any agent.

It is important to note the distinction here: the count of actions must be done separately for each agent, whereas the count of state visits needs to be done for both of the agents.

In the following theorem we will assume to have access to a generative model, an assumption that is common in initial theoretical works on IRL [Metelli et al., 2021, Lindner et al., 2022, Metelli et al., 2023]. We will discuss potential directions that loosen this assumption in Appendix C.

Theorem 3.9. *Let Assumption 3.8 hold true. Then, allocating the samples uniformly over $\mathcal{S} \times \mathcal{A} \times \mathcal{B}$ and using the empirical estimators introduced in Eq.* (12) *and Eq.* (13)*, we can stop the sampling procedure with a probability of at least $1 - \delta$ after iteration τ and satisfy the optimality criterion Definition F.2, where the sample complexity is of order $\tilde{\mathcal{O}}\left(\frac{\gamma^2 R_{\max}^2 |\mathcal{S}||\mathcal{A}||\mathcal{B}|}{(1-\gamma)^4 \epsilon^2 \Delta_{\min}^4}\right)$.*

Some remarks are in order for this complexity bound. We observe that the sample complexity bound depends on the product of the action space of both players. Translating this to the n-player setting would result in an exponential dependency in the number of players. Although this might seem unfavorable, it is generally known that learning an NE in the worst case has an exponential bound. Zhang et al. [2023] show that even in model-based two player zero-sum games with access to a generative model, where the reward knowledge can not be used during learning the sample complexity depends on $|\mathcal{A}||\mathcal{B}|$. Therefore, the derived bound for the MAIRL setting aligns with the bounds derived for learning NE in the MARL setting. Additionally, we can see that the sample complexity bound is related to estimating the expert policies. This requires to estimate the log probabilities as well as the inverse probabilities of specific action taken by one player. To do so, we need the assumptions that the probabilities are bounded away from zero (Assumption 3.8).

4 Identifiability in Multi-agent Games?

At the beginning of Section 3.2, we noted that entropy regularization has been introduced in the single-agent setting to derive conditions to identify the reward (up to constants). Unlike the single-agent case, multi-agent systems introduce additional challenges due to the interplay between agents' strategies and the underlying reward structure. These challenges make identifiability like in single-agent settings not possible, unless further assumptions on the underlying Markov Game are posed. We split this section into two parts. First, we consider average reward characterization. Then, we introduce linear separable Markov Games and investigate identification in this setting.

Average Reward identification. First, let us revisit the derivations from the last section. In particular, note that the left-hand side of Eq. (2) shows, that agent'1 average reward depends on agent 2's policy $\nu(b' \mid s)$ which averages over agent 2's action. Additionally, note that the reward function is in the multi-agent setting has dimension $|\mathcal{S}||\mathcal{A}||\mathcal{B}|$, while Eq. (1) only gives condition for $|\mathcal{S}||\mathcal{A}|$ and $|\mathcal{S}||\mathcal{B}|$ respectively. This shows immediately that the resulting system of equations is under-determined.

However, the left hand side can be interpreted in terms of the induced MDP. In particular we derive an explicit reward function for the average reward $R^{\nu}(s, a) = \sum_{b' \in \mathcal{B}} \nu(b' \mid s) R(s, a, b')$. Next, let us additionally define the induced transition function, keeping the strategy of agent 2 fixed. Then, we have $P^{\nu}(\cdot \mid s, a) := \sum_{b' \in \mathcal{B}} \nu(b' \mid s) P(s' \mid s, a, b')$. Thus, agent 1's decision problem, when facing a fixed opponent policy ν, is equivalent to a single-agent MDP with the average reward function R^{ν} and the transition function P^{ν}.

These findings indicate that single-agent IRL theory [Cao et al., 2021, Rolland et al., 2022] can be applied. However, note that for identifiability in single-agent settings at least two environments with different transition dynamics and discount factors that induce experts with the same reward functions are necessary. In multi-agent systems, one obtains a different transition model by varying the policy of the opponent and observing the best responds to this change of dynamics. This can be less restrictive than requiring a new environment as in the single-agent case. Next, we state the result for the average reward case for multi-agents, this is similar to Theorem 3 by [Rolland et al., 2022].

Theorem 4.1. *Let a Markov Game be given with two different opponents ν_1, ν_2 that induce different dynamics P^{ν_1}, P^{ν_2} and discount factors γ_1, γ_2. Suppose that in both Games we observe QRE equilibrium policy pairs (μ_1, ν_1) and a different ν_2 with a best responding policy μ_2 such that they have same average reward functions $R^{\nu_1} = R^{\nu_2}$. Additionally, define $P_a^{\nu_i} \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}}$ the induced transition matrix of expert $i \in \{1, 2\}$. Then, the* average *reward player 1 receives can be recovered up to a constant if and only if*

$$\mathit{rank}\begin{pmatrix} I - \gamma_1 P_{a_1}^{\nu_1} & I - \gamma_2 P_{a_1}^{\nu_2} \\ \vdots & \vdots \\ I - \gamma_1 P_{a_{|\mathcal{A}|}}^{\nu_1} & I - \gamma_2 P_{a_{|\mathcal{A}|}}^{\nu_2} \end{pmatrix} = 2|\mathcal{S}| - 1. \tag{3}$$

Analogously this holds for player 2.

Therefore, for the average case this closely resembles the single-agent case. However, if we want to estimate the underlying problem, which is the more realistic setting, things change. In particular the error in the estimated induced transition $\|P^{\nu} - P^{\tilde{\nu}}\|$ is bounded by terms dependent on the policy estimation error and the underlying transition model error. Using the $L1$ norm we receive for a given $(s,a) : \|P^{\nu}(\cdot \mid s,a) - P^{\tilde{\nu}}(\cdot \mid s,a)\| \leq \|\nu(\cdot \mid s) - \hat{\nu}(\cdot \mid s)\|_1 + \max_{b'} \|P(s,a,b') - \hat{P}(\cdot \mid s,a,b')\|_1$. We can derive the following sample complexity for this.

Theorem 4.2 (Sample Complexity for Induced Transitions)**.** *To estimate the induced transition model P^{ν} for Player 1 such that the maximum L_1 error over all (s,a) rows is bounded by ϵ with probability at least $1-\delta$, the total number of samples N_{total} is in the order of $\mathcal{O}\left(\frac{|\mathcal{S}||\mathcal{A}||\mathcal{B}|}{\epsilon^2}\right)$.*

Reward identification in linearly separable Markov Games. To get an identifiable reward not only in the average sense one needs to disentangle the reward from player 2's strategy. This shows that the multi-agent case is fundamentally harder than then the single-agent case. Intuitively, the reason is that the definition of the NE only ensures optimality against a fixed opponent strategy. Additionally, note that the reward function is a matrix of size $|\mathcal{S}||\mathcal{A}||\mathcal{B}|$ while the optimal policy is only defined on $|\mathcal{S}||\mathcal{A}|$ and $|\mathcal{S}||\mathcal{B}|$ respectively.

A potential assumption that disentangles the joint dependency of the reward is fulfilled in *linearly separable reward Markov Games*, first introduced in the seminal work of Parthasarathy et al. [1984]. This framework has also been leveraged in more recent studies (e.g., Pérolat et al. [2021]). A reward function is said to be *linearly separable* if it can be decomposed into two independent terms, each depending solely on one player's action $R^1(s,a,b) = R^1_A(s,a) + R^1_B(s,b)$. This formulation immediately mitigates the complexity introduced by the product space dependency, as the reward function now operates over individual action spaces instead. Now, we can rewrite the condition on a reward for a specific state action pair (s,a). We give the formulation directly for two observed environments, meaning that for every $(s,a) \in \mathcal{S} \times \mathcal{A}$

$$\begin{aligned} R_A(s,a) &= \lambda \log(\mu_1^*(a \mid s)) + V_1(s) - \gamma \sum_{s'} P^{\nu_1^*}(s' \mid, s,a) V_1(s') - \sum_{b \in \mathcal{B}} \nu_1^*(b \mid s) R_B(s,b) \\ &= \lambda \log(\mu_2^*(a \mid s)) + V_2(s) - \gamma \sum_{s'} P^{\nu_2^*}(s' \mid, s,a) V_2(s') - \sum_{b \in \mathcal{B}} \nu_2^*(b \mid s) R_B(s,b) \end{aligned} \tag{4}$$

We can observe the following two cases. If we have two environments for which the reward $R_A(s,a)$ is the same, and player 1 faces the same opponent strategy ν^*, but different transition dynamics P_1 and P_2, then we notice that $\sum_{b \in \mathcal{B}} \nu^*(b \mid s) R_B(s,b)$ cancels out. Instead if we have different opponent policies $\nu_1^* \neq \nu_2^*$, then we get a new rank requirement for the resulting system of equation. We give a detailed discussion in Section 4 and summarize the findings in the following proposition.

Proposition 4.3 (Identifiability with Linearly Separable Rewards)**.** *Let two Markov Games with linearly separable rewards be given. Then, the resulting system of equations from Eq.* (4) *is solvable, i.e. the reward is identifiable up to a constant, if the matrix has rank* $2|\mathcal{S}| - 1$ *and* $\nu_1^* = \nu_2^*$ *or rank* $2|\mathcal{S}|(|\mathcal{B}|+1) - 1$ *in case* $\nu_1^* \neq \nu_2^*$ *and for one action* $b_0 \in \mathcal{B}$ *we have* $R_B(s,b_0) = 0$.

Summarized, we have showed that identifiability is not possible in general in the multi-agent setting for a particular $R(s,a,b)$. Additionally, we have given scenarios and conditions that make identifiability possible.

5 Conclusion and Future Work

We formalized Multi-Agent Inverse Reinforcement Learning (MAIRL), highlighting its unique challenges over single-agent IRL, particularly the insufficiency of single equilibrium observations for meaningful reward construction due to equilibrium multiplicity and selection ambiguity. To address this, we focused on regularized Markov Games to ensure equilibrium uniqueness. In this setting, we extended single-agent IRL bounds Metelli et al. [2021] to analyze error propagation from these estimations to the recovered reward set, resulting in a sample complexity bound for the Uniform Sampling algorithm for Quantal Response Equilibria. Additionally, we addressed the question of reward identifiability in multi-agent systems. This work gives theoretical foundation of MAIRL and opens many potential directions for future work. We outline a few of them in Appendix C.

References

P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 1, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385. doi: 10.1145/1015330.1015430. URL `https://doi.org/10.1145/1015330.1015430`.

D. Amodei, C. Olah, J. Steinhardt, P. F. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. *CoRR*, abs/1606.06565, 2016. URL `http://arxiv.org/abs/1606.06565`.

B. Baker, I. Kanitscheider, T. M. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch. Emergent tool use from multi-agent autocurricula. *CoRR*, abs/1909.07528, 2019. URL `http://arxiv.org/abs/1909.07528`.

D. Berend and A. Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18, 10 2012. doi: 10.1214/ECP.v18-2359.

H. Cao, S. N. Cohen, and L. Szpruch. Identifiability in inverse reinforcement learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Q. Cui and S. S. Du. When are offline two-player zero-sum markov games solvable? In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25779–25791. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/a57483b394a3654f4317051e4ce3b2b8-Paper-Conference.pdf`.

M. Dawood, S. Pan, N. Dengler, S. Zhou, A. P. Schoellig, and M. Bennewitz. Safe multi-agent reinforcement learning for formation control without individual reference targets, 2023. URL `https://arxiv.org/abs/2312.12861`.

A. M. Fink. Equilibrium in a stochastic n-person game. *Journal of Science of the Hiroshima University*, 28:89–93, 1964. URL `https://api.semanticscholar.org/CorpusID:120600263`.

J. Fu, A. Tacchetti, J. Perolat, and Y. Bachrach. Evaluating strategic structures in multi-agent inverse reinforcement learning. *J. Artif. Int. Res.*, 71:925–951, Sept. 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12594. URL `https://doi.org/10.1613/jair.1.12594`.

D. Goktas, A. Greenwald, S. Zhao, A. Koppel, and S. Ganesh. Efficient inverse multiagent learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=JzvIWvC9MG`.

D. Hadfield-Menell, S. Milli, P. Abbeel, S. Russell, and A. D. Dragan. Inverse reward design. *CoRR*, abs/1711.02827, 2017. URL `http://arxiv.org/abs/1711.02827`.

J. Hu and M. P. Wellman. Nash q-learning for general-sum stochastic games. *J. Mach. Learn. Res.*, 4 (null):1039–1069, dec 2003. ISSN 1532-4435.

J. Jin, C. Song, H. Li, K. Gai, J. Wang, and W. Zhang. Real-time bidding with multi-agent reinforcement learning in display advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18. ACM, Oct. 2018. doi: 10.1145/3269206.3272021. URL `http://dx.doi.org/10.1145/3269206.3272021`.

S. Kalyanaraman and C. Umans. The complexity of rationalizing matchings. In *Proceedings of the 19th International Symposium on Algorithms and Computation*, ISAAC '08, page 171–182, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 9783540921813. doi: 10.1007/978-3-540-92182-0_18. URL `https://doi.org/10.1007/978-3-540-92182-0_18`.

S. Kalyanaraman and C. Umans. The complexity of rationalizing network formation. In *2009 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 485–494, 2009. doi: 10.1109/FOCS.2009.48.

E. Kaufmann, P. Ménard, O. D. Domingues, A. Jonsson, E. Leurent, and M. Valko. Adaptive reward-free exploration. In *Algorithmic Learning Theory*, pages 865–891. PMLR, 2021.

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`.

V. Kuleshov and O. Schrijvers. Inverse game theory: Learning utilities in succinct games. In *Workshop on Internet and Network Economics*, 2015. URL `https://api.semanticscholar.org/CorpusID:2405324`.

F. Lazzati, M. Mutti, and A. M. Metelli. How does inverse rl scale to large state spaces? a provably efficient approach. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 54820–54871. Curran Associates, Inc., 2024a. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/62a9c80248963f348778a9c0bec060dd-Paper-Conference.pdf`.

F. Lazzati, M. Mutti, and A. M. Metelli. Offline inverse rl: new solution concepts and provably efficient algorithms. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024b.

S. Leonardos, G. Piliouras, and K. Spendlove. Exploration-exploitation in multi-agent competition: Convergence with bounded rationality. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26318–26331. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/dd1970fb03877a235d530476eb727dab-Paper.pdf`.

S. Levine, Z. Popovic, and V. Koltun. Nonlinear inverse reinforcement learning with gaussian processes. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL `https://proceedings.neurips.cc/paper-files/paper/2011/file/c51ce410c124a10e0db5e4b97fc2af39-Paper.pdf`.

X. Lin, P. A. Beling, and R. Cogill. Multi-agent inverse reinforcement learning for zero-sum games. *CoRR*, abs/1403.6508, 2014. URL `http://arxiv.org/abs/1403.6508`.

X. Lin, S. C. Adams, and P. A. Beling. Multi-agent inverse reinforcement learning for general-sum stochastic games. *CoRR*, abs/1806.09795, 2018. URL `http://arxiv.org/abs/1806.09795`.

D. Lindner, A. Krause, and G. Ramponi. Active exploration for inverse reinforcement learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

A. M. Metelli, G. Ramponi, A. Concetti, and M. Restelli. Provably efficient learning of transferable rewards. 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 7665–7676. PMLR, 7 2021. URL `https://proceedings.mlr.press/v139/metelli21a.html`.

A. M. Metelli, F. Lazzati, and M. Restelli. Towards theoretical understanding of inverse reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

S. Natarajan, G. Kunapuli, K. Judah, P. Tadepalli, K. Kersting, and J. W. Shavlik. Multi-agent inverse reinforcement learning. *2010 Ninth International Conference on Machine Learning and Applications*, pages 395–400, 2010. URL `https://api.semanticscholar.org/CorpusID:3440496`.

A. Y. Ng and S. J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 663–670, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

T. Parthasarathy, S. Tijs, and O. J. Vrieze. Stochastic games with state independent transitions and separable rewards. *Lecture Notes in Economics and Mathematical Systems*, pages 262–271, 1984. URL `https://api.semanticscholar.org/CorpusID:122647253`.

J. Pérolat, S. Perrin, R. Elie, M. Laurière, G. Piliouras, M. Geist, K. Tuyls, and O. Pietquin. Scaling up mean field games with online mirror descent. *CoRR*, abs/2103.00623, 2021. URL `https://arxiv.org/abs/2103.00623`.

D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 2586–2591, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

G. Ramponi, P. Kolev, O. Pietquin, N. He, M. Laurière, and M. Geist. On imitation in mean-field games. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

N. Ratliff, D. Bradley, J. A. Bagnell, and J. Chestnutt. Boosting structured prediction for imitation learning. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'06, page 1153–1160, Cambridge, MA, USA, 2006a. MIT Press.

N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich. Maximum margin planning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 729–736, New York, NY, USA, 2006b. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143936. URL `https://doi.org/10.1145/1143844.1143936`.

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.

M. Samvelyan, T. Rashid, C. Schroeder de Witt, G. Farquhar, N. Nardelli, T. G. J. Rudner, C.-M. Hung, P. H. S. Torr, J. Foerster, and S. Whiteson. The starcraft multi-agent challenge. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 2186–2188, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450363099.

S. Shalev-Shwartz, S. Shammah, and A. Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *CoRR*, abs/1610.03295, 2016. URL `http://arxiv.org/abs/1610.03295`.

L. S. Shapley. Stochastic games*. *Proceedings of the National Academy of Sciences*, 39:1095 – 1100, 1953. URL `https://api.semanticscholar.org/CorpusID:263414073`.

J. Song, H. Ren, D. Sadigh, and S. Ermon. Multi-agent generative adversarial imitation learning. *CoRR*, abs/1807.09936, 2018. URL `http://arxiv.org/abs/1807.09936`.

M. Takahashi. Equilibrium points of stochastic non-cooperative \$n\$-person games. *Journal of Science of the Hiroshima University*, 28:95–99, 1964. URL `https://api.semanticscholar.org/CorpusID:118906641`.

J. Tang, G. Swamy, F. Fang, and Z. S. Wu. Multi-agent imitation learning: Value is easy, regret is hard. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 27790–27816. Curran Associates, Inc., 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/3103b25853719847502559bf67eb4037-Paper-Conference.pdf`.

X. Wang, L. Ke, Z. Qiao, and X. Chai. Large-scale traffic signal control using a novel multi-agent reinforcement learning. *CoRR*, abs/1908.03761, 2019. URL `http://arxiv.org/abs/1908.03761`.

J. Wu, W. Shen, F. Fang, and H. Xu. Inverse game theory for stackelberg games: the blessing of bounded rationality. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 32186–32198. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/cfce833814505906445f8df2f65ab548-Paper-Conference.pdf`.

L. Yu, J. Song, and S. Ermon. Multi-agent adversarial inverse reinforcement learning. In *International Conference on Machine Learning*, pages 7194–7201. PMLR, 2019.

A. Zanette, M. J. Kochenderfer, and E. Brunskill. Almost horizon-free structure-aware best policy identification with a generative model. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/a724b9124acc7b5058ed75a31a9c2919-Paper.pdf.

K. Zhang, S. M. Kakade, T. Başar, and L. F. Yang. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *J. Mach. Learn. Res.*, 24(1), Jan. 2023. ISSN 1532-4435.

L. Zhao, M. Wang, and Y. Bai. Is inverse reinforcement learning harder than standard reinforcement learning? a theoretical perspective. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

H. Zhong, W. Xiong, J. Tan, L. Wang, T. Zhang, Z. Wang, and Z. Yang. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27117–27142. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhong22b.html.

M. Zhou, J. Luo, J. Villela, Y. Yang, D. Rusu, J. Miao, W. Zhang, M. Alban, I. Fadakar, Z. Chen, A. C. Huang, Y. Wen, K. Hassanzadeh, D. Graves, D. Chen, Z. Zhu, N. M. Nguyen, M. Elsayed, K. Shao, S. Ahilan, B. Zhang, J. Wu, Z. Fu, K. Rezaee, P. Yadmellat, M. Rohani, N. P. Nieves, Y. Ni, S. Banijamali, A. I. Cowen-Rivers, Z. Tian, D. Palenicek, H. Bou-Ammar, H. Zhang, W. Liu, J. Hao, and J. Wang. SMARTS: scalable multi-agent reinforcement learning training school for autonomous driving. *CoRR*, abs/2010.09776, 2020. URL https://arxiv.org/abs/2010.09776.

B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, page 1433–1438. AAAI Press, 2008. ISBN 9781577353683.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims stated in the introduction are answered in Proposition 3.4, Theorem 3.9 and in Theorem 4.1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations about the assumption of a generative model are discussed and justified above and also outlined in Appendix C.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Each Theorem states the assumptions clearly and contains a proof in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper is mainly theoretical. However, simple empirical validations are given and the details for these examples can be found in Appendix H.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The main contributions of this paper are theoretical and we only provide numerical verifications of the theory. However, details on the empirical validations are given in Appendix H. The full code will be released after publication.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper does not have a part that requires training. However, details on how to reproduce the plots are given in Appendix H.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All plots contain error bars where applicable Fig. 2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: No specific requirements are needed regarding compute resources as the experiments are only on a small scale.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper is conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The main contributions of this work are theoretical. However, reward learning in multi-agents can have a societal impact as for example recovering a reward that is not exploitable, see Fig. 2 and the introduction for a discussion.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Contents of Appendix

This appendix provides supplementary material to support the main findings of the paper. First, we give an overview of the used notations, including some additional notations needed for the proofs in the appendix. Then, we present the omitted analysis for the feasible reward set under a single equilibrium observation in Appendix D. We then present the complete proofs for key results for the regularized Markov Game setting in Appendix F. Afterwards, in Appendix G we give the missing proofs for the identifiability section. Then, we give the details for our presented numerical validations and some additional experiments that show that MAIRL can be superior to BC. Finally, the appendix compiles a list of technical results, along with their proofs, that are referenced throughout this work. For a better overview we provide a table of contents.

A Notation and Symbols

In this part of the appendix, we include notation used in the main paper and some additional notation used for the proofs in the appendix.

Notation	Description
$\mathcal{X}$	Finite set
$\mathbb{R}^{\mathcal{X}}$	Set of all functions mapping from $\mathcal{X}$ to $\mathbb{R}$
$\Delta^{\mathcal{X}}$	Set of probability measures over $\mathcal{X}$
$[n]$	Set $\{1, ..., n\}$
$(\mathcal{X}, d)$	(Pre)metric space
$\mathcal{H}_d$	Hausdorff (pre)metric
$\mathcal{G}$	Markov Game without reward function $(n, \mathcal{S}, \mathcal{A}, P, \gamma, \rho)$
n	Number of players
$\mathcal{S}$	Finite state space
$\mathcal{A}^i$	Action space for player i
$\mathcal{A}$	Joint action space $\mathcal{A}^1 \times ... \times \mathcal{A}^n$
$P(s'\|s, a)$	Transition model (probability of next state s' given state s and joint action a)
γ	Discount factor
ρ	Initial state distribution
π^i	Policy (strategy) for player $i : \mathcal{S} \to \Delta^{\mathcal{A}^i}$
$\boldsymbol{\pi}$	Joint policy $(\pi^1, ..., \pi^n) = (\pi^i, \pi^{-i})$
$\boldsymbol{\pi}(\mathbf{a}\|s)$	Probability of joint action a under joint policy π in state s, $\prod_{j=1}^n \pi^j(a^j\|s)$
Π^i	Set of all policies for agent i
Π	Set of all joint policies
$\bar{w}_{s,a}^{\boldsymbol{\pi},\rho}$	Discounted probability of visiting state-action pair (s, a) starting from ρ under $\boldsymbol{\pi}$
$R^i(s, a)$	Reward function for agent $i : \mathcal{S} \times \mathcal{A} \to [-R^i_{max}, R^i_{\max}]$
R	Joint reward function $(R^1, ..., R^n)$
$R_{\max}$	Maximum absolute reward across all agents, $\max_{i\in[n]} R^i_{max}$
$\mathcal{G} \cup R$	Standard Markov Game with reward
$Q^{i,\pi}_{\mathcal{G}\cup R}(s, a)$	Q-function for agent i under policy π
$V^{i,\pi}_{\mathcal{G}\cup R}(s)$	Value function for agent i under policy π
$\boldsymbol{\pi}^{\text{Nash}}$	Nash Equilibrium policy
$\Pi^{Nash}(R)$	Set of Nash Equilibria for reward R
$V^{i,(\mu,\nu)}_{\lambda}(s)$	Value function for player i in λ_i-entropy regularized MG
$Q^{i,(\mu,\nu)}_{\lambda}(s, a, b)$	Q-function for player i in λ_i-entropy regularized MG
$(\mathcal{G}, \boldsymbol{\pi}^{\text{Nash}})$	MAIRL problem definition
$(\hat{\mathcal{G}}, \hat{\boldsymbol{\pi}}^{\text{Nash}})$	Recovered MAIRL problem
$\mathcal{R}_{(\mathcal{G},\pi^{Nash})}$	Feasible reward set for a single observed NE π^{Nash}
$\hat{\pi}$	Policy recovered from MAIRL problem
$\mathcal{E}(\hat{\pi})$	Nash Imitation Gap for MAIRL
$\mathcal{R}_{(\mathcal{G},\Pi^{Nash})}$	Feasible reward set for the set of all NE Π^{Nash}
μ^*, ν^*	QRE equilibrium policies
$R^\nu(s, a)$	Average reward for player 1 when player 2 plays ν: $\sum_{b'\in\mathcal{B}} \nu(b'\|s)R(s, a, b')$
$P^\nu(s'\|s, a)$	Induced transition for player 1 when player 2 plays ν: $\sum_{b'\in\mathcal{B}} \nu(b'\|s)P(s'\|s, a, b')$
$R^1_A(s, a) + R^1_B(s, b)$	Linearly separable reward for player 1
Δ_{min}	Minimum probability bound for QRE policies
$N_k(s, a, b, s')$	Count of visits to (s, a, b, s') up to iteration k
$N_k(s, a, b)$	Count of visits to (s, a, b) up to iteration k
$N_k(s, a)$	Count of player 1 taking action a in state s up to iteration k
$N_k(s, b)$	Count of player 2 taking action b in state s up to iteration k
$N_k(s)$	Count of visits to state s up to iteration k
$\hat{P}_k(s'\|s, a, b)$	Empirical estimate of transition probability at iteration k
$\hat{\mu}_k(a\|s)$	Empirical estimate of policy μ at iteration k
$\hat{\nu}_k(b\|s)$	Empirical estimate of policy ν at iteration k

In this section, we introduce the additional notation needed for the matrix expression of the Q-function, the value function, and in particular, for an additional implicit condition for the feasible reward function (Theorem D.1) similar to the one derived in Lin et al. [2018]. To achieve this we use a similar notation from Lin et al. [2018], adjusted to this work. First, we introduce for every agent $i \in [n]$ the stacked reward $\boldsymbol{R}^i$. For every state $s \in \mathcal{S}$ the reward can be seen as a matrix of dimension $|\mathcal{A}^i| \times |\prod_{j\neq i}^n |\mathcal{A}^j|$. Doing this for every state and stacking them, results in a vector $\boldsymbol{R}^i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $|\mathcal{A}|$ is the dimension of the joint action space. We additionally introduce the operator $\boldsymbol{\pi}$, which can be written as a $|\mathcal{S}| \times |\mathcal{S}||\mathcal{A}|$ matrix, structured in the following way. First, we need to fix an arbitrary order on the joint action space $[|\mathcal{A}|]$ in the same way as already done for stacking the Reward for every agent. Given the order, we have that for $k \in [|S|]$, the k-th row is given by

$$\Phi_1^\pi(k), \dots, \Phi_{|\mathcal{A}|}^\pi(k),$$

where for $j \in [|\mathcal{A}|]$ we have

$$\Phi_j^\pi(k) = \left[\underbrace{0, \dots, 0}_{k-1}, \prod_{i=1}^n \pi^i(a_j^i \mid k), \underbrace{0, \dots, 0}_{|\mathcal{S}|-k}\right].$$

Therefore, the resulting matrix has in its first $|\mathcal{S}|$ columns a diagonal matrix of size $|\mathcal{S}| \times |\mathcal{S}|$ with the corresponding probabilities of playing the first joint action in all possible states.

$$\begin{pmatrix} \prod_{i=1}^n \pi^i(a_1 \mid 1) & 0 & 0 & \cdots & 0 & \cdots \\ 0 & \prod_{i=1}^n \pi^i(a_1 \mid 2) & 0 & \cdots & 0 & \cdots \\ 0 & 0 & \prod_{i=1}^n \pi^i(a_1 \mid 3) & \cdots & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \cdots \\ 0 & 0 & 0 & \cdots & \prod_{i=1}^n \pi^i(a_1 \mid S) & \cdots \end{pmatrix}$$

The transition matrix $\boldsymbol{P}$ of a Markov Game also depends on the joint actions, making the resulting transition matrix of dimension $|\mathcal{S}||\mathcal{A}| \times \mathcal{S}$. This allows us to write the value function as a column vector of dimension $\mathbb{R}^{|\mathcal{S}|}$ and the Q-value function as a vector, identically as the reward vector, of dimension $|\mathcal{S}||\mathcal{A}| \times 1$. Therefore, we can write:

$$\boldsymbol{Q}^{i,\boldsymbol{\pi}} = \boldsymbol{R}^i + \gamma \boldsymbol{P} \boldsymbol{V}^{i,\boldsymbol{\pi}}, \quad \boldsymbol{V}^{i,\boldsymbol{\pi}} = \boldsymbol{\pi} \boldsymbol{Q}^{i,\boldsymbol{\pi}}.$$

B Related Work

This work intersects with several fields of research, particularly **Inverse Reinforcement Learning**, **Multi-Agent Inverse Reinforcement Learning** , and **Inverse (Algorithmic) Game Theory**.

Theoretical Understanding of IRL. IRL was first introduced by Ng and Russell [2000], emphasizing its ill-posed nature. Subsequent work tackled ambiguity via reformulations [Abbeel and Ng, 2004, Ziebart et al., 2008, Ramachandran and Amir, 2007, Ratliff et al., 2006a, Levine et al., 2011]. To avoid the ambiguity in IRL, recent research has addressed to characterize the set of feasible rewards instead of picking a single reward function. Theoretical efforts to characterize the feasible reward set were pioneered by Metelli et al. [2021]. The authors provide an explicit reward formulation, that shows that the reward depends on the expert policy and the transition dynamics. As in realistic scenarios it is not common to know the transition model and the expert policy, the authors provide an error propagation analysis on how estimation errors in these quantities transfer to the recovered reward. Additionally, they provide a uniform sampling algorithm with access to a generative model combined with a sample complexity analysis on how many samples are required to find a suitable reward function from the set of feasible rewards, that is also transferable to new environments. In Lindner et al. [2022] the authors extend this to the finite horizon setting. Additionally, the authors provide the first algorithm that removes the assumption of a generative model and instead create exploration policies to mitigate the reward uncertainty dubbed active inverse reinforcement learning. Additionally, they provide a sample complexity analysis for the most general case and a problem-dependent variant. Further insights on the theoretical insights of IRL have been provided by [Metelli et al., 2023]. The authors investigate different metrics for the IRL problem, leading to a more nuanced analysis and the requirement of refined concentration inequalities. Additionally, they provide the first lower bound for IRL, addressing an open question , that IRL is not harder than forward RL. The first offline algorithm for IRL combined with a sample complexity analysis has been provided in [Zhao et al., 2024]. The authors note limitations of so far introduced metrics in settings without a generative model. Additionally, the authors show that IRL is not harder than standard RL in the offline setting. The offline setting has also been considered in Lazzati et al. [2024b]. The authors provide a new formulation of the feasible reward set, more suitable for the offline setting. They introduce two new efficient algorithms designed for the offline setting, overcoming the new introduced challenges as the data coverage cannot be controlled anymore. An investigation how IRL translates to large state spaces has been obtained in Lazzati et al. [2024a]. The authors provide the negative result, that the feasible reward set cannot be learned efficiently in large state spaces without additional assumptions. Instead of the feasible reward set, they provide a new framework, rewards compatibility and an efficient algorithm for this setting.

While these works analyze distances to value functions and expert policies, applying them to Markov Games remains challenging due to the need for equilibrium-based solutions. This implies that the standard objectives for IRL, namely value based gaps, cannot be applied in multi-agent settings. Additionally, we show that MAIRL introduces new challenges due to the multiplicity of equilibria.

Another line of works, considers the case of reward identifiability. In Cao et al. [2021] the authors give an explicit reward characterization in the setting of regularized MDPs. In particular, the authors note that the value function can be chosen arbitrarily for a given optimal policy and therefore identifiability is not possible by a single expert policy. However, if one considers two MDPS with different transition dynamics and discount factors and the same optimal reward function, then identifiablity is possible if the MDPs are value-distinguishable. Based on these observations, Rolland et al. [2022] derived explicit conditions what value-distinguishable MDPs are. In particular, they rewrote the reward identifiability problem as system of equations and derived rank conditions that need to be fulfilled to identify the rewards up to constants. Additionally, they provide insights for the case with unknown transition functions and transferability to new environments.

In [Kim et al., 2021] the authors study the type of MDPs under which reward identifiability is possible. Considering a deterministic MDP with an entropy regularized objective, the authors provide necessary and sufficient conditions whether and MDP is identifiable. Additionally, building on these findings they provide efficient algorithms to check if an MDP is identifiable.

In this work we address the question of identifiability in the context of multi-agents. We show that it is not possible to identify the reward function up to constants without additional assumptions. Instead it is only possible to obtain an average reward function or one poses additional structural assumptions on the underlying Markov Game as e.g. linear separable rewards.

Multi-Agent Inverse Reinforcement Learning. The first extension of IRL to multi-agent settings was introduced by Natarajan et al. [2010], focusing on a centralized controller in an average reward RL framework, but without addressing competitive settings requiring game-theoretic solutions. Lin et al. [2014] extended this to Zero-Sum Markov Games, introducing a Bayesian MAIRL framework based on observed Nash equilibria, later expanded by Lin et al. [2018] to incorporate various solution concepts, though without sample complexity bounds. Yu et al. [2019] extended Maximum Entropy IRL to multi-agent settings via the logistic best response equilibrium, focusing on recovering a single reward function rather than analyzing the feasible reward set. More recently, Goktas et al. [2024] explored Inverse Multi-agent Learning with parameter-dependent payoffs, simplifying the problem by assuming access to samples from the reward function. Fu et al. [2021] approached MAIRL by decomposing it into multiple single-agent IRL tasks, applying utility-matching IRL algorithms on the induced MDPs. Additionally, this work is the first that notes that single Nash equilibrium observations can be limited. Instead of considering a single reward function, we formalize that that single equilibrium observations lead to an uninformative feasible reward set. Additionally, their approach does not address equilibrium multiplicity and lacks a sample complexity analysis. Finally, Tang et al. [2024] highlighted the inadequacy of value-based gaps in Multi-agent Imitation Learning, proposing regret as a more suitable objective when observing Correlated Equilibrium experts. In this work, we consider the Multi-Agent Inverse Reinforcement Learning framework and consider the Nash equilibrium as well as the QRE.

Inverse (Algorithmic) Game Theory. There is significant overlap between *Multi-Agent Inverse Reinforcement Learning* and *inverse algorithmic game theory*. Many works in this area apply game-theoretic solution concepts to rationalize the behavior of observed players in specific types of games [Kalyanaraman and Umans, 2008, 2009]. A related work is by Kuleshov and Schrijvers [2015], who developed polynomial-time algorithms for coarse correlated equilibria in succinct games, where the structure of the game is known and noted that in cases where the game structure is unknown, the problem is NP-hard. Their theorems indicate that without additional assumptions or more specific settings, polynomial-time algorithms cannot be expected for inversely solving Nash equilibria. In a more recent work, [Wu et al., 2022] introduce bounded rationality, i.e. considering QRE as the observed behavior, in the context of Stackelberg Games. In particular, the authors show that QRE observations are more informative than Nash equilibrium observations. This means that bounded rationality helps to be more robust against an irrational opponent. Additionally, this makes it possible to construct algorithms that have exponential dependencies on neither the number of leader actions nor the number of follower actions.

C Limitations and Future Directions

Our work establishes a first theoretical understanding of MAIRL, but several challenges remain open. In this section we outline a few of them.

Removing assumption of generative model. In the provided sample complexity analysis we have assumed having to a generative model and sampled uniformly from this. While this is a common assumption in initial works also in single-agent IRL [Metelli et al., 2021, Lindner et al., 2022, Metelli et al., 2023], this can be restrictive in general. Therefore, it could be of interest to explore the offline setting as done in [Lazzati et al., 2024b, Zhao et al., 2024] for multi-agents. However, this might not be easy as in offline learning for multi-agents different coverage assumptions are needed compared to the single-agent setting even in case of reward knowledge [Cui and Du, 2022, Zhong et al., 2022]. Another direction would be to consider the active exploration direction, meaning how to construct policies that actively the environment as done in Lindner et al. [2022] for the single-agent case.

Designing Algorithms. Our analysis showed that single Nash equilibrium observations allow for an uninformative feasible reward set. One potential algorithmic implication of this finding could be to design an algorithm that guarantees that the observed NE is the only equilibrium under this reward function. This algorithm could be in a similar spirit as the Max-Gap IRL algorithm for single-agents [Ratliff et al., 2006b], but now on an equilibrium level instead of a value-based design.

Observing multiple equilibria. As pointed out in the discussion around Definition 3.5 and in Fu et al. [2021], if one would be able to observe multiple or all equilibria from the set of Nash equilibria of the underlying game, the characterized feasible reward set is more meaningful. As this might be restrictive in practice it could give important theoretical insights.

D Omitted analysis for Section 3

The first theorem serves as an extension of the two player version theorem by Lin et al. [2018] (see section 4.6 in Lin et al. [2018]) to n-person games and general Nash equilibria. It makes use of the notation introduced in Appendix A.

Theorem D.1. *Let $\mathcal{G} \cup R$ be a n-person general-sum Markov Game. A policy $\boldsymbol{\pi}$ is an NE strategy if and only if*

$$(\boldsymbol{\pi}^{\mathrm{Nash}} - \tilde{\boldsymbol{\pi}})(I - \gamma \boldsymbol{P} \boldsymbol{\pi}^{\mathrm{Nash}})^{-1} \boldsymbol{R}^i \geq 0.$$

with the meaning that without (s, a) symbols a matrix notation and $\tilde{\boldsymbol{\pi}}$ is the policy with $\pi^{-i} = \pi^{-i,\mathrm{Nash}}$ and π^i plays action a with probability 1.

Proof. In the first step of the proof we state the theorem for the case where $n = 2$ with the use of the definition of an NE. We only write the condition for agent 1 to understand the structure. For every action $a^1 \in \mathcal{A}^1$ and every state $s \in \mathcal{S}$ it must hold true that:

$$\sum_{a^2 \in \mathcal{A}^2} \pi^{2,\mathrm{Nash}}(a^2 \mid s) R^1(s, a^1 a^2) + \gamma \sum_{a^2 \in \mathcal{A}^2} \pi^{2,\mathrm{Nash}}(a^2 \mid s) \sum_{s'} P(s' \mid s, a^1 a^2) V^{\boldsymbol{\pi}^{\mathrm{Nash}}}(s) \leq V^{\boldsymbol{\pi}^{\mathrm{Nash}}}(s)$$

If we now want to generalize this to a n-person Markov Game, we get that for every player $i \in [n]$, every action a^i and every state $s \in \mathcal{S}$ it must hold true that:

$$\begin{aligned} &\sum_{a^{-i} \in \mathcal{A}^{-i}} \pi^{-i,\mathrm{Nash}}(a^{-i} \mid s) R^1(s, a^i a^{-i}) \\ &\quad + \gamma \sum_{a^{-i} \in \mathcal{A}^{-i}} \pi^{-i,\mathrm{Nash}}(a^{-i} \mid s) \sum_{s'} P(s' \mid s, a^i a^{-i}) V^{\boldsymbol{\pi}^{\mathrm{Nash}}}(s) \leq V^{\boldsymbol{\pi}^{\mathrm{Nash}}}(s) \end{aligned}$$

We can rewrite this equation in terms of the Q-function and get

$$\sum_{a^{-i} \in \mathcal{A}^{-i}} \pi^{-i}(a^{-i} \mid s) Q^{\boldsymbol{\pi}^{\mathrm{Nash}}}(s, \mathbf{a}) \leq V^{\boldsymbol{\pi}^{\mathrm{Nash}}}(s). \tag{5}$$

Now we want to rewrite the equation for all states simultaneously. Therefore we recall the notation introduced in Appendix A. We have that

$$\boldsymbol{Q}^{i,\boldsymbol{\pi}} = \boldsymbol{R}^i + \gamma \boldsymbol{P} \boldsymbol{V}^{i,\boldsymbol{\pi}}, \quad \boldsymbol{V}^{i,\boldsymbol{\pi}} = \boldsymbol{\pi} \boldsymbol{Q}^{i,\boldsymbol{\pi}}.$$

Rewriting this equation for the Nash Policy $\boldsymbol{\pi}^{\mathrm{Nash}}$ gives us

$$\boldsymbol{Q}^{i,\boldsymbol{\pi}^{\mathrm{Nash}}} = (I - \gamma \boldsymbol{P} \boldsymbol{\pi}^{\mathrm{Nash}})^{-1} \boldsymbol{R}^i.$$

Plugging in the derived equations in (5) using matrix notation for all states $s \in \mathcal{S}$ simultaneously and additionally denote the joint policy, where agent i plays action a^i with probability 1 and the other agents execute their Nash strategy $\pi^{-i,\mathrm{Nash}}$ as $\tilde{\boldsymbol{\pi}}$, we get

$$(\tilde{\boldsymbol{\pi}} - \boldsymbol{\pi}^{\mathrm{Nash}})(I - \gamma \boldsymbol{P} \boldsymbol{\pi}^{\mathrm{Nash}})^{-1} \boldsymbol{R}^i \leq 0.$$

□

The next lemma restates the condition by directly using the expectation of the advantage function with respect to the policy.

Lemma D.2 (Feasible Reward Set Implicit)**.** *A reward function $R = (R^1, \dots, R^n)$ is feasible if and only if for a Nash policy $\boldsymbol{\pi}^{\mathrm{Nash}}$, for every agent $i \in [n]$ and all $(s, a^i) \in \mathcal{S} \times \mathcal{A}^i$, it holds true that:*

$$\sum_{a^{-i} \in \mathcal{A}^{-i}} \pi^{-i,\mathrm{Nash}}(a^{-i} \mid s) A^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R}(s, a^i, a^{-i}) = 0, \textit{if } \pi^{i,\mathrm{Nash}}(a^i \mid s) > 0, a^{-i} \in \mathrm{supp}(\pi^{-i,\mathrm{Nash}}(\cdot \mid s)).$$
$$\sum_{a^{-i} \in \mathcal{A}^{-i}} \pi^{-i,\mathrm{Nash}}(a^{-i} \mid s) A^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R}(s, a^i, a^{-i}) \le 0, \textit{if } \pi^{i,\mathrm{Nash}}(a^i \mid s) = 0, a^{-i} \in \mathrm{supp}(\pi^{-i,\mathrm{Nash}}(\cdot \mid s)).$$

Proof. As we know that $a^{-i} \in \mathrm{supp}(\pi^{-i,\mathrm{Nash}}(\cdot \mid s))$ for both cases, we get for all agents $i \in [n]$ and all actions $a^{i,Nash} \in \mathcal{A}^i$ that fulfill $\pi^{i,\mathrm{Nash}}(a^{i,Nash} \mid s) > 0$, that $\sum_{a^{-i} \in \mathcal{A}^{-i}} \pi^{-i,\mathrm{Nash}}(a^{-i} \mid s) Q^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}(s, a^{i,\mathrm{Nash}} a^{-i}) > \sum_{a^{-i} \in \mathcal{A}^{-i}} \pi^{-i,\mathrm{Nash}}(a^{-i} \mid s) Q^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}(s, a^i a^{-i})$. Additionally, we have that for all $a^{i,\mathrm{Nash}}$ with $\pi^{i,\mathrm{Nash}}(a^{i,\mathrm{Nash}} \mid s) > 0$ that $\sum_{a^{-i} \in \mathcal{A}^{-i}} \pi^{-i,\mathrm{Nash}}(a^{-i} \mid s) Q^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}(s, a^{i,\mathrm{Nash}} a^{-i}) = V^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}(s)$. □

In the following we will constrain to the case of pure NE, therefore it holds true, that for some $a^{-i} \in \mathcal{A}^{-i}$, we have $\pi(a^{-i} \mid s) = 1\, \forall s \in \mathcal{S}$.

Lemma D.3. *Let $i \in [n]$ be an arbitrary agent. Then the Q-function of player i satisfies the optimality conditions of Lemma D.2 for a pure Nash equilibrium if and only if for every $(s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}$ there exists a function $A^i \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\ge 0}$ and $V^i \in \mathbb{R}^{\mathcal{S}}$ such that:*

$$Q^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R}(s, \boldsymbol{a}) = -A^i(s, \boldsymbol{a}) \mathbf{1}_{\{\pi^{i,\mathrm{Nash}}(a^i \mid s) = 0\}} \mathbf{1}_{\{\pi^{-i,\mathrm{Nash}}(a^{-i} \mid s) = 1\}} + V^i(s)$$

Proof. First we assume that the Q-function can be expressed as

$$Q^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R}(s, \mathbf{a}) = -A^i(s, \mathbf{a}) \mathbf{1}_{\{\pi^{i,\mathrm{Nash}}(a^i \mid s) = 0\}} \mathbf{1}_{\{\pi^{-i,\mathrm{Nash}}(a^{-i} \mid s) = 1\}} + V^i(s).$$

We note that

$$\begin{aligned}
V^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R}(s) &= \sum_{\mathbf{a} \in \mathcal{A}} \boldsymbol{\pi}^{\mathrm{Nash}}(\mathbf{a} \mid s) Q^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R}(s, \mathbf{a}) \\
&= \sum_{\mathbf{a} \in \mathcal{A}} \boldsymbol{\pi}^{\mathrm{Nash}}(\mathbf{a} \mid s) \left(-A^i(s, \mathbf{a}) \mathbf{1}_{\{\pi^{i,\mathrm{Nash}}(a^i \mid s) = 0\}} \mathbf{1}_{\{\pi^{-i,\mathrm{Nash}}(a^{-i} \mid s) = 1\}} + V^i(s) \right) \\
&= V^i(s),
\end{aligned}$$

where the last equality follows from the fact, that if $\boldsymbol{\pi}^{\mathrm{Nash}}(\mathbf{a} \mid s) > 0$, then $\mathbf{1}_{\{\pi^{i,\mathrm{Nash}}(a^i \mid s) = 0\}} = 0$ and vice versa. Additionally, $V^i(s)$ is independent of $\mathbf{a}$ and as the sum is over the joint action space it holds true that $\sum_{\mathbf{a} \in \mathcal{A}} \boldsymbol{\pi}^{\mathrm{Nash}}(\mathbf{a} \mid s) = 1$. We now have to consider two cases. The first one is if $\mathbf{1}_{\{\pi^{i,\mathrm{Nash}}(a^i \mid s) = 0\}} = 0$ and $\mathbf{1}_{\{\pi^{-i,\mathrm{Nash}}(a^{-i} \mid s) = 1\}} = 1$. Then it holds true that

$$\sum_{\mathbf{a} \in \mathcal{A}} \boldsymbol{\pi}^{\mathrm{Nash}}(\mathbf{a} \mid s) Q^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R}(s, \mathbf{a}) = V^i(s) = V^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R}.$$

The second case is if $\mathbf{1}_{\{\pi^{i,\mathrm{Nash}}(a^i \mid s) = 0\}} = 1$ and $\mathbf{1}_{\{\pi^{-i,\mathrm{Nash}}(a^{-i} \mid s) = 1\}} = 1$ for one action $\tilde{a}^{-i}$ with $\pi^{-i}(\tilde{a}^{-i} \mid s)$ as we assumed it is a pure NE. Then it holds true that

$$\begin{aligned}
&\sum_{\mathbf{a} \in \mathcal{A}} \pi^{-i,\mathrm{Nash}}(a^{-i} \mid s) Q^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R}(s, a^i a^{-i}) \\
&= Q^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R}(s, a^i \tilde{a}^{-i}) \\
&\quad = -A^i(s, \mathbf{a}) \mathbf{1}_{\{\pi^{i,\mathrm{Nash}}(a^i \mid s) = 0\}} \mathbf{1}_{\{\pi^{-i,\mathrm{Nash}}(a^{-i} \mid s) = 1\}} + V^i(s) \\
&\quad \le V^i(s) = V^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R},
\end{aligned}$$

where we used the fact that $-A^i(s, \mathbf{a}) \le 0$.

If we now assume that the conditions of Lemma D.2 hold, we can set for every $(s, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$ $V^i(s) = V^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R}$ and $A^i(s, \mathbf{a}) = V^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R}(s) - Q^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R}(s, \mathbf{a})$. □

Lemma D.4 (Feasible Reward Set Explicit)**.** *A reward function R is feasible if and only if, for each agent $i \in [n]$, there exist a function $A^i \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\geq 0}$ and a function $V^i \in \mathbb{R}^{\mathcal{S}}$ such that for all $(s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}$, the following holds:*

$$R^i(s, \boldsymbol{a}) = -A^i(s, \boldsymbol{a})\mathbf{1}_{\{\pi^{i,\mathrm{Nash}}(a^i|s)=0\}}\mathbf{1}_{\{\pi^{-i,\mathrm{Nash}}(a^{-i}|s)=1\}} + V^i(s) - \gamma \sum_{s'} P(s' \mid s, \boldsymbol{a})V^i(s').$$

Proof. Remembering that we can express the Q-function as $Q^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R}(s, \mathbf{a}) = R^i(s, \mathbf{a}) + \gamma \sum_{s'} P(s' \mid s, \mathbf{a})V^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R}(s')$ and applying Lemma D.3 to express the Q-function for an NE policy, we can conclude

$$\begin{aligned} R^i(s, \mathbf{a}) &= Q^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R}(s, \mathbf{a}) - \gamma \sum_{s'} P(s' \mid s, a)V^{i,\boldsymbol{\pi}^{\mathrm{Nash}}}_{\mathcal{G}\cup R}(s') \\ &= -A^i(s, \mathbf{a})\mathbf{1}_{\{\pi^{i,\mathrm{Nash}}(a^i|s)=0\}}\mathbf{1}_{\{\pi^{-i,\mathrm{Nash}}(a^{-i}|s)=1\}} + V^i(s) - \gamma \sum_{s'} P(s' \mid s, \mathbf{a})V^i(s'). \end{aligned}$$

$\square$

Theorem D.5 (Error Propagation)**.** *Let $(\mathcal{G}, \boldsymbol{\pi}^{\mathrm{Nash}})$ and $(\hat{\mathcal{G}}, \hat{\boldsymbol{\pi}}^{\mathrm{Nash}})$ be the true and the recovered MAIRL problem. Then, for every agent $i \in [n]$ and any $R_i \in \mathcal{R}_{\mathcal{B}}$ there exists $\hat{R}_i \in \mathcal{R}_{\hat{\mathcal{B}}}$ such that:*

$$|R_i(s, \boldsymbol{a}) - \hat{R}_i(s, \boldsymbol{a})| \leq A^i(s, \boldsymbol{a})|\mathbf{1}_E - \mathbf{1}_{\hat{E}}| + \gamma \sum_{s'} V^i(s')|P(s' \mid s, \boldsymbol{a}) - \hat{P}(s' \mid s, \boldsymbol{a})|,$$

where $E := \{\{\pi^{i,\mathrm{Nash}}(a^i \mid s) = 0\} \cap \{\pi^{-i,\mathrm{Nash}}(a^{-i} \mid s) > 0\}\}$ and $\hat{E} := \{\{\hat{\pi}^{i,\mathrm{Nash}}(a^i \mid s) = 0\} \cap \{\hat{\pi}^{-i,\mathrm{Nash}}(a^{-i} \mid s) = 1\}\}$.

Proof. From the explicit expression of a feasible reward D.4, we know that we can write the reward function of any agent $i \in [n]$ as

$$R^i(s, \mathbf{a}) = -A^i(s, \mathbf{a})\mathbf{1}_{\{\pi^{i,\mathrm{Nash}}(a^i|s)=0\}}\mathbf{1}_{\{\pi^{-i,\mathrm{Nash}}(a^{-i}|s)=1\}} + V^i(s) - \gamma \sum_{s'} P(s' \mid s, \mathbf{a})V^i(s') \tag{6}$$

$$\hat{R}^i(s, \mathbf{a}) = -\hat{A}^i(s, \mathbf{a})\mathbf{1}_{\{\hat{\boldsymbol{\pi}}^{i,\mathrm{Nash}}(a^i|s)=0\}}\mathbf{1}_{\{\hat{\boldsymbol{\pi}}^{-i,\mathrm{Nash}}(a^{-i}|s)=1\}} + \hat{V}^i(s) - \gamma \sum_{s'} \hat{P}(s' \mid s, \mathbf{a})\hat{V}^i(s') \tag{7}$$

As pointed out in Metelli et al. [2023], the rewards $\hat{R}^i(s, \mathbf{a})$ do not have to be bounded by the same $R^i(s, \mathbf{a})$ and therefore also not by the same $R_{\max}$. To fix this issue the authors point out, that the reward needs to be rescaled such that the recovered feasible reward set is bounded by the same value. In our case we have to be a bit more careful with the choice of the scaling, as we did not assume that the reward is bounded by 1. As we proof the existence of such reward function, we can choose $\tilde{V}^i(s) = V^i(s)$ for every $s \in \mathcal{S}$ and $\tilde{A}^i(s, \mathbf{a}) = A^i(s, \mathbf{a})$ for every $(s, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$, which results in a reward

$$\tilde{R}^i(s, \mathbf{a}) = -A^i(s, \mathbf{a})\mathbf{1}_{\{\hat{\pi}^{i,\mathrm{Nash}}(a^i|s)=0\}}\mathbf{1}_{\{\hat{\pi}^{-i,\mathrm{Nash}}(a^{-i}|s)=1\}} + V^i(s) + \gamma \sum_{s'} \hat{P}(s' \mid s, \mathbf{a})V^i(s').$$

Now we need to rescale the reward with $R_{\max} + |\epsilon^i(s, \mathbf{a})|$, where

$$\begin{aligned} \epsilon^i(s, \mathbf{a}) = {} & -A^i(s, \mathbf{a})\left(\mathbf{1}_{\{\pi^{i,\mathrm{Nash}}(a^i|s)=0\}}\mathbf{1}_{\{\pi^{-i,\mathrm{Nash}}(a^{-i}|s)=1\}} - \mathbf{1}_{\{\hat{\pi}^{i,\mathrm{Nash}}(a^i|s)=0\}}\mathbf{1}_{\{\hat{\pi}^{-i,\mathrm{Nash}}(a^{-i}|s)=1\}}\right) \\ & + \gamma \sum_{s'} (P(s' \mid s, \mathbf{a}) - \hat{P}(s, \mathbf{a}))V^i(s'), \end{aligned}$$

such that it remains bounded by $R_{\max}$, we receive

$$\begin{aligned} \hat{R}^i(s, \mathbf{a}) &= \tilde{R}^i(s, \mathbf{a})\frac{R_{\max}}{R_{\max} + |\epsilon^i(s, \mathbf{a})|} \\ &= -A^i(s, \mathbf{a})\frac{R_{\max}}{R_{\max} + |\epsilon^i(s, \mathbf{a})|}\mathbf{1}_{\{\hat{\pi}^{i,\mathrm{Nash}}(a^i|s)=0\}}\mathbf{1}_{\{\hat{\pi}^{-i,\mathrm{Nash}}(a^{-i}|s)=1\}} \\ &\quad + \frac{R_{\max}V^i(s)}{R_{\max} + |\epsilon^i(s, \mathbf{a})|} + \gamma \sum_{s'} \hat{P}(s' \mid s, \mathbf{a})\frac{R_{\max}V^i(s')}{R_{\max} + |\epsilon^i(s, \mathbf{a})|} \end{aligned}$$

It then follows that:

$$
\begin{aligned}
|R^i(s,\mathbf{a}) - \hat{R}^i(s,\mathbf{a})| &= |R^i(s,\mathbf{a}) - \frac{R_{\max}\tilde{R}^i(s,\mathbf{a})}{R_{\max} + |\epsilon^i(s,\mathbf{a})|}| \\
&\leq \frac{R_{\max}}{R_{\max} + |\epsilon^i(s,\mathbf{a})|} \left| \left(\frac{R_{\max} + |\epsilon^i(s,\mathbf{a})|}{R_{\max}} \right) R^i(s,\mathbf{a}) - \tilde{R}^i(s,\mathbf{a}) \right| \\
&\leq \frac{R_{\max}}{R_{\max} + |\epsilon^i(s,\mathbf{a})|} \left(|R^i(s,\mathbf{a}) - \tilde{R}^i(s,\mathbf{a})| + |\frac{\epsilon^i(s,\mathbf{a})}{R_{\max}} R^i(s,\mathbf{a})| \right) \\
&\leq \frac{R_{\max}}{R_{\max} + |\epsilon^i(s,\mathbf{a})|} \left(|\epsilon^i(s,\mathbf{a})| + |\epsilon^i(s,\mathbf{a})| \right) \leq \frac{R_{\max}}{R_{\max}} \left(|\epsilon^i(s,\mathbf{a})| + |\epsilon^i(s,\mathbf{a})| \right) \\
&= 2 \left(A^i(s,\mathbf{a})| \left(\mathbf{1}_E - \mathbf{1}_{\hat{E}} \right) | + \gamma \left| \sum_{s'} (P(s' \mid s,\mathbf{a}) - \hat{P}(s' \mid s,\mathbf{a})) V^i(s') \right| \right)
\end{aligned}
$$

□

Lemma D.6. *Let $\mathcal{G} \cup R$ be a n-person general-sum Markov Game, $P, \hat{P}$ two transition probabilities and $R, \hat{R}$ two reward functions, such that $\hat{\boldsymbol{\pi}}$ is a Nash equilibrium strategy in $\hat{\mathcal{G}} \cup \hat{R}$. Then, it holds true that:*

$$
\begin{aligned}
&V^i(\pi^i, \hat{\pi}^{-i}) - V^i(\hat{\pi}_i, \hat{\pi}^{-i}) \\
&\leq \sum_{s,\boldsymbol{a}} \overline{w}^{\hat{\boldsymbol{\pi}}}_{s,a} (R^i(s,\boldsymbol{a}) - \hat{R}^i(s,\boldsymbol{a}) + \gamma \sum_{s'} (\hat{P}(s' \mid s,\boldsymbol{a}) - P(s' \mid s,a) V^{i,\hat{\pi}}(s'))) \\
&\quad + \sum_{s,\boldsymbol{a}} \overline{w}^{\tilde{\boldsymbol{\pi}}}_{s,\boldsymbol{a}} (R^i(s,\boldsymbol{a}) - \hat{R}^i(s,\boldsymbol{a}) + \gamma \sum_{s'} (\hat{P}(s' \mid s,\boldsymbol{a}) - P(s' \mid s,\boldsymbol{a}) V^{i,\tilde{\boldsymbol{\pi}}}(s'))),
\end{aligned}
$$

where $\tilde{\boldsymbol{\pi}} = (\pi^i, \hat{\pi}^{-i})$.

Proof.

$$
\begin{aligned}
&V^i(\pi^i, \hat{\pi}^{-i}) - V^i(\hat{\pi}^i, \hat{\pi}^{-i}) \\
&= V^i(\pi^i, \hat{\pi}^{-i}) - \hat{V}^i(\pi^i, \hat{\pi}^{-i}) + \hat{V}^i(\hat{\pi}^i, \hat{\pi}^{-i}) - V^i(\hat{\pi}^i, \hat{\pi}^{-i}) + \hat{V}^i(\pi^i, \hat{\pi}^{-i}) - \hat{V}^i(\hat{\pi}^i, \hat{\pi}^{-i}) \\
&\leq V^i(\pi^i, \hat{\pi}^{-i}) - \hat{V}^i(\pi^i, \hat{\pi}^{-i}) + \hat{V}^i(\hat{\pi}^i, \hat{\pi}^{-i}) - V^i(\hat{\pi}^i, \hat{\pi}^{-i}) \\
&= \quad \sum_{s,\mathbf{a}} \overline{w}^{\hat{\boldsymbol{\pi}}}_{s,a} (R^i(s,\mathbf{a}) - \hat{R}^i(s,\mathbf{a}) + \gamma \sum_{s'} (\hat{P}(s' \mid s,\mathbf{a}) - P(s' \mid s,a)) V^{i,\hat{\pi}}(s')) \\
&\quad + \sum_{s,\mathbf{a}} \overline{w}^{\tilde{\boldsymbol{\pi}}}_{s,\mathbf{a}} (R^i(s,\mathbf{a}) - \hat{R}^i(s,\mathbf{a}) + \gamma \sum_{s'} (\hat{P}(s' \mid s,\mathbf{a}) - P(s' \mid s,\mathbf{a})) V^{i,\tilde{\boldsymbol{\pi}}}(s')),
\end{aligned}
$$

where we used that $\hat{V}^i(\hat{\pi}^i, \hat{\pi}^{-i}) - V^i(\hat{\pi}^i, \hat{\pi}^{-i}) \leq 0$ as $\hat{\boldsymbol{\pi}}$ is a NE policy and in the last equation we applied I.2. □

D.1 Sample Complexity analysis of the Uniform Sampling algorithm

In this section we give the proofs of the sample complexity for Uniform Sampling algorithm and lemmas derived in Theorem D.12 and Lemma D.8 The structure is as follows:

1. We first state the Uniform Sampling algorithm.
2. Then, we state the optimality criterion based on the Nash Imitation Gap.
3. Next, we present the Good Event Lemma bounds, using Hoeffding's inequality.
4. We define the reward uncertainty.
5. Then, we state a theorem that provides conditions—dependent on the derived confidence bounds—under which the optimality criterion holds.
6. Finally, we consolidate all results to prove the sample complexity bound for the uniform sampling algorithm.

Algorithm 1 MAIRL Uniform Sampling Algorithm with Generative Model

Require: Significance $\delta \in (0,1)$, target accuracy ϵ

```
Initialize k ← 0, ϵ_0 ← +∞
while ϵ_k > ϵ do
    Generate one sample for each (s, a) ∈ S × A
    Update P̂_k as described in (12)
    Update accuracy ϵ_k ← 1/(1−γ) max_(s,a) Ĉ_k(s, a)
end while
```

The algorithm, that we are evaluating in this section is given in Algorithm 1.

Now, we can restate the optimality criterion of the algorithm.

Definition D.7 (Optimality Criterion)**.** Let $\mathcal{R}_{(\mathcal{G},\boldsymbol{\pi}^{\text{Nash}})}$ be the exact feasible set and $\mathcal{R}_{(\hat{\mathcal{G}},\hat{\boldsymbol{\pi}}^{\text{Nash}})}$ the recovered feasible set after sampling $N \geq 0$ from $(\mathcal{G},\boldsymbol{\pi}^{\text{Nash}})$. We consider an algorithm to be (ϵ,δ,N)-correct after observing N samples if it holds with a probability of at least $1-\delta$:

$$\sup_{R\in\mathcal{R}_{(\mathcal{G},\boldsymbol{\pi}^{\text{Nash}})}} \inf_{\hat{R}^i\in\mathcal{R}_{(\hat{\mathcal{G}},\hat{\boldsymbol{\pi}}^{\text{Nash}})}} \max_{i\in[n]} \max_{\pi^i\in\pi^i} V^i_{\mathcal{G}\cup R}(\pi^i,\hat{\pi}^{-i}) - V^i_{\mathcal{G}\cup R}(\hat{\pi}^i,\hat{\pi}^{-i}) \leq \epsilon$$

$$\sup_{\hat{R}\in\mathcal{R}_{(\mathcal{G},\boldsymbol{\pi}^{\text{Nash}})}} \inf_{R\in\mathcal{R}_{(\hat{\mathcal{G}},\hat{\boldsymbol{\pi}}^{\text{Nash}})}} \max_{i\in[n]} \max_{\pi^i\in\pi^i} V^i_{\mathcal{G}\cup R}(\pi^i,\hat{\pi}^{-i}) - V^i_{\mathcal{G}\cup R}(\hat{\pi}^i,\hat{\pi}^{-i}) \leq \epsilon$$

We are now introducing the empirical estimator for the transition dynamic. For each iteration $k\in[K]$, let $n_k(s,\mathbf{a},s') = \sum_{t=1}^{k} \mathbf{1}_{(s_t,\mathbf{a}_t,s'_t)=(s,\mathbf{a},s')}$ denote the count of visits to the triplet $(s,\mathbf{a},s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, and let $n_k(s,\mathbf{a}) = \sum_{s'\in\mathcal{S}} n_k(s,\mathbf{a},s')$ denote the count of visits to the state-action pair $(s,\mathbf{a})$.

It is important to note the distinction here: the count of actions must be done separately for each agent, whereas the count of state visits needs to be done for any one of the agents. The cumulative count over all iterations $k\in[K]$ can then be written as:

$$N_k(s,\mathbf{a},s') = \sum_{j\in[k]} n_j(s,\mathbf{a},s'),$$

The cumulative state visit count are given by:

$$N_k^i(s,a^i) = \sum_{j\in[k]} n_j^i(s,a^i) \quad \forall i\in[n]$$

After introducing the empirical counts, we can now state the empirical estimators for the transition model:

$$\hat{P}_k(s' \mid s,\mathbf{a}) = \begin{cases} \frac{N_k(s,\mathbf{a},s')}{N_k(s,\mathbf{a})} & \text{if} \quad N_k(s,\mathbf{a})>0 \\ \frac{1}{S} & \text{otherwise} \end{cases} \tag{8}$$

Next, we state the lemma that derives the good event by applying Hoeffding's inequality. As the Nash equilibrium is assumed to be a pure one, meaning it is deterministic for every $s\in\mathcal{S}$, we only have to define the good event for the transition probability and for the experts policy we require, that for each agent we have seen each state only once.

Lemma D.8 (Good Event)**.** *Let k be the number of iterations and $\boldsymbol{\pi}^{\text{Nash}}$ be the stochastic expert policy. Furthermore let $\hat{\boldsymbol{\pi}}^{\text{Nash}}$ and $\hat{P}$ be the empirical estimates of the transition probability after k iterations as defined in Eq.* (8) *respectively. Then for $\delta\in(0,1)$, define the good event $\mathcal{E}$ as the event such that the following inequalities hold simultaneously for all $(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}$ and $k\geq 1$:*

$$\sum_{s'} \left| (P(s' \mid s,\boldsymbol{a}) - \hat{P}_k(s' \mid s,\boldsymbol{a}))V^i(s') \right| \leq \frac{R_{\max}}{1-\gamma}\sqrt{\frac{8l_k(s,\boldsymbol{a})}{N_k^+(s,\boldsymbol{a})}},$$

$$\sum_{s'} \left| (P(s' \mid s,\boldsymbol{a}) - \hat{P}_k(s' \mid s,\boldsymbol{a}))\hat{V}^i(s') \right| \leq \frac{R_{\max}}{1-\gamma}\sqrt{\frac{8l_k(s,\boldsymbol{a})}{N_k^+(s,\boldsymbol{a})}}.$$

where we introduced $l_k(s, \boldsymbol{a}) := \log\left(\frac{12|\mathcal{S}|\prod_i |\mathcal{A}^i|(N_k^+(s,\boldsymbol{a}))^2}{\delta}\right)$.

Proof. We start with bound the two last equations. Therefore we define $l_k(s, \mathbf{a}) = \log\left(\frac{12S\prod_i |\mathcal{A}^i|(N_k^+(s,\mathbf{a}))^2}{\delta}\right)$ and additionally we denote $\beta_{N_k(s,\mathbf{a})} = \frac{R_{\max}}{1-\gamma}\sqrt{\frac{2l_k(s,\mathbf{a})}{N_k^+(s,\mathbf{a})}}$. Now we define the set

$$\mathcal{E}^{\text{trans}} := \left\{\forall k \in \mathbb{N},\, \forall (s, \mathbf{a}) \in \mathcal{S} \times \mathcal{A} : \sum_{s' \in \mathcal{S}} |(P(s' \mid s, \mathbf{a}) - \hat{P}(s' \mid s, \mathbf{a}))V^i(s')| \leq \beta_{N_k(s,\mathbf{a})}\right\}.$$

Then we get for V^i with probability of $1 - \delta$:

$$\begin{aligned}
&\mathbb{P}\left((\mathcal{E}^{\text{trans}})^C\right) \\
&= \mathbb{P}\left(\exists k \geq 1, \exists (s, \mathbf{a}) \in \mathcal{S} \times \mathcal{A} : \sum_{s'} \left|\left(P(s' \mid s, \mathbf{a}) - \hat{P}_k(s' \mid s, \mathbf{a})\right) V^i(s')\right| > \beta_{N_k(s,\mathbf{a})}(s, \mathbf{a})\right) \\
&\overset{(a)}{\leq} \sum_{(s,\mathbf{a}) \in \mathcal{S}\times\mathcal{A}} \mathbb{P}\left(\exists k \geq 1 : \sum_{s'} \left|\left(P(s' \mid s, \mathbf{a}) - \hat{P}_k(s' \mid s, \mathbf{a})\right) V^i(s')\right| > \beta_{N_k(s,\mathbf{a})}(s, \mathbf{a})\right) \\
&\overset{(b)}{\leq} \sum_{(s,\mathbf{a}) \in \mathcal{S}\times\mathcal{A}} \mathbb{P}\left(\exists m \geq 0 : \sum_{s'} \left|\left(P(s' \mid s, \mathbf{a}) - \hat{P}_k(s' \mid s, \mathbf{a})\right) V^i(s')\right| > \beta_{N_k(s,\mathbf{a})}(s, \mathbf{a})\right) \\
&\overset{(c)}{\leq} \sum_{(s,\mathbf{a}) \in \mathcal{S}\times\mathcal{A}} \sum_{m \geq 0} 2\exp\left(\frac{\beta^2_{N_k(s,\mathbf{a})} m^2 (1-\gamma)^2}{4m\gamma^2 R^2_{\max}}\right) \\
&\leq \sum_{(s,\mathbf{a}) \in \mathcal{S}\times\mathcal{A}} \sum_{m \geq 0} \frac{\delta}{6|\mathcal{S}|(\prod_{i=1}^n |\mathcal{A}^i|)(m^+)^2} \\
&\leq \frac{\delta}{6}\left(1 + \frac{\pi^2}{6}\right) \leq \frac{\delta}{2},
\end{aligned}$$

where (a) uses a union bound over the state and joint-action space, (b) uses that we only consider the m-times, where we updated the estimated transition model and (c) uses an union bound over the update times m and an application of Hoeffding's inequality combined with the fact that we can bound the value function, i.e. $V^i(s') \leq \frac{R_{\max}}{1-\gamma}$ for every $s' \in \mathcal{S}$. □

Following, we present the reward uncertainty metric, which allows us to demonstrate that the difference between the recovered reward function and the true reward function is bounded.

Definition D.9 (Reward Uncertainty)**.** Let k be the number of iterations. Then the reward uncertainty after k iterations for any $(s, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$ is defined as

$$C_k(s, \mathbf{a}) := \frac{4\gamma R_{\max}}{1-\gamma}\left(\sqrt{\frac{8l_k(s, \mathbf{a})}{N_k^+(s, \mathbf{a})}}\right).$$

Theorem D.10. *Let the reward uncertainty be defined as in D.9. Under the good event it holds for any* $(s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}$ *that:*

$$|R^i(s, \boldsymbol{a}) - \hat{R}^i(s, \boldsymbol{a})| \leq C_k(s, \boldsymbol{a}).$$

Proof. The theorem is an application of the error propagation Theorem D.5, followed by Lemma D.8

$$
\begin{aligned}
|R^i(s,\mathbf{a}) - \hat{R}^i(s,\mathbf{a})| &\le 2\left(A^i(s,\mathbf{a})|\mathbf{1}_E - \mathbf{1}_{\hat{E}}| + \gamma \sum_{s'} |(P(s' \mid s,\mathbf{a}) - \hat{P}(s' \mid s,\mathbf{a}))V^i(s')|\right) \\
&\le \frac{4\gamma R_{\max}}{1-\gamma}\left(\sqrt{\frac{2l_k(s,\mathbf{a})}{N_k^+(s,\mathbf{a})}}\right) \\
&\le \frac{4\gamma R_{\max}}{1-\gamma}\left(\sqrt{\frac{8l_k(s,\mathbf{a})}{N_k^+(s,\mathbf{a})}}\right) \\
&= C_k(s,\mathbf{a}).
\end{aligned}
$$

□

Corollary D.11. *Let k be the number of iterations for any allocation of the samples over the state-action space $\mathcal{S}\times\mathcal{A}$. Furthermore, let $\mathcal{R}_{(\mathcal{G},\boldsymbol{\pi}^{\text{Nash}})}$ be the true feasible set and $\mathcal{R}_{(\hat{\mathcal{G}},\hat{\boldsymbol{\pi}}^{\text{Nash}})}$ the recovered one. Then the optimality criterion 3.3 holds true, if*

$$
\frac{1}{1-\gamma}\max_{(s,\boldsymbol{a})} C_k(s,\boldsymbol{a}) \le \frac{\epsilon}{2}.
$$

Proof. We complete the proof for the first case of the optimality criterion, the second one follows analogously.

$$
\begin{aligned}
&\sup_{R\in\mathcal{R}_{(\mathcal{G},\boldsymbol{\pi}^{\text{Nash}})}} \inf_{\hat{R}^i\in\mathcal{R}_{(\hat{\mathcal{G}},\hat{\boldsymbol{\pi}}^{\text{Nash}})}} \max_{i\in[n]}\max_{\pi^i\in\Pi^i}\left(V^i_{\mathcal{G}\cup R}(\pi^i,\hat{\pi}^{-i}) - V^i_{\mathcal{G}\cup R}(\hat{\pi}^i,\hat{\pi}^{-i})\right) \\
&\overset{(a)}{\le} \sup_{R\in\mathcal{R}_{(\mathcal{G},\boldsymbol{\pi}^{\text{Nash}})}} \inf_{\hat{R}^i\in\mathcal{R}_{(\hat{\mathcal{G}},\hat{\boldsymbol{\pi}}^{\text{Nash}})}} \max_{i,\pi^i}\left(\sum_{s,\mathbf{a}}\overline{w}^{\hat{\pi}}_{s,\mathbf{a}}\left(R^i(s,\mathbf{a}) - \hat{R}^i(s,\mathbf{a}) + \gamma\sum_{s'}(\hat{P}-P)(s'\mid s,\mathbf{a})V^{i,\hat{\pi}}(s')\right)\right. \\
&\qquad\qquad \left. + \sum_{s,\mathbf{a}}\overline{w}^{\tilde{\pi}}_{s,\mathbf{a}}\left(R^i(s,\mathbf{a}) - \hat{R}^i(s,\mathbf{a}) + \gamma\sum_{s'}(\hat{P}-P)(s'\mid s,\mathbf{a})V^{i,\tilde{\pi}}(s')\right)\right) \\
&\overset{(b)}{\le} \sup_{R\in\mathcal{R}_{(\mathcal{G},\boldsymbol{\pi}^{\text{Nash}})}} \inf_{\hat{R}^i\in\mathcal{R}_{(\hat{\mathcal{G}},\hat{\boldsymbol{\pi}}^{\text{Nash}})}} \max_{i,\pi^i} 2\sum_{s,\mathbf{a}}\overline{w}^{\pi}_{s,\mathbf{a}}\left(R^i(s,\mathbf{a}) - \hat{R}^i(s,\mathbf{a}) + \gamma\sum_{s'}(\hat{P}-P)(s'\mid s,\mathbf{a})V^{i,\pi}(s')\right) \\
&\overset{(c)}{\le} \sup_{R\in\mathcal{R}_{(\mathcal{G},\boldsymbol{\pi}^{\text{Nash}})}} \inf_{\hat{R}^i\in\mathcal{R}_{(\hat{\mathcal{G}},\hat{\boldsymbol{\pi}}^{\text{Nash}})}} \max_{i,\pi^i} 2\sum_{s,\mathbf{a}}\overline{w}^{\pi}_{s,\mathbf{a}}\left(A^i(s,\mathbf{a})\left|\mathbf{1}_E - \mathbf{1}_{\hat{E}}\right| + 2\gamma\sum_{s'}\left|(\hat{P}-P)(s'\mid s,\mathbf{a})V^i(s')\right|\right) \\
&\overset{(d)}{\le} \frac{2}{1-\gamma}\max_{(s,\mathbf{a})} C_k(s,\mathbf{a}) \le \epsilon.
\end{aligned}
$$

where in (a) we applied D.6; in (b) we used the fact that $a+b\le 2\max\{a,b\}$ and denoted the corresponding policy as $\boldsymbol{\pi}$; in (c) we used the error propagation Theorem D.5 and in (d) we used D.10. □

We can combine the derived results to now state the main theorem regarding the sample complexity of allocating the samples uniformly over the state action space.

Theorem D.12 (Sample Complexity of Uniform Sampling MAIRL)**.** *Allocating the samples uniformly (see Algorithm 1) over the state and (joint-) action space stops with a probability of at least $1-\delta$ after iteration τ and satisfies the optimality criterion (see 3.3), where the sample complexity is of order*

$$
\tilde{\mathcal{O}}\left(\frac{\gamma^2 R_{\max}^2|\mathcal{S}|\prod_{i=1}^n|\mathcal{A}^i|}{(1-\gamma)^4\epsilon^2}\right)
$$

Proof. We know from D.11, that we need

$$\frac{1}{1-\gamma}\max_{(s,\mathbf{a})} C_k(s,\mathbf{a}) \leq \frac{\epsilon_k}{2}$$
$$\Leftrightarrow \frac{2R_{\max}}{(1-\gamma)^2}\max_{(s,\mathbf{a})}\left(\gamma\sqrt{\frac{8l_k(s,\mathbf{a})}{N_k^+(s,\mathbf{a})}}\right) \leq \frac{\epsilon_k}{2}$$

This is satisfied if

$$\frac{4\gamma R_{\max}}{(1-\gamma)^2}\sqrt{\frac{8l_k(s,\mathbf{a})}{N_k^+(s,\mathbf{a})}} \leq \frac{\epsilon_k}{2}$$

To achieve the first condition, we get

$$N_k(s,\mathbf{a}) \geq \frac{R_{\max}}{(1-\gamma)^4}\gamma^2 8 l_k(s,\mathbf{a})\frac{8}{\epsilon_k^2} = \frac{\gamma^2 64 R_{\max}}{(1-\gamma)^4\epsilon_k^2}\log\left(\frac{12S\prod_i|\mathcal{A}^i|(N_k^+(s,\mathbf{a}))^2}{\delta}\right)$$

Applying Lemma B.8 by Metelli et al. [2021] we get that

$$N_k(s,\mathbf{a}) \leq \frac{256\gamma^2 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2}\log\left(\frac{128\gamma^2 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2}\sqrt{\frac{12|\mathcal{S}|\prod_i|\mathcal{A}^i|}{\delta}}\right).$$

At each iteration we are allocating the samples uniformly over $\mathcal{S}\times\mathcal{A}$ and recalling that $\tau_{s,a} = S\prod_i|\mathcal{A}^i|N_k(s,\mathbf{a})$ therefore we get

$$\tau \leq \frac{256|\mathcal{S}|\prod_i|\mathcal{A}^i|\gamma^2 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2}\log\left(\frac{128\gamma^2 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2}\sqrt{\frac{12|\mathcal{S}|\prod_i|\mathcal{A}^i|}{\delta}}\right)$$

Now we only have to achieve that we have seen each state at least once, to correctly estimate the policies for every agent. Therefore, we force that $N_k(s) \geq 1$. As we here need to allocate samples uniformly over the state space only but for every agent separately and recalling that $\tau_s = |\mathcal{S}|N_k(s)$, we get

$$\tau_s \leq n|S|$$

With $\tau = \tau_{s,a} + \tau_s$ we get in total

$$\tau \leq \frac{128|\mathcal{S}|\prod_i|\mathcal{A}^i|\gamma^2 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2}\log\left(\frac{64a\gamma^2 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2}\sqrt{\frac{12|\mathcal{S}|\prod_i|\mathcal{A}^i|}{\delta}}\right) + n|\mathcal{S}|.$$

This is exactly of order

$$\tilde{\mathcal{O}}\left(\frac{\gamma^2 R_{\max}^2|\mathcal{S}|\prod_{i=1}^n|\mathcal{A}^i|}{(1-\gamma)^4\epsilon^2}\right)$$

□

E Hardness result

In this section, we want to quantify the non-expressiveness of the recovered feasible reward set under a single NE observation. Therefore, we give the following hardness result. The idea is that the observed NE only covers a part of the environment and the definition of NE only ensures robustness against single agent deviations.

Theorem E.1. *Let us consider any IRL algorithm* $\mathrm{Alg}_{\mathrm{IRL}}$ *that chooses* $\hat{R} \in \mathcal{R}_{(\hat{\mathcal{G}},\hat{\boldsymbol{\pi}}^{\mathrm{Nash}})}$ *that is not a constant reward, i.e.* $\hat{R} \neq C$ *for* $C \in [-R_{\max}, R_{\max}]$*. Furthermore consider a forward MARL algorithm* $\mathrm{Alg}_{\mathrm{MARL}}$ *that guarantees learning a policy* $\tilde{\boldsymbol{\pi}} \in \Pi^{\mathrm{Nash}}$*. Then, there exists a Markov Game, such that even if* $\hat{\boldsymbol{\pi}} \in \Pi_{\mathrm{Nash}}$ *and* $\hat{R} \in \mathcal{R}_{(\mathcal{G},\boldsymbol{\pi}^{\mathrm{Nash}})}$ *it holds true that* $\mathcal{E}(\boldsymbol{\pi}')$ *is of order* $(1-\gamma)^{-1}$*.*

Proof. We consider the following 2-player general-sum Markov Game $\mathcal{G}\cup R$. Let $\mathcal{S} = \{s_0, s_1, s_2, s_3, s_4\}$ with $\mathcal{A}_1 = \mathcal{A}_2 = \{a_1, a_2\}$. Furthermore, let the transition dynamics be given by

$$P(\cdot \mid s_0, \mathbf{a}) = \begin{cases} s_1 & \text{if } \mathbf{a} = (a_1 a_1), \\ s_2 & \text{if } \mathbf{a} \in \{(a_1 a_2), (a_2 a_1)\}, \\ s_3 & \text{if } \mathbf{a} = (a_2, a_2), \end{cases}$$

In states s_1, s_2, s_3 the transition is defined to stay in the respective state with probability 1. Furthermore, let the true reward of the Markov Game be given by

$$R^1(s_0, \mathbf{a}) = \begin{cases} 3 & \text{if } \mathbf{a} = (a_1 a_1), \\ 0 & \text{if } \mathbf{a} \in \{(a_1 a_2), (a_2 a_1)\}, \\ 2 & \text{if } \mathbf{a} = (a_2, a_2), \end{cases} \qquad R^2(s_0, \mathbf{a}) = \begin{cases} 2 & \text{if } \mathbf{a} = (a_1 a_1), \\ 0 & \text{if } \mathbf{a} \in \{(a_1 a_2), (a_2 a_1)\}, \\ 3 & \text{if } \mathbf{a} = (a_2, a_2), \end{cases}$$

For the other states we have that $R(s_1, \mathbf{a}) = R(s_3, \mathbf{a}) = 1 \, \forall \mathbf{a} \in \mathcal{A}$ and $R(s_2, \mathbf{a}) = 0 \, \forall \mathbf{a} \in \mathcal{A}$. This indicates that the Markov Game has two pure NE strategies ${\boldsymbol{\pi}^{\mathrm{Nash}}}_1$ with$\pi_1(a_1 \mid s_0) = \pi_2(a_1 \mid s_0) = 1$ and ${\boldsymbol{\pi}^{\mathrm{Nash}}}_2$ with $\pi_1(a_2 \mid s_0) = \pi_2(a_2 \mid s_0) = 1$ and any distribution in states s_1, s_2, s_3. Note that this game can be seen as a Markov game extension of the Normal Form Game Battle of the Saxes that rewards the NE strategies in subsequent states. An illustration of this game can be found in Fig. 2. Let us assume that the observed Nash equilibrium is ${\boldsymbol{\pi}^{\mathrm{Nash}}}_1$. Next, we apply any$\mathrm{Alg}_{\mathrm{IRL}}$ that returns $\hat{R} \in \mathcal{R}_{(\hat{\mathcal{G}}, \hat{\boldsymbol{\pi}}^{\mathrm{Nash}})}$. Note, that the Nash equilibrium for the state s_0 and s_1 will be recovered perfectly and a potential $\hat{R}$ is given by

$$\hat{R}^1(s_0, \mathbf{a}) = \begin{cases} 2 & \text{if } \mathbf{a} = (a_1 a_1), \\ -1 & \text{if } \mathbf{a} = (a_1 a_2), \\ 2 & \text{if } \mathbf{a} = (a_2 a_1), \\ -2 & \text{if } \mathbf{a} = (a_2, a_2), \end{cases} \qquad \hat{R}^2(s_0, \mathbf{a}) = \begin{cases} 2 & \text{if } \mathbf{a} = (a_1 a_1), \\ 2 & \text{if } \mathbf{a} = (a_1 a_2), \\ -1 & \text{if } \mathbf{a} = (a_2 a_1), \\ -2 & \text{if } \mathbf{a} = (a_2, a_2), \end{cases}$$

Additionally, the rewards $\hat{R}(s_1, \mathbf{a}) = 1, \hat{R}(s_3, \mathbf{a}) = -1 \, \forall \mathbf{a} \in \mathcal{A}$ and $R(s_2, \mathbf{a}) = 1 \, \forall \mathbf{a} \in \mathcal{A}$. Then, we note that $\hat{\boldsymbol{\pi}}$ is indeed a NE under $\hat{R}$ and also in the true underlying environment $\mathcal{G} \cup R$.

However, in the recovered Markov Game $\hat{\mathcal{G}} \cup \hat{R}$ there exists another pure equilibrium solutions ${\boldsymbol{\pi}^{\mathrm{Nash}}}_3, {\boldsymbol{\pi}^{\mathrm{Nash}}}_4 \notin \Pi^{\mathrm{Nash}}_{\mathcal{G} \cup R}$, where for ${\boldsymbol{\pi}^{\mathrm{Nash}}}_3$ we have $\pi^1_3(a_1 \mid s_0) = 1$ and $\pi^2_3(a_2 \mid s_0) = 1$ and ${\boldsymbol{\pi}^{\mathrm{Nash}}}_4$ is given by $\pi^1_3(a_2 \mid s_0) = 1$ and $\pi^2_3(a_1 \mid s_0) = 1$.

If one now applies a forward MARL algorithm that guarantees convergence to a any (pure) NE defined by $\tilde{\boldsymbol{\pi}}$, i.e. satisfying

$$\mathcal{E}_{\hat{R}} = \max_{i \in \{1,2\}} \max_{\pi^i \in \Pi^i} V^i_{\hat{R}}(\pi^i, \tilde{\pi}^{-i}) - V^i_{\hat{R}}(\tilde{\boldsymbol{\pi}}) = 0.$$

Then, assuming that $\tilde{\boldsymbol{\pi}}$ in the true Markov Game it holds true that

$$\mathbb{E}_{\mathrm{Alg}_{\mathrm{MARL}}}[\mathcal{E}_R(\tilde{\boldsymbol{\pi}})] \geq \frac{1}{(1-\gamma)} \mathbb{P}(\tilde{\boldsymbol{\pi}} \neq {\boldsymbol{\pi}^{\mathrm{Nash}}}_1).$$

This holds true because for $\tilde{\boldsymbol{\pi}} = {\boldsymbol{\pi}^{\mathrm{Nash}}}_1$, we have that $\mathcal{E}({\boldsymbol{\pi}^{\mathrm{Nash}}}_1) = 0$ as this is also a NE in the original Markov Game. However for ${\boldsymbol{\pi}^{\mathrm{Nash}}}_3$ and ${\boldsymbol{\pi}^{\mathrm{Nash}}}_4$ which both go to state s_2 a Best response would either be to go s_1 or s_3 resulting in a exploitability for 1 for all future states. Assuming that the algorithm returns a NE uniformly across the set of NE, we get

$$\mathbb{E}_{\mathrm{Alg}_{\mathrm{MARL}}}[\mathcal{E}_R(\tilde{\boldsymbol{\pi}})] \geq \frac{1}{(1-\gamma)} \mathbb{P}(\tilde{\boldsymbol{\pi}} \neq {\boldsymbol{\pi}^{\mathrm{Nash}}}_1) = \frac{2}{3(1-\gamma)}.$$

This is exactly of the order $(1-\gamma)^{-1}$ and completes the proof. □

Next, we want to provide further intuition on this phenomenon by giving the Normal Form Game that is the origin of the considered Markov Game instance.

Example E.1. We consider the general form of a coordination game as a Normal Form Game (NFG):

	Player 2: Stag	Player 2: Hare
Player 1: Stag	(A, A)	(C, B)
Player 1: Hare	(B, C)	(D, D)

In general coordination games, we have that $D > B$ and $D - B < A + D - B - C$. Assume we observe the pure Nash equilibrium strategy $(\text{Stag}, \text{Stag})$. The feasible reward set $\mathcal{R}$ then contains all rewards that satisfy:

$$R^1(\text{Stag}, \text{Stag}) \geq R^1(\text{Hare}, \text{Stag}) \wedge R^2(\text{Stag}, \text{Stag}) \geq R^2(\text{Stag}, \text{Hare}),$$

while for all other reward values, **any** rewards are feasible, i.e., $R^1(\text{Stag}, \text{Hare})$, $R^1(\text{Hare}, \text{Hare}) \in \mathbb{R}^{\mathcal{A}_1 \times \mathcal{A}_2}$.

This flexibility in reward specification allows for undesirable scenarios (see Example E.1), such as:

- **Changing the nature of the game**: The game can transform into an anti-coordination variant with additional pure Nash equilibria not present in the original game.
- **Losing equilibria**: Rewards can be defined so that "Stag" becomes the unique dominant strategy for player 1.

The following are two examples of feasible rewards if observing the NE expert $(\text{Stag}, \text{Stag})$:

	Player 2: Stag	Player 2: Hare
Player 1: Stag	$(2, 2)$	$(0, 0)$
Player 1: Hare	$(0, 0)$	$(-1, -1)$

	Player 2: Stag	Player 2: Hare
Player 1: Stag	$(2, 2)$	$(-1, 2)$
Player 1: Hare	$(2, -1)$	$(-10, -10)$

This example highlights that even in simple NFGs, the feasible reward set encompasses too many reward functions, including those that significantly alter the game's equilibria. This contrasts with the single-agent IRL setting, where the feasible reward set contains degenerate rewards like constant ones, but due to the fact that all equilibria obtain the same value, preserving the meaning of the environment. In the multi-agent setting, this second source of ambiguity allows for strategic behavior entirely absent in the original game, which is highly undesirable if the goal is to recover meaningful reward functions for transfer to new environments.

F Proofs of Section 3.2

In this section, we will give the missing proofs of Section 3.2. We start by giving again the definition of a feasible reward function for an observed pair of expert policies. In particular, if the observed expert policy is a QRE equilibrium.

Definition F.1 (Feasible Reward Set (regularized))**.** A reward function R is feasible for an MAIRL problem $(\mathcal{G}, (\mu^*, \nu*))$ if and only if the observed policy pair forms an equilibrium in $\mathcal{G} \cup R$.

Additionally, we will restate Definition 3.3 in terms of regularized games.

Definition F.2 (Regularized Optimality Criterion)**.** Let $\mathcal{R} := \mathcal{R}_{(\mathcal{G},(\mu^*,\nu^*))}$ be the exact feasible set and $\hat{\mathcal{R}} := \mathcal{R}_{(\hat{\mathcal{G}},(\hat{\mu}^*,\hat{\nu}^*))}$ the recovered feasible set after observing $N \geq 0$ samples from the underlying MAIRL problem $(\mathcal{G}, \boldsymbol{\pi}^{\text{Nash}})$. We consider an algorithm to be (ϵ, δ, N)-correct after observing N samples if with a probability of at least $1 - \delta$ it holds:

$$\sup_{R \in \mathcal{R}} \inf_{\hat{R} \in \hat{\mathcal{R}}} \max\{\max_{\mu} V_\lambda^1(\mu, \hat{\nu}) - V_\lambda^1(\hat{\mu}, \hat{\nu}), \max_{\nu} V_\lambda^2(\hat{\mu}, \nu) - V_\lambda^2(\hat{\mu}, \hat{\nu})\} \leq \epsilon$$
$$\sup_{\hat{R} \in \hat{\mathcal{R}}} \inf_{R \in \mathcal{R}} \max\{\max_{\mu} V_\lambda^1(\mu, \hat{\nu}) - V_\lambda^1(\hat{\mu}, \hat{\nu}), \max_{\nu} V_\lambda^2(\hat{\mu}, \nu) - V_\lambda^2(\hat{\mu}, \hat{\nu})\} \leq \epsilon.$$

Next, note that a policy is considered optimal, i.e. a QRE equilibrium, if the policies satisfy

$$\mu^*(a \mid s) = \frac{\exp\left(\frac{1}{\lambda}\sum_{b' \in \mathcal{B}} \nu^*(b' \mid s) Q_\lambda^{*,1}(s, a, b')\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\frac{1}{\lambda}\sum_{b' \in \mathcal{B}} \nu^*(b' \mid s) Q_\lambda^{*,1}(s, a', b')\right)}, \tag{9}$$

$$\nu^*(b \mid s) = \frac{\exp\left(\frac{1}{\lambda}\sum_{a' \in \mathcal{A}} \mu^*(a' \mid s) Q_\lambda^{*,2}(s, a', b)\right)}{\sum_{b' \in \mathcal{A}} \exp\left(\frac{1}{\lambda}\sum_{a' \in \mathcal{A}} \mu^*(a' \mid s) Q_\lambda^{*,2}(s, a', b')\right)}. \tag{10}$$

Similarly to the analysis done in the single-agent setting the goal is to derive an explicit characterization of the reward function [Metelli et al., 2023, Lindner et al., 2022, Metelli et al., 2023, Zhao et al., 2024, Cao et al., 2021]. The idea is to rewrite the formulation of the optimal policy in terms of the reward function by using the definition of the value function and the Q-function. We will present the analysis only for player 1, it holds analogously for player 2. For a better readability we drop the superscript for the player.

Lemma F.3 (Feasible Explicit (regularized))**.** *A reward function R for the regularized Markov Game i feasible if and only if there exists $V \in \mathbb{R}^{\mathcal{S}}$ and $|\mathcal{B}|-1$ many functions $R' \in [-R_{\max}, R_{\max}]^{\mathcal{S}\times\mathcal{A}\times\mathcal{B}}$ such that for all (s,a,b) it holds that*

$$R(s,a,b) = \tfrac{1}{\nu^*(b|s)}\left(\lambda \log(\mu^*(a \mid s)) + V(s) - \gamma \textstyle\sum_{s'} \sum_{b' \in \mathcal{B}} \nu^*(b' \mid s) P(s' \mid s,a,b') V(s') - \sum_{b' \neq b} \nu^*(b' \mid s) R(s,a,b')\right).$$

Proof. First, assume that the reward function R is feasible. By definition, this implies that μ^* is an optimal policy for agent 1 under R when agent 2 plays ν^*. Let $V_\lambda^*(s)$ be the corresponding unique entropy-regularized optimal value function for agent 1. The optimal policy $\mu^*(a \mid s)$ (see Eq. (9)) is given by

$$\mu^*(a \mid s) = \frac{\exp\left(\frac{1}{\lambda}\sum_{b'\in\mathcal{B}} \nu^*(b' \mid s) Q_\lambda^*(s,a,b')\right)}{\sum_{a'\in\mathcal{A}} \exp\left(\frac{1}{\lambda}\sum_{b'\in\mathcal{B}} \nu^*(b' \mid s) Q_\lambda^*(s,a',b')\right)}.$$

Recognizing that the denominator relates to the soft value function $V_\lambda^*(s) = \lambda \log \sum_{a\in\mathcal{A}} \exp\left(\frac{1}{\lambda}\sum_{b'\in\mathcal{B}} \nu^*(b' \mid s) Q_\lambda^*(s,a,b')\right)$, we can write

$$\mu^*(a \mid s) = \exp\left(\frac{1}{\lambda}\left(\sum_{b'\in\mathcal{B}} \nu^*(b' \mid s) Q_\lambda^*(s,a,b') - V_\lambda^*(s)\right)\right).$$

Using the definition of the Q-function, $Q_\lambda^*(s,a,b') = R(s,a,b') + \gamma \sum_{s'\in\mathcal{S}} P(s' \mid s,a,b') V_\lambda^*(s')$, we substitute this into the expression for $\mu^*(a \mid s)$:

$$\begin{aligned}&\mu^*(a \mid s)\\ &= \exp\left(\frac{1}{\lambda}\left(\sum_{b'\in\mathcal{B}} \nu^*(b' \mid s) R(s,a,b') + \gamma \sum_{b'\in\mathcal{B}} \nu^*(b' \mid s) \sum_{s'\in\mathcal{S}} P(s' \mid s,a,b') V_\lambda^*(s') - V_\lambda^*(s)\right)\right).\end{aligned}$$

Taking the logarithm and rearranging terms yields

$$\textstyle\sum_{b'\in\mathcal{B}} \nu^*(b' \mid s) R(s,a,b') = \lambda \log(\mu^*(a \mid s)) + V_\lambda^*(s) - \gamma \sum_{b'\in\mathcal{B}} \nu^*(b' \mid s) \sum_{s'\in\mathcal{S}} P(s' \mid s,a,b') V_\lambda^*(s').$$

Let $K_{V_\lambda^*}(s,a)$ denote the right-hand side of the equation above. Then, for any specific action $b \in \mathcal{B}$ such that $\nu^*(b \mid s) > 0$, we can express $R(s,a,b)$ as

$$R(s,a,b) = \frac{1}{\nu^*(b \mid s)}\left(K_{V_\lambda^*}(s,a) - \sum_{b'\in\mathcal{B}\setminus\{b\}} \nu^*(b' \mid s) R(s,a,b')\right).$$

This matches the form specified in the lemma, where $V(s)$ is taken as $V_\lambda^*(s)$, and the $|\mathcal{B}|-1$ functions R' correspond to the components $R(s,a,b')$ for $b' \neq b$ from the original feasible reward R.

For the opposing direction, assume there exists an arbitrary function $V \in \mathbb{R}^{\mathcal{S}}$ and a reward function R (composed of $|\mathcal{B}|-1$ given functions $R'(s,a,b')$ for $b' \neq b$ that are within $[-R_{\max}, R_{\max}]$, and the remaining component $R(s,a,b)$ defined by the formula) such that for all (s,a,b) it holds that

$$\begin{aligned}R(s,a,b) = \frac{1}{\nu^*(b \mid s)}\Bigg(&\lambda \log(\mu^*(a \mid s)) + V(s) - \gamma \sum_{s'\in\mathcal{S}} \sum_{b''\in\mathcal{B}} \nu^*(b'' \mid s) P(s' \mid s,a,b'') V(s')\\ &- \sum_{b'\in\mathcal{B}\setminus\{b\}} \nu^*(b' \mid s) R(s,a,b')\Bigg).\end{aligned}$$

This structural definition implies that the expected reward for agent 1, $R^{\nu^*}(s,a) = \sum_{b'\in\mathcal{B}} \nu^*(b' \mid s) R(s,a,b')$, satisfies

$$R^{\nu^*}(s,a) = \lambda \log(\mu^*(a \mid s)) + V(s) - \gamma \sum_{s'\in\mathcal{S}} \sum_{b'\in\mathcal{B}} \nu^*(b' \mid s) P(s' \mid s,a,b') V(s'). \tag{11}$$

Let $P^{\nu^*}(s' \mid s, a) = \sum_{b' \in \mathcal{B}} \nu^*(b' \mid s) P(s' \mid s, a, b')$ be the expected transition probability from agent 1's perspective. Then (11) becomes $R^{\nu^*}(s,a) = \lambda \log(\mu^*(a \mid s)) + V(s) - \gamma \sum_{s' \in \mathcal{S}} P^{\nu^*}(s' \mid s, a) V(s')$. We now show that R is feasible by demonstrating that $V(s)$ is the value function of policy μ^* for agent 1 (given agent 2 plays ν^*) and that μ^* is the optimal policy.

First, let $V^{\mu^*}(s)$ be the value function for agent 1 when it follows policy μ^* and agent 2 follows ν^*, with rewards $R(s, a, b)$. The Bellman equation for $V^{\mu^*}(s)$ is given by

$$V^{\mu^*}(s) = \sum_{a \in \mathcal{A}} \mu^*(a|s) \left(R^{\nu^*}(s,a) + \gamma \sum_{s' \in \mathcal{S}} P^{\nu^*}(s'|s,a) V^{\mu^*}(s') - \lambda \log \mu^*(a|s) \right).$$

Substituting the expression for $R^{\nu^*}(s,a)$ from (11):

$$\begin{aligned} V^{\mu^*}(s) &= \sum_{a \in \mathcal{A}} \mu^*(a|s) \left(\left[\lambda \log(\mu^*(a \mid s)) + V(s) - \gamma \sum_{s' \in \mathcal{S}} P^{\nu^*}(s' \mid s,a) V(s') \right] \right. \\ &\qquad \left. + \gamma \sum_{s' \in \mathcal{S}} P^{\nu^*}(s'|s,a) V^{\mu^*}(s') - \lambda \log \mu^*(a|s) \right) \\ &= \sum_{a \in \mathcal{A}} \mu^*(a|s) \left(V(s) - \gamma \sum_{s' \in \mathcal{S}} P^{\nu^*}(s' \mid s,a) V(s') + \gamma \sum_{s' \in \mathcal{S}} P^{\nu^*}(s'|s,a) V^{\mu^*}(s') \right). \end{aligned}$$

Since $\sum_{a \in \mathcal{A}} \mu^*(a|s) = 1$, we have

$$V^{\mu^*}(s) = V(s) + \gamma \sum_{a \in \mathcal{A}} \mu^*(a|s) \sum_{s' \in \mathcal{S}} P^{\nu^*}(s' \mid s,a)(V^{\mu^*}(s') - V(s')).$$

Let $g(s) = V^{\mu^*}(s) - V(s)$. Then $g(s) = \gamma \sum_{a \in \mathcal{A}} \mu^*(a|s) \sum_{s' \in \mathcal{S}} P^{\nu^*}(s' \mid s,a) g(s')$. This equation, $g = \gamma \mathcal{P}_{\mu^*,\nu^*} g$, where $\mathcal{P}_{\mu^*,\nu^*}$ is the Bellman operator for policy evaluation, implies that $g(s) = 0$ is the unique solution since $\gamma \in [0,1)$ ensures $\mathcal{P}_{\mu^*,\nu^*}$ is a contraction. Thus, $V^{\mu^*}(s) = V(s)$ for all $s \in \mathcal{S}$.

Next, we show that $\mu^*(a|s)$ is the entropy-regularized optimal policy for agent 1. The optimal policy, $\pi^{*,1}(a|s)$, is given by $\pi^{*,1}(a|s) \propto \exp\left(\frac{1}{\lambda} E_Q^{*,1}(s,a)\right)$, where $E_Q^{*,1}(s,a)$ is the expected Q-value using the optimal value function $V(s)$:

$$E_Q^{*,1}(s,a) = R^{\nu^*}(s,a) + \gamma \sum_{s' \in \mathcal{S}} P^{\nu^*}(s'|s,a) V(s').$$

Substituting the expression for $R^{\nu^*}(s,a)$ from (11):

$$E_Q^{*,1}(s,a) = \left[\lambda \log(\mu^*(a \mid s)) + V(s) - \gamma \sum_{s' \in \mathcal{S}} P^{\nu^*}(s' \mid s,a) V(s') \right] + \gamma \sum_{s' \in \mathcal{S}} P^{\nu^*}(s'|s,a) V(s').$$

Rewriting this expression gives exactly the form of an optimal policy in the entropy regularized Markov Game. Since $\mu^*(a|s)$ is the optimal policy for agent 1 under the reward R (when agent 2 plays ν^*), the reward function R is feasible. This completes the proof. □

In the following, we want to investigate how an error in estimating the expert policy pair and the transition function translates to the recovered reward function. Note that compared to the single-agent case it is required to estimate the policy of opponent accurately as well.

The next lemma is of great importance, to derive the error propagation. It states how estimating the induced transition probability is related to the joint transition probability.

Lemma F.4. *Let $V : \mathcal{S} \to \mathbb{R}$ be a function. Furthermore, let P^ν be the induced transition probability and $\widehat{P}^\nu$ the estimated one of an underlying Markov Game with transition dynamic P. With out loss of generality, we assume that we are fixing the policy of agent 2, i.e. ν. Then it holds true, that*

$$\begin{aligned} &| \sum_{s'} P^\nu(s' \mid s,a) V(s') - \sum_{s'} \widehat{P}^\nu(s' \mid s,a) V(s') | \\ &\leq \max_b | \sum_{s'} V(s') \left(P(s' \mid s,a,b) - \hat{P}(s' \mid s,a,b) \right) | \end{aligned}$$

Proof.

$$
\begin{aligned}
&|\sum_{s'} P^{\nu}(s' \mid s,a)V(s) - \sum_{s'} \widehat{P}^{\nu}(s' \mid s,a)V(s')| \\
&= |\sum_{s'}\sum_{b\in\mathcal{B}} \nu(b \mid s)P(s' \mid s,a,b)V(s') - \sum_{s'}\sum_{b\in\mathcal{B}} \hat{\nu}(b \mid s)\widehat{P}(s' \mid s,a,b)V(s')| \\
&\leq \sum_{b\in\mathcal{B}} \max(\nu(b \mid s), \hat{\nu}(b \mid s))|\sum_{s'} P(s' \mid s,a,b)V(s') - \sum_{s'} \widehat{P}(s' \mid s,a,b)V(s')| \\
&\leq \max_b |\sum_{s'} P(s' \mid s,a,b)V(s') - \sum_{s'} \widehat{P}(s' \mid s,a,b)V(s')| \\
&= \max_b |\sum_{s'} V(s')\left(P(s' \mid s,a,b) - \widehat{P}(s' \mid s,a,b)\right)|
\end{aligned}
$$

□

With the introduced Lemma, we can now introduce an error propagation theorem. The idea is that we use the explicit reward function from Lemma F.3 and bound the individual terms of the true underlying MAIRL problem and the estimated one.

Theorem F.5 (Error propagation)**.** *Let the MAIRL problem be given by $(\mathcal{G}, (\mu^*, \nu^*))$ for a Markov Game and let $(\hat{\mathcal{G}}, (\hat{\mu}^*, \hat{\nu}^*)$ be another MAIRL problem. Then, we have that*

$$
\begin{aligned}
|R(s,a,b) - \hat{R}(s,a,b)| \leq \frac{2}{\nu^*(b \mid s)\hat{\nu}^*(b \mid s)}\Bigg(&\lambda|\log\mu^*(a \mid s) - \log\hat{\mu}^*(a \mid s)| \\
&+ \gamma \max_b |\sum_{s'} V(s')P(s' \mid s,a,b) - \hat{P}(s' \mid s,a,b)| + R_{\max}TV(\nu,\hat{\nu})\Bigg)
\end{aligned}
$$

Proof. In the first step, we use the derived explicit form of the reward derived in Lemma F.3.

$$
\begin{aligned}
\hat{R}(s,a,b) = \frac{1}{\hat{\nu}^*}(b \mid s)\Bigg(&\lambda\log(\hat{\mu}^*(a \mid s)) + \hat{V}(s) \\
&- \gamma\sum_{s}\sum_{b'} \hat{\nu}^*(b' \mid s)\hat{P}(s' \mid s,a,b')\hat{V}(s') - \sum_{b'} \nu^*(b' \mid s)\hat{R}(s,a,b')\Bigg)
\end{aligned}
$$

As pointed out in Metelli et al. [2023], the rewards $\hat{R}(s,a,b)$ do not have to be bounded by the same $R_{\max}$ as $R(s,a,b)$. To fix this issue the authors point out, that the reward needs to be rescaled such that the recovered feasible reward set is bounded by the same value. In our case we have to be a bit more careful with the choice of the scaling, as we did not assume that the reward is bounded by 1. As we proof the existence of such reward function, we can choose $\tilde{V}(s) = V(s)$ for every $s \in \mathcal{S}$ and $\tilde{R}(s,a,b') = R(s,a,b')\forall b' \neq b$ for every $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$, which results in rewards

$$
\begin{aligned}
&\tilde{R}(s,a,b) = \\
&\frac{\lambda\log(\hat{\mu}^*(a \mid s)) + V(s) - \gamma\sum_s\sum_{b'}\hat{\nu}^*(b \mid s)\hat{P}(s' \mid s,a,b')V(s') - \sum_{b'\neq b}\hat{\nu}^*(b' \mid s)R(s,a,b')}{\hat{\nu}^*(b \mid s)}.
\end{aligned}
$$

Now we need to rescale the reward with $R_{\max} + |\epsilon^i(s,a,b)|$ respectively,

$$
\begin{aligned}
&\epsilon(s,a,b) \\
&= \frac{\lambda\log(a \mid s) + V(s) - \gamma\sum_{s'}\sum_{b'\in\mathcal{B}}\nu^*(b' \mid s)P(s' \mid s,a,b')V(s') - \sum_{b'\neq b}\nu^*(b' \mid s)R(s,a,b')}{\nu^*(b \mid s)} \\
&\quad - \frac{\left(\lambda\log(\hat{\mu}^*(a \mid s)) + \hat{V}(s) - \gamma\sum_s\sum_{b'}\hat{\nu}^*(b' \mid s)\hat{P}(s' \mid s,a,b')\hat{V}(s') - \sum_{b'}\hat{\nu}^*(b' \mid s)\hat{R}(s,a,b')\right)}{\hat{\nu}^*(b \mid s)},
\end{aligned}
$$

such that it remains bounded by $R_{\max}$, we receive $\hat{R}(s,a,b) = \tilde{R}(s,a,b)\frac{R_{\max}}{R_{\max}+|\epsilon(s,a,b)|}$.

It then follows that:

$$
\begin{aligned}
|R(s,a,b)-\hat{R}(s,a,b)| &= |R(s,a,b)-\frac{R_{\max}\tilde{R}(s,a,b)}{R_{\max}+|\epsilon(s,a,b)|}| \\
&\leq \frac{R_{\max}}{R_{\max}+|\epsilon(s,a,b)|}\left|\left(\frac{R_{\max}+|\epsilon(s,a,b)|}{R_{\max}}\right)R(s,a,b)-\tilde{R}(s,a,b)\right| \\
&\leq \frac{R_{\max}}{R_{\max}+|\epsilon(s,a,b)|}\left(|R(s,a,b)-\tilde{R}(s,a,b)|+|\frac{\epsilon(s,a,b)}{R_{\max}}R(s,a,b)|\right) \\
&\leq \frac{R_{\max}}{R_{\max}+|\epsilon(s,a,b)|}\left(|\epsilon(s,a,b)|+|\epsilon(s,a,b)|\right) \\
&\leq \frac{R_{\max}}{R_{\max}}\left(|\epsilon(s,a,b)|+|\epsilon(s,a,b)|\right) \\
&= 2|\epsilon(s,a,b)|
\end{aligned}
$$

In the next step, we bound the $|\epsilon(s,a,b)|$

$$
\begin{aligned}
&|\epsilon(s,a,b)| \\
&= \frac{\lambda\log(a\mid s)+V(s)-\gamma\sum_{s'}\sum_{b'\in\mathcal{B}}\nu^*(b'\mid s)P(s'\mid s,a,b')V(s')-\sum_{b'\neq b}\nu^*(b'\mid s)R(s,a,b')}{\nu^*(b\mid s)} \\
&- \frac{\left(\lambda\log(\hat{\mu}^*(a\mid s))+\hat{V}(s)-\gamma\sum_{s}\sum_{b'}\hat{\nu}^*(b'\mid s)\hat{P}(s'\mid s,a,b')\hat{V}(s')-\sum_{b'}\hat{\nu}^*(b'\mid s)\hat{R}(s,a,b')\right)}{\hat{\nu}^*(b\mid s)} \\
&\leq \frac{1}{\nu^*(b\mid s)\hat{\nu}^*(b\mid s)}\Bigg(\lambda|\log\mu^*(a\mid s)-\log\hat{\mu}^*(a\mid s)| \\
&+\gamma\left|\sum_{s'}\sum_{b'}(\nu(b'\mid s)P(s'\mid s,a,b')-\hat{\nu}^*(b'\mid s)\hat{P}(s,a,b'))V(s')\right| \\
&+\left|\sum_{b'\neq b}R(s,a,b')(\nu^*(b'\mid s)-\hat{\nu}^*(b\mid s))\right|\Bigg) \\
&\overset{(i)}{\leq} \frac{1}{\nu^*(b\mid s)\hat{\nu}^*(b\mid s)}\Bigg(\lambda|\log\mu^*(a\mid s)-\log\hat{\mu}^*(a\mid s)| \\
&+\gamma\left|\max_{b'\in\mathcal{B}}|\sum_{s'}P(s'\mid s,a,b')-\hat{P}(s,a,b'))V(s')\right|+\left|\sum_{b'\neq b}R(s,a,b')(\nu^*(b'\mid s)-\hat{\nu}^*(b\mid s))\right|\Bigg) \\
&\overset{(ii)}{\leq} \frac{1}{\nu^*(b\mid s)\hat{\nu}^*(b\mid s)}\Bigg(\lambda|\log\mu^*(a\mid s)-\log\hat{\mu}^*(a\mid s)| \\
&+\gamma\left|\max_{b'\in\mathcal{B}}\sum_{s'}P(s'\mid s,a,b')-\hat{P}(s,a,b'))V(s')\right|+R_{\max}\sum_{b'\neq b}|(\nu^*(b'\mid s)-\hat{\nu}^*(b\mid s))|\Bigg) \\
&\leq \frac{1}{\nu^*(b\mid s)\hat{\nu}^*(b\mid s)}\Bigg(\lambda|\log\mu^*(a\mid s)-\log\hat{\mu}^*(a\mid s)| \\
&\quad+\gamma\left|\max_{b'\in\mathcal{B}}|\sum_{s'}P(s'\mid s,a,b')-\hat{P}(s,a,b'))V(s')\right|+R_{\max}\sum_{b'\neq b}\mathrm{TV}(\nu^*,\hat{\nu}^*)\Bigg),
\end{aligned}
$$

where we used Lemma F.4 for (i), then the assumption that $R(s,a,b)$ is bounded by $R_{\max}$ in (ii) and last, we added $|\nu^*(b\mid s)-\hat{\nu}^*(b\mid s)|$ to obtain the definition of the total variation with the triangle inequality. $\square$

We now again, want to use the empirical estimators to do a sample complexity analysis. The part of the transition probability can be obtained similar to the case of the pure NE, while for the policy we need to bound with high probability

$$|\log(\mu^*(a \mid s)) - \log(\hat{\mu}^*(a \mid s))|, \, |\log(\nu^*(a \mid s)) - \log(\hat{\nu}^*(a \mid s))|.$$

Let us first introduce the assumption, also common in single-agent IRL, that the lowest probability of an action taken from the expert is bounded away from zero by some constant.

Assumption F.6. Let μ^*, ν^* be the QRE equilibrium expert policies. Then we assume that

$$\min_{a \in \mathcal{A}, b \in \mathcal{B}} (\mu^*(a \mid s), \nu^*(b \mid s)) \geq \Delta_{\min}.$$

Now we are introducing the empirical estimators, used for recovering the MAIRL problem.

For both estimation tasks, the expert policies and the transition probability, we employ empirical estimators. For each iteration $k \in [K]$, let $n_k(s,a,b,s') = \sum_{t=1}^{k} \mathbf{1}_{(s_t,a_t,b_t,s_t')=(s,a,b,s')}$ denote the count of visits to the triplet $(s,a,b,s') \in \mathcal{S} \times (\mathcal{A} \times \mathcal{B}) \times \mathcal{S}$, and let $n_k(s,a,b) = \sum_{s' \in \mathcal{S}} n_k(s,a,b,s')$ denote the count of visits to the state-action pair (s,a). Additionally, we introduce $n_k(s,a) = \sum_{t=1}^{k} \mathbf{1}_{(s_t,a_t)=(s,a)}$ and $n_k(s,b) = \sum_{t=1}^{k} \mathbf{1}_{(s_t,b_t)=(s,b)}$ as the count of times action a and respectively b was sampled in state $s \in \mathcal{S}$ for each agent i, and $n_k(s) = \sum_{a \in \mathcal{A}} n_k(s,a)$ as the count of visits to state s for any agent.

It is important to note the distinction here: the count of actions must be done separately for each agent, whereas the count of state visits needs to be done for both of the agents.

The cumulative count of actions for each agent and the cumulative state visit count are given by:

$$N_k(s,a,b) = \sum_{j \in [k]} n_j(s,a,b) \quad N(s) = \sum_{j \in [k]} n_j(s).$$

After introducing the empirical counts, we can now state the empirical estimators for the transition model and the expert's policy:

$$\hat{P}_k(s' \mid s,a,b) = \begin{cases} \frac{N_k(s,a,b,s')}{N_k(s,a,b)} & \text{if} \quad N_k(s,a,b) > 0 \\ \frac{1}{S} & \text{otherwise} \end{cases} \tag{12}$$

$$\hat{\mu}_k(a \mid s) = \begin{cases} \frac{N_k(s,a)}{N_k(s)} & \text{if} \quad N_k(s) > 0 \\ \frac{1}{|\mathcal{A}|} & \text{otherwise.} \end{cases} \tag{13}$$

$$\hat{\nu}_k(b \mid s) = \begin{cases} \frac{N_k(s,b)}{N_k(s)} & \text{if} \quad N_k(s) > 0 \\ \frac{1}{|\mathcal{B}|} & \text{otherwise.} \end{cases} \tag{14}$$

Next, we state the lemma that derives the good event. Note that here we prove something stronger regarding the transition model, i.e. that the good event holds for all s,a,b, therefore also for $\max_{b \in \mathcal{B}}$.

Lemma F.7 (Good Event Regularized Games)**.** *Let k be the number of iterations and (μ^*, ν^*) be the QRE expert policies. Furthermore let $(\hat{\mu}, \hat{\nu})$ and $\hat{P}$ be the empirical estimates of the Nash expert and the transition probability after k iterations as defined in Eq.* (13) *and Eq.* (12) *respectively. Then for $\delta \in (0,1)$, define the good event $\mathcal{E}$ as the event such that the following inequalities hold* *simultaneously for all* $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ *and* $k \geq 1$*, which holds with probability at least* $1-\delta$*:*

$$\begin{aligned}
|\log(\mu(a \mid s)) - \log(\hat{\mu}(a \mid s))| &\leq \frac{1}{\Delta_{\min}} \sqrt{\frac{2\log(10|\mathcal{S}||\mathcal{A}|(N_k^+(s))^2/\delta)}{N_k^+(s)}}, \\
\sum_{s'} \left| (P(s' \mid s,a,b) - \hat{P}_k(s' \mid s,a,b)) V(s') \right| &\leq \frac{R_{\max}}{1-\gamma} \sqrt{\frac{8 l_k(s,a,b)}{N_k^+(s,a,b)}}, \\
\sum_{s'} \left| (P(s' \mid s,a,b) - \hat{P}_k(s' \mid s,a,b)) \hat{V}(s') \right| &\leq \frac{R_{\max}}{1-\gamma} \sqrt{\frac{8 l_k(s,a,b)}{N_k^+(s,a,b)}} \\
\sum_{b \in \mathcal{B}} |\nu(b \mid s) - \hat{\nu}(b \mid s)| &\leq \sqrt{\frac{2|\mathcal{B}| \log(10|\mathcal{S}||\mathcal{B}|(N_k^+(s))^2/\delta)}{N_k^+(s)}} \\
\frac{1}{\hat{\nu}(b \mid s)\nu(b \mid s)} &\leq \frac{1}{\Delta_{\min}^2} \sqrt{2\frac{\log(10|\mathcal{S}||\mathcal{B}|(N_k^+(s))^2/\delta)}{N_k^+(s)}}
\end{aligned}$$

where we introduced $l_k(s,a,b) := \log\left(\frac{30|\mathcal{S}||\mathcal{A}||\mathcal{B}|(N_k^+(s,a,b))^2}{\delta}\right)$.

Proof. We start the proof by defining the good event for the transition model, which proceeds in a similar way as already seen in Lemma D.8. We start with bound of the transition dynamics. Note that here we prove something stronger, that the good event holds for all s, a, b, therefore also for $\max_{b \in \mathcal{B}}$. Therefore we define $l_k(s,a,b) := \log\left(\frac{30|\mathcal{S}||\mathcal{A}||\mathcal{B}|(N_k^+(s,a,b))^2}{\delta}\right)$ and additionally we denote $\beta_{N_k(s,a,b)} = \frac{\gamma R_{\max}}{1-\gamma}\sqrt{\frac{2 l_k(s,a,b)}{N_k^+(s,a,b)}}$. Now we define the set

$$\mathcal{E}^{\text{trans}} := \left\{\forall k \in \mathbb{N}, \forall (s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} : \sum_{s'} \left| P(s' \mid s,a,b) - \hat{P}(s' \mid s,a,b) \right| V(s') \leq \beta_{N_k(s,a,b)} \right\}.$$

Then we get for V with probability of $1-\delta$:

$$\begin{aligned}
&\mathbb{P}\left((\mathcal{E}^{\text{trans}})^C\right) \\
&= \mathbb{P}\Big(\exists k \geq 1, \exists (s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} : \\
&\qquad \sum_{s'} \left| \left(P(s' \mid s,a,b) - \hat{P}_k(s' \mid s,a,b)\right) V(s') \right| > \beta_{N_k(s,a,b)}(s,a,b) \Big) \\
&\overset{(a)}{\leq} \sum_{(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}} \mathbb{P}\left(\exists k \geq 1 : \sum_{s'} \left| \left(P(s' \mid s,a,b) - \hat{P}_k(s' \mid s,a,b)\right) V(s') \right| > \beta_{N_k(s,a,b)} \right) \\
&\overset{(b)}{\leq} \sum_{(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}} \mathbb{P}\left(\exists m \geq 0 : \sum_{s'} \left| \left(P(s' \mid s,a,b) - \hat{P}_k(s' \mid s,a,b)\right) V(s') \right| > \beta_{N_k(s,a,b)} \right) \\
&\overset{(c)}{\leq} \sum_{(s,a,b) \in \mathcal{S} \times \mathcal{A} \mathcal{B}} \sum_{m \geq 0} 2 \exp\left(\frac{\beta_{N_k(s,a,b)}^2 m^2 (1-\gamma)^2}{4 m \gamma^2 R_{\max}^2}\right) \\
&\leq \sum_{(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}} \sum_{m \geq 0} \frac{\delta}{15|\mathcal{S}||\mathcal{A}||\mathcal{B}|(m^+)^2} \\
&\leq \frac{\delta}{15}\left(1 + \frac{\pi^2}{6}\right) \leq \frac{\delta}{5},
\end{aligned}$$

where (a) uses a union bound over the state and joint-action space, (b) uses that we only consider the m-times, where we updated the estimated transition model and (c) uses an union bound over the update times m and an application of Hoeffding's inequality combined with the fact that we can bound the value function, i.e. $V^i(s') \leq \frac{\gamma R_{\max}}{1-\gamma}$ for every $s' \in \mathcal{S}$. Next, we will consider the first equation regarding estimating the log probability of the expert policy. In a first step we define $\beta_2(s) := \frac{1}{\Delta_{\min}}\sqrt{\frac{\log(10|\mathcal{S}||\mathcal{A}|(N_k^+(s))^2/\delta)}{2N_k^+(s)}}$ the following set

$$\mathcal{E}^{\log} := \{\forall k \in \mathbb{N},\, \forall (s,a) \in \mathcal{S}\times\mathcal{A} : |\log(\mu(a \mid s)) - \log(\hat{\mu}(a \mid s))| \leq \beta_2(s)\}\,.$$

$$\begin{aligned}
\mathbb{P}\left((\mathcal{E}^{\log})^C\right) &= \mathbb{P}\left(\exists k \geq 1, \exists (s,a) \in \mathcal{S}\times\mathcal{A} : |\log(\mu(a \mid s)) - \log(\hat{\mu}_k(a \mid s))| > \beta_2(s)\right) \\
&\overset{(a)}{\leq} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbb{P}\left(\exists k \geq 1 : |\log(\mu(a \mid s)) - \log(\hat{\mu}_k(a \mid s))| > \beta_2(s)\right) \\
&\overset{(b)}{\leq} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbb{P}\left(\exists m \geq 0 : |\log(\mu(a \mid s)) - \log(\hat{\mu}_m(a \mid s))| > \beta_2(s)\right) \\
&\overset{(c)}{\leq} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{m\geq 0} \frac{\delta}{10|\mathcal{S}||\mathcal{A}|m^2} \\
&\leq \frac{\delta}{10}\left(1 + \frac{\pi^2}{6}\right) \leq \frac{\delta}{5},
\end{aligned}$$

where (a) uses a union bound over the state and action space of player 1, (b) uses that we only consider the m-times, where we updated the estimated transition model and (c) we can applied Lemma I.3. To be precise, (c) only holds if N is large enough, however, we will late only consider this case, therefore we use it directly.

For the second last step we define the good event for the total variation

$$\mathcal{E}^{\mathrm{TV}} := \left\{\forall k \in \mathbb{N},\, \forall (s,b) \in \mathcal{S}\times\mathcal{B} : \mathrm{TV}(\nu, \hat{\nu}) \leq \sqrt{\frac{|\mathcal{B}|\log(10|\mathcal{S}||\mathcal{B}|(N_k^+(s))^2/\delta)}{N_k^+(s)}}\right\}.$$

We will the bound the probability of the complement of this event by δ and can then take the intersection of both events to get the total result. We will skip some intermediate steps, as they are similar to the ones obtained above.

$$\begin{aligned}
\mathbb{P}((\mathcal{E}^{\mathrm{TV}})^C) &= \mathbb{P}\left(\exists k \in \mathbb{N},\, \exists (s,b) \in \mathcal{S}\times\mathcal{B} : \mathrm{TV}(\nu,\hat{\nu}) > \sqrt{\frac{|\mathcal{B}|\log(10|\mathcal{S}||\mathcal{B}|(N_k^+(s))^2/\delta)}{N_k^+(s)}}\right) \\
&\overset{(a'')}{\leq} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{m\geq 0} \mathbb{P}\left(\mathrm{TV}(\nu,\hat{\nu}) > \sqrt{\frac{|\mathcal{B}|\log(10|\mathcal{S}||\mathcal{B}|m^2/\delta)}{m}}\right) \\
&\overset{(b'')}{\leq} \frac{\delta}{5},
\end{aligned}$$

where (a") uses a union bound over the state and action space, (b") uses Lemma I.5. Also here to be precise, (b'') only holds if N is large enough, however, we will late only consider this case, therefore we use it directly. To bound the second last event, we can apply Lemma I.4. To complete this proof, we first define $\beta_3(s) := \frac{1}{\Delta_{\min}^2}\sqrt{\frac{2\log(10|\mathcal{S}||\mathcal{B}|(N_k^+(s))^2/\delta)}{N_k^+(s)}}$ and once again apply the same argument for the good event,

$$\mathcal{E}^{\mathrm{invprop}} := \left\{\forall k \in \mathbb{N},\, \forall (s,b) \in \mathcal{S}\times\mathcal{B} : \frac{1}{\nu(b \mid s)\hat{\nu}(b \mid s)} \leq \beta_3(s)\right\},$$

now combined with Lemma I.4. As all the good events holds with probability $\delta/5$, we get that

$$\mathbb{P}\left(\mathcal{E}^{\mathrm{TV}} \cap \mathcal{E}^{\log} \cap \mathcal{E}^{trans} \cap \mathcal{E}^{\mathrm{invprop}}\right) > 1 - \delta.$$

$\square$

Following, we present the reward uncertainty metric, which allows us to demonstrate that the difference between the recovered reward function and the true reward function is bounded.

Definition F.8 (Reward Uncertainty)**.** Let k be the number of iterations. Then the reward uncertainty after k iterations for any $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ is defined as

$$C_k(s,a,b) := \frac{4\gamma R_{\max}}{(1-\gamma)\Delta_{\min}^2}\left(\sqrt{\frac{8|\mathcal{B}|\log(10|\mathcal{S}||\mathcal{B}|(N_k^+(s))^2/\delta)}{N_k^+(s)}}\right.$$
$$\left. + \sqrt{\frac{2\log(10|\mathcal{S}||\mathcal{A}|(N_k^+(s))^2/\delta)}{N_k^+(s)}} + \sqrt{\frac{2 l_k(s,a,b)}{N_k^+(s,a,b)}}\right).$$

Theorem F.9. *Let the reward uncertainty be defined as in D.9. Under the good event it holds for any $(s,a,b) \in \mathcal{S} \times \mathcal{A}$ that:*

$$|R^i(s,a,b) - \hat{R}^i(s,a,b)| \leq C_k(s,a,b).$$

Proof. The theorem is an application of the error propagation Theorem F.5, followed by Lemma F.7

$$\begin{aligned}
&|R^i(s,a,b) - \hat{R}^i(s,a,b)| \\
&\leq \frac{2}{\nu^*(b \mid s)\hat{\nu}^*(b \mid s)}\left(\lambda|\log\mu^*(a \mid s) - \log\hat{\mu}^*(a \mid s)|\right. \\
&\quad \left. +\gamma \max_b |\sum_{s'} V(s')P(s' \mid s,a,b) - \hat{P}(s' \mid s,a,b)| + R_{\max} TV(\mu,\hat{\mu})\right) \\
&\leq \frac{4R_{\max}}{1-\gamma}\left(\frac{1}{\Delta_{\min}^2}\sqrt{2\frac{\log(10|\mathcal{S}||\mathcal{B}|(N_k^+(s))^2/\delta)}{N_k^+(s)}} + \sqrt{\frac{2|\mathcal{B}|\log(10|\mathcal{S}||\mathcal{B}|(N_k^+(s))^2/\delta)}{N_k^+(s)}}\right. \\
&\quad \left. + \frac{1}{\Delta_{\min}}\sqrt{\frac{2\log(10|\mathcal{S}||\mathcal{A}|(N_k^+(s))^2/\delta)}{N_k^+(s)}} + \gamma\sqrt{\frac{2l_k(s,a,b)}{N_k^+(s,a,b)}}\right) \\
&\leq \frac{4\gamma R_{\max}}{(1-\gamma)\Delta_{\min}^2}\left(\sqrt{\frac{8|\mathcal{B}|\log(10|\mathcal{S}||\mathcal{B}|(N_k^+(s))^2/\delta)}{N_k^+(s)}} + \sqrt{\frac{2\log(10|\mathcal{S}||\mathcal{A}|(N_k^+(s))^2/\delta)}{N_k^+(s)}}\right. \\
&\quad \left. + \sqrt{\frac{2l_k(s,a,b)}{N_k^+(s,a,b)}}\right) \\
&= C_k(s,a,b).
\end{aligned}$$

□

Before stating the correctness of the algorithm, we want to mention that the derivations of Lemma D.6 also hold for the regularized case. Therefore, we can continue with the the correctness result for the regularized case.

Corollary F.10. *Let k be the number of iterations for any allocation of the samples over the state-action space $\mathcal{S} \times \mathcal{A}$. Furthermore, let $\mathcal{R}_{(\mathcal{G},\boldsymbol{\pi}^{\text{Nash}})}$ be the true feasible set and $\mathcal{R}_{(\hat{\mathcal{G}},\hat{\boldsymbol{\pi}}^{\text{Nash}})}$ the recovered one. Then the optimality criterion 3.3 holds true, if*

$$\frac{1}{1-\gamma}\max_{(s,a,b)} C_k(s,a,b) \leq \frac{\epsilon}{2}.$$

Proof. We complete the proof for the first case of the optimality criterion, the second one follows analogously.

$$\begin{aligned}
&\sup_{R \in \mathcal{R}_{(\mathcal{G},\boldsymbol{\pi}^{\text{Nash}})}} \inf_{\hat{R} \in \mathcal{R}_{(\hat{\mathcal{G}},\hat{\boldsymbol{\pi}}^{\text{Nash}})}} \max\{\max_\mu V_\lambda^1(\mu,\hat{\nu}) - V_\lambda^1(\hat{\mu},\hat{\nu}), \max_\nu V_\lambda^1(\hat{\mu},\nu) - V_\lambda^1(\hat{\mu},\hat{\nu})\} \\
&\leq \frac{2}{1-\gamma}\max_{(s,a,b)} C_k(s,a,b) \leq \epsilon,
\end{aligned}$$

where we applied D.6 and we used the fact that $a+b \leq 2\max\{a,b\}$ followed by the error propagation Theorem F.5 and then we used D.10. □

We can combine the derived results to now state the main theorem regarding the sample complexity of the Uniform Sampling

Theorem F.11. *Let Assumption 3.8 hold true. Then, allocating the samples uniformly over $\mathcal{S}\times\mathcal{A}\times\mathcal{B}$ and using the empirical estimators introduced in Eq.* (13) *and Eq.* (12)*, we can stop the sampling procedure with a probability of at least $1-\delta$ after iteration τ and satisfy the optimality criterion Definition F.2, where the sample complexity is of order*

$$\tilde{\mathcal{O}}\left(\frac{\gamma^2 R_{\max}^2 |\mathcal{S}||\mathcal{A}||\mathcal{B}|}{(1-\gamma)^4\epsilon^2\Delta_{\min}^4}\right)$$

Proof. We know from D.11, that we need

$$\frac{1}{1-\gamma}\max_{(s,a,b)} C_k(s,a,b) \leq \frac{\epsilon_k}{2}$$

By the definition of $C_k(s,a,b)$ this is satisfied if

$$\frac{4\gamma R_{\max}}{(1-\gamma)^2\Delta_{\min}^2}\sqrt{\frac{2l_k(s,a,b)}{N_k^+(s,a,b)}} \leq \frac{\epsilon_k}{6}$$

$$\frac{4\gamma R_{\max}}{(1-\gamma)^2\Delta_{\min}^2}\sqrt{\frac{\log(10|\mathcal{S}||\mathcal{A}|(N_k^+(s))^2/\delta)}{N_k^+(s)}} \leq \frac{\epsilon_k}{6},$$

$$\frac{4\gamma R_{\max}}{(1-\gamma)^2\Delta_{\min}^2}\sqrt{\frac{8|\mathcal{B}|\log(|\mathcal{S}||\mathcal{B}|(N_k^+(s))^2/\delta)}{N_k^+(s)}} \leq \frac{\epsilon_k}{6}.$$

To achieve the first condition, we get

$$\begin{aligned}N_k(s,a,b) &\geq \frac{R_{\max}^2}{(1-\gamma)^4\Delta_{\min}^4}\gamma^2 24^2 l_k(s,a,b)\frac{2}{\epsilon_k^2}\\ &= \frac{\gamma^2 1152 R_{\max}^2}{(1-\gamma)^4\Delta_{\min}^4\epsilon_k^2}\log\left(\frac{12|\mathcal{S}|\mathcal{B}||\mathcal{A}|(N_k^+(s,a,b))^2}{\delta}\right)\end{aligned}$$

Applying Lemma B.8 by Metelli et al. [2021] we get that

$$N_k(s,a,b) \leq \frac{4608\gamma^2 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2\Delta_{\min}^4}\log\left(\frac{2304\gamma^2 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2}\sqrt{\frac{12|\mathcal{S}\mathcal{B}||\mathcal{A}|}{\delta}}\right).$$

At each iteration we are allocating the samples uniformly over $\mathcal{S}\times\mathcal{A}\times\mathcal{B}$ and recalling that $\tau_{s,a,b} = |\mathcal{S}||\mathcal{B}||\mathcal{A}|N_k(s,a,b)$ therefore we get

$$\tau_{s,a,b} \leq \frac{4608\gamma^2|\mathcal{S}|\mathcal{B}||\mathcal{A}|R_{\max}^2}{(1-\gamma)^4\Delta_{\min}^4\epsilon_k^2}\log\left(\frac{2304\gamma^2 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2\Delta_{\min}^4}\sqrt{\frac{12|\mathcal{S}||\mathcal{B}||\mathcal{A}|}{\delta}}\right)$$

Now we have to achieve the second condition,

$$\begin{aligned}N_k(s) &\geq \frac{24^2\gamma^2 R_{\max}^2}{(1-\gamma)^4\Delta_{\min}^4}\frac{\log(10|\mathcal{S}||\mathcal{A}|(N_k^+(s))^2/\delta)}{\epsilon_k^2}\\ &= \frac{\gamma^2 576 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2\Delta_{\min}^4}\log(10|\mathcal{S}||\mathcal{A}|(N_k^+(s))^2/\delta)\\ &= \frac{\gamma^2 576 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2\Delta_{\min}^4}\left(\log(10|\mathcal{S}||\mathcal{A}|/\delta) + 2\log((N_k^+(s)))\right)\end{aligned}$$

If we additionally force that $N_k(s) \geq 1$ and apply Lemma 15 of Kaufmann et al. [2021] with $1/\Delta^2 = \frac{\gamma^2 576 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2\Delta_{\min}^4}$, $a = \log(10|\mathcal{S}||\mathcal{A}|/\delta), b = 2, c = 0, d = 1$, we get that

$$\begin{aligned}N_k(s) \leq 1 + \frac{\gamma^2 576 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2\Delta_{\min}^4}\Bigg(&\log(10|\mathcal{S}||\mathcal{A}|/\delta)\\ &+ 2\log\left(\left(\frac{\gamma^2 576 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2\Delta_{\min}^4}\right)^2(\log(10|\mathcal{S}||\mathcal{A}|/\delta)+2)\right)\Bigg)\end{aligned}$$

As we here need to allocate samples uniformly over the state space only but for every agent separately and recalling that $\tau_{s_1} = SN_k(s)$, we get

$$\tau_{s_1} \leq |\mathcal{S}| + \frac{|\mathcal{S}|\gamma^2 576 R_{\max}^2}{(1-\gamma)^4 \epsilon_k^2 \Delta_{\min}^4}\Bigg(\log(10|\mathcal{S}||\mathcal{A}|/\delta) + 2\log\left(\left(\frac{\gamma^2 576 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2\Delta_{\min}^4}\right)^2 (\log(10|\mathcal{S}||\mathcal{A}|/\delta)+2)\right)\Bigg).$$

Lastly, we can proceed in a similar way compared to the last step,

$$\begin{aligned} N_k(s) &\geq \frac{R_{\max}^2}{(1-\gamma)^4\Delta_{\min}^4}\gamma^2 24^2 \frac{8|\mathcal{B}|\log(10|\mathcal{S}||\mathcal{B}|(N_k^+(s))^2/\delta)}{\epsilon_k^2} \\ &= \frac{\gamma^2 4608 R_{\max}^2|\mathcal{B}|}{(1-\gamma)^4\epsilon_k^2\Delta_{\min}^4}\log(10|\mathcal{S}||\mathcal{B}|(N_k^+(s))^2/\delta) \end{aligned}$$

If we additionally force that $N_k(s) \geq 1$ and again apply Lemma 15 of Kaufmann et al. [2021] with $1/\Delta^2 = \frac{|\mathcal{B}|\gamma^2 4608 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2\Delta_{\min}^4}, a = \log(10|\mathcal{S}||\mathcal{A}|/\delta), b = 2, c = 0, d = 1$ we get that

$$N_k(s) \leq 1 + \frac{|\mathcal{B}|\gamma^2 4608 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2\Delta_{\min}^4}\Bigg(\log(10|\mathcal{S}||\mathcal{B}|/\delta) + 2\log\left(\left(\frac{|\mathcal{B}|\gamma^2 4608 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2\Delta_{\min}^4}\right)^2 (\log(10|\mathcal{S}||\mathcal{B}|/\delta)+2)\right)\Bigg)$$

As we here need to allocate samples uniformly over the state space only but for every agent separately and recalling again that $\tau_{s_2} = SN_k(s)$, we get

$$\tau_{s_2} \leq |\mathcal{S}| + \frac{|\mathcal{S}||\mathcal{B}|\gamma^2 4608 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2\Delta_{\min}^4}\Bigg(\log(10|\mathcal{S}||\mathcal{B}|/\delta) + 2\log\left(\left(\frac{|\mathcal{B}|\gamma^2 4608 R_{\max}^2}{(1-\gamma)^4\epsilon_k^2\Delta_{\min}^4}\right)^2 (\log(10|\mathcal{S}||\mathcal{B}|/\delta)+2)\right)\Bigg).$$

Finally, as $\tau = \tau_{s,a} + \tau_{s_1} + \tau_{s_2}$ we get that this is exactly of order

$$\tilde{\mathcal{O}}\left(\frac{\gamma^2 R_{\max}^2|\mathcal{S}||\mathcal{A}||\mathcal{B}|}{(1-\gamma)^4\epsilon^2\Delta_{\min}^4}\right)$$

□

One can see that the problem scales with $|\mathcal{A}||\mathcal{B}|$, which is due to the used union bound. Scaling this up to n-player games the union bound implies that the bound will scale exponentially in the number of players. Additionally, as we have to bound the inverse probability the term $\Delta_{\min}$ appears on the bound. A sufficiently large $\Delta_{\min}$ can e.g. be obtained if the regularization parameter λ is small.

G Identifiability in Multi-agent Games?

This appendix provides supplementary details and proofs for the identifiability results discussed in Section 4. We follow the notation established in the main text, where applicable, using $R^1(s,a,b)$ to denote Player 1's reward, $R^\nu(s,a) = \sum_{b' \in \mathcal{B}} \nu(b' \mid s) R^1(s,a,b')$ for the average reward received by Player 1 when Player 2 uses policy ν, and P_a^ν for the induced transition matrix for Player 1 under ν. λ_1 denotes Player 1's entropy regularization parameter.

Average Identifiability. We start with the case, where we try to identify the average reward function (up to constants) for any player. The theorem is a direct consequence of Theorem 3 by Rolland et al. [2022].

Theorem G.1. *Let a Markov Game be given with two different opponents ν_1, ν_2 that induce different dynamics P^{ν_1}, P^{ν_2} and discount factors γ_1, γ_2. Suppose that in both Games we observe QRE equilibrium policy pairs (μ_1, ν_1) and a different ν_2 with a best responding policy μ_2 such that they have same average reward functions $R^{\nu_1} = R^{\nu_2}$. Additionally, define $P_a^{\nu_i} \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}}$ the induced transition matrix of expert $i \in \{1,2\}$. Then, the* average *reward player 1 receives can be recovered up to a constant if and only if*

$$\textit{rank}\begin{pmatrix} I-\gamma_1 P_{a_1}^{\nu_1} & I-\gamma_2 P_{a_1}^{\nu_2} \\ \vdots & \vdots \\ I-\gamma_1 P_{a_{|\mathcal{A}|}}^{\nu_1} & I-\gamma_2 P_{a_{|\mathcal{A}|}}^{\nu_2} \end{pmatrix} = 2|\mathcal{S}|-1.$$

Analogously this holds for player 2.

Theorem G.2 (Sample Complexity for Induced Transitions)**.** *To estimate the induced transition model P^{ν^*} for Player 1 (where ν^* is Player 2's true policy) such that the maximum L_1 error over all (s,a) rows is bounded by ϵ, i.e., $\max_{s,a}\|P^{\nu^*}(\cdot\mid s,a) - \widehat{P}^{\hat{\nu}}(\cdot\mid s,a)\|_1 \leq \epsilon$, with probability at least $1-\delta$, the total number of samples N is of the order:*

$$\mathcal{O}\left(\frac{|\mathcal{S}|^2|\mathcal{A}||\mathcal{B}|\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|/\delta)}{\epsilon^2}\right)$$

where $\widehat{P}^{\hat{\nu}}(s'|s,a) = \sum_{b\in\mathcal{B}}\hat{\nu}(b|s)\hat{P}(s'|s,a,b)$ uses empirical estimates $\hat{\nu}$ of ν^* and $\hat{P}$ of the true dynamics P.

Proof. The estimated induced transition is $\widehat{P}^{\hat{\nu}}(\cdot|s,a) = \sum_b \hat{\nu}(b|s)\hat{P}(\cdot|s,a,b)$. The true induced transition is $P^{\nu^*}(\cdot|s,a) = \sum_b \nu^*(b|s)P(\cdot|s,a,b)$. We want to bound the L_1 error. We use the triangle inequality and properties of the L_1 norm:

$$\begin{aligned} &\|P^{\nu^*}(\cdot|s,a) - \widehat{P}^{\hat{\nu}}(\cdot|s,a)\|_1 \\ &= \left\|\sum_b \nu^*(b|s)P(\cdot|s,a,b) - \sum_b \hat{\nu}(b|s)\hat{P}(\cdot|s,a,b)\right\|_1 \\ &= \left\|\sum_b(\nu^*(b|s)-\hat{\nu}(b|s))P(\cdot|s,a,b) + \sum_b \hat{\nu}(b|s)(P(\cdot|s,a,b)-\hat{P}(\cdot|s,a,b))\right\|_1 \\ &\leq \sum_b |\nu^*(b|s)-\hat{\nu}(b|s)|\cdot\|P(\cdot|s,a,b)\|_1 + \sum_b \hat{\nu}(b|s)\|P(\cdot|s,a,b)-\hat{P}(\cdot|s,a,b)\|_1 \end{aligned}$$

Since $\|P(\cdot|s,a,b)\|_1 = 1$ and $\sum_b \hat{\nu}(b|s) = 1$:

$$\|P^{\nu^*}(\cdot|s,a) - \widehat{P}^{\hat{\nu}}(\cdot|s,a)\|_1 \leq \|\nu^*(\cdot|s) - \hat{\nu}(\cdot|s)\|_1 + \max_{b'\in\mathcal{B}}\|P(\cdot|s,a,b') - \hat{P}(\cdot|s,a,b')\|_1$$

To ensure $\|P^{\nu^*}(\cdot|s,a) - \widehat{P}^{\hat{\nu}}(\cdot|s,a)\|_1 \leq \epsilon$ with high probability, we need to ensure both terms on the right are sufficiently small, e.g., $\leq \epsilon/2$.

Estimating the multinomial distribution $\nu^*(\cdot|s)$ over $|\mathcal{B}|$ actions requires $N_\nu(s)$ samples of Player 2's actions at state s. Applying Lemma I.5 gives us that to achieve $\|\nu^*(\cdot|s) - \hat{\nu}(\cdot|s)\|_1 \leq \epsilon/2$ with probability $1-\delta'$, requires that $N_\nu(s)$ is of the order $\mathcal{O}\left(\frac{|\mathcal{B}|}{\epsilon^2}\right)$ samples.

Similarly, we can bound the transition dynamics. In particular, if we apply Lemma I.5, then we get that the amount of samples required to minimize it with high probability is of the order $\mathcal{O}\left(\frac{|\mathcal{S}||\mathcal{A}||\mathcal{B}|}{\epsilon^2}\right)$. As the part of the transition dynamics dominates, the total number of samples is then of the order is then given by

$$\mathcal{O}\left(\frac{|\mathcal{S}||\mathcal{A}||\mathcal{B}|}{\epsilon^2}\right).$$

□

Reward identification in Linear separable Markov Games. As the so far discussed theorems only work for the average reward case, we want to identify conditions that allow us to identify the rewards in the multi-agent case. As discussed in Section 4 one potential solution to achieve this is to assume linear separable rewards that naturally disentangle the rewards into a part that only depends on action a and a part that only depends on action b.

We can immediately see that $R^{\nu^*}(s,a) = \sum_{b\in\mathcal{B}} \nu^*(b \mid s) R(s,a,b) = R_A(s,a) + \sum_{b\in\mathcal{B}} \nu^*(b \mid s) R_B(s,b)$. This means, that the average reward equation Eq. (2), can be rewritten. In particular, we get for every $(s,a)\in\mathcal{S}\times\mathcal{A}$

$$R_A(s,a) = \lambda \log(\mu^*(a \mid s)) + V(s) - \gamma \sum_{s'} P^{\nu^*}(s' \mid, s, a) V(s') - \sum_{b\in\mathcal{B}} \nu^*(b \mid s) R_B(s,b).$$

Next, we again consider the case, where we have two Markov Games where we have the same reward function for player 1, in particular for action a. Using the explicit formulation of the reward, we get the following:

$$\begin{aligned} R_A(s,a) &= \lambda \log(\mu_1^*(a \mid s)) + V_1(s) - \gamma \sum_{s'} P_1^{\nu_1^*}(s' \mid, s, a) V_1(s') - \sum_{b\in\mathcal{B}} \nu_1^*(b \mid s) R_B(s,b) \\ &= \lambda \log(\mu_2^*(a \mid s)) + V_2(s) - \gamma \sum_{s'} P_2^{\nu_2^*}(s' \mid, s, a) V_2(s') - \sum_{b\in\mathcal{B}} \nu_2^*(b \mid s) R_B(s,b) \end{aligned}$$

Let us consider two cases. The first case is that in both environments the opponent policy is the same, meaning that we have $\nu_1^* = \nu_2^*$. Then, we see immediately that $\sum_{b\in\mathcal{B}} \nu^*(b \mid s) R_B(s,b)$ cancels out and we get

$$\lambda \log(\mu_1^*(a \mid s)) + V_1(s) - \gamma \textstyle\sum_{s'} P_1^{\nu^*}(s' \mid s, a) V_1(s') = \lambda \log(\mu_2^*(a \mid s)) + V_2(s) - \gamma \textstyle\sum_{s'} P_2^{\nu^*}(s' \mid s, a) V_2(s').$$

Therefore, we reconstruct the single-agent case, as we obtain for every $(s,a)\in\mathcal{S}\times\mathcal{A}$:

$$V_1(s) - V_2(s) - \gamma \textstyle\sum_{s'} P_1^{\nu^*}(s' \mid, s, a) V_1(s') + \gamma \textstyle\sum_{s'} P_2^{\nu^*}(s' \mid, s, a) V_2(s') = \lambda(\log(\mu_2^*(a \mid s)) - \log(\mu_1^*(a \mid s))).$$

Therefore, we can again write this as a system of equations but this time for the same ν^*. We can summarize these findings in the following result.

Proposition G.3. *Let a Markov Game with a QRE equilibrium policy pair (μ_1^*, ν_1^*) be given. Additionally, let another Markov Game with the same ν_1^* but a different transition Model P_2 and discount factor γ_2 and therefore also different best response μ_2^* be given such that R_A^1 is the same for both environments. Then, R_A^1 is identifiable if and only if*

$$\mathit{rank} \begin{pmatrix} I - \gamma_1 P_{1,a_1}^{\nu_1^*} & I - \gamma_2 P_{2,a_1}^{\nu_1^*} \\ \vdots & \vdots \\ I - \gamma_1 P_{1,a_{|\mathcal{A}|}}^{\nu_1} & I - \gamma_2 P_{2,a_{|\mathcal{A}|}}^{\nu_1^*} \end{pmatrix} = 2|\mathcal{S}| - 1.$$

Analogously this holds for player 2.

For the second case, we assume that the opponent policies are different in the two Markov Games, meaning that $\nu_1^* \neq \nu_2^*$. In this case we need an additional assumption, namely $R_B(s, b_0) = 0$. his constraint fixes the baseline for $R_B(s,b)$, allowing us to determine how much player 2's different actions contribute to player 1's reward, relative to this baseline.

$$\begin{aligned} V_1(s) - V_2(s) - \gamma \sum_{s'} P_1^{\nu_1^*}(s' \mid, s, a) V_1(s') + \gamma \sum_{s'} P_2^{\nu_2^*}(s' \mid, s, a) V_2(s') \\ + \sum_{b\neq b_0} R_B(s,b)(\nu_2^*(b \mid s) - \nu_1^*(b \mid s)) = \lambda \log(\mu_2^*(a \mid s)) - \lambda \log(\mu_1^*(a \mid s)). \end{aligned}$$

This we can now again express as a system of equations as the above holds for every $(s,a)\in\mathcal{S}\times\mathcal{A}_1$ and we get:

$$\begin{pmatrix} M_{R_B} & I - \gamma_1 P_{a_1}^{\nu_1} & I - \gamma_2 P_{a_1}^{\nu_2} \\ \vdots & \vdots & \vdots \\ M_{R_B} & I - \gamma_1 P_{a_{|\mathcal{A}|}}^{\nu_1} & I - \gamma_2 P_{a_{|\mathcal{A}|}}^{\nu_2} \end{pmatrix} \begin{pmatrix} R_B \\ V_1 \\ V_2 \end{pmatrix} = \begin{pmatrix} \lambda(\log(\mu_2(\cdot \mid a_1)) - \log(\mu_1(\cdot \mid a_1))) \\ \vdots \\ \lambda(\log(\mu_2(\cdot \mid a_{|\mathcal{A}|})) - \log(\mu_1(\cdot \mid a_{|\mathcal{A}|}))) \end{pmatrix},$$

where R_B is the reward vector for every $s \in \mathcal{S}$ and $b \in \mathcal{B} \setminus \{b_0\}$ and each M_{R_B} is an $|\mathcal{S}| \times |\mathcal{S}||\mathcal{B}|$ block diagonal matrix with the following structure

$$M_{R_B} = \begin{pmatrix} \boldsymbol{\nu}(s_1) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\nu}(s_2) & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\nu}(s_{|\mathcal{S}|}) \end{pmatrix},$$

where each $\boldsymbol{\nu}(s_i)$ for $s_i \in \mathcal{S}$ is a $1 \times (|\mathcal{B}| - 1)$ row vector where every element is the difference of the policies for the actions $\nu_2(b \mid s) - \nu_1(b \mid s)$ for $b \in \mathcal{B} \setminus \{b_0\}$.

This means that the matrix on the left has shape $|\mathcal{A}||\mathcal{S}| \times |\mathcal{S}|(|\mathcal{B}| + 1)$. As we require identification up to constants we need that the matrix has rank $|\mathcal{S}|(|\mathcal{B}| + 1) - 1$. We can conclude our findings in the following proposition.

Proposition G.4. *Let us assume that we have two-player general-sum Markov Games with different transition functions P_1, P_2 and γ_1, γ_2 be given. Additionally, assume that the rewards for player 1 are the same under the QRE (μ_1^*, ν_1^*) and the other observed best responding policy μ_2^* to ν_2^*. Additionally, suppose player 1's reward function is linearly separable: $R_1(s, a, b) = R_A(s, a) + R_B(s, b)$. To ensure unique decomposition between R_A and R_B, the normalization $R_B(s, b_0) = 0$ is imposed for some fixed $b_0 \in \mathcal{B}$ and for all $s \in \mathcal{S}$. Then, player 1's reward function $R_1(s, a, b)$ (and its components $R_A(s, a)$ and $R_B(s, b)$ under the given normalization) can be recovered up to a single additive constant if and only if:*

$$\operatorname{rank} \begin{pmatrix} M_{R_B} & I - \gamma_1 P_{a_1}^{\nu_1} & I - \gamma_2 P_{a_1}^{\nu_2} \\ \vdots & \vdots & \vdots \\ M_{R_B} & I - \gamma_1 P_{a_{|\mathcal{A}|}}^{\nu_1} & I - \gamma_2 P_{a_{|\mathcal{A}|}}^{\nu_2} \end{pmatrix} = |\mathcal{S}|(|\mathcal{B}| + 1) - 1.$$

H Experimental Evaluation

In this section, we first give the details for the figures in Fig. 2 and Fig. 3. Then, we aim to demonstrate the advantages of IRL in the multi-agent setting compared to Behavior Cloning.

It is important to emphasize that the primary goal of this paper is to address the IRL problem from a theoretical perspective by defining a new objective and presenting the first algorithm to characterize the feasible set of rewards. In particular, we will demonstrate the case of pure NE strategies in a simple environment. Although, we have shown that in the general case one needs to observe multiple equilibria to infer a meaningful reward function, we will demonstrate that one can still obtain a good performance in simpler environments. We motivate this simple example given that calculating the NE is PPAD-hard. The idea is to emphasize the need for MAIRL framework and motivate future research on computationally more feasible equilibrium solutions.

H.1 Numerical verifications for Nash and QRE equilibrium observations.

In this section, we give the details for the numerical examples provided in Fig. 2 and Fig. 3 respectively. The idea of both experiments is the same. The considered environment is the one used in Proposition 3.4 and illustrated in Fig. 2.

Expert observations. In the case of Nash equilibrium experts, we can simply take any Nash equilibrium solver to calculate the equilibrium for state s_0. This holds true because in the following states the Nash equilibrium actions of the Normal Form Game in s_0 are rewarded (s_1 and s_3), while the other actions are not s_2. Regarding the equilibrium observations for the QRE, we again only consider state s_0 and run a simple algorithm that iteratively computes the expected reward of a player keeping the other players strategy fixed and then updates the policy. This is repeated until the strategies of both players are not changing anymore.

Calculating new equilibria and exploitability. For both expert observations, we then use any IRL algorithm that picks a reward function, such that the observed equilibrium is feasible under this reward. Then, we compute again the new equlbria and compare the list of original Nash equilibria with the ones under the recovered reward function. From this list, we then randomly select an equilibrium and calculate the exploitability of the picked strategy in the original Markov Game. This we repeat for 10000 iterations and compute the average exploitability and the average correlations.

H.2 MAIRL vs. Behavior Cloning.

In this section, we empirically evaluate the benefits of MAIRL compared to BC and describe the details on the environments in the following paragraphs. One of the advantages of IRL over BC lies in its ability to transfer the recovered reward function to new environments with different transition probabilities. This is particularly significant in a multi-agent setting, where even minor changes in transition probabilities can alter not only the individual behavior of agents but also the interactions between them.

For our experiments, we utilize the 3×3 Gridworld example, also considered in Hu and Wellman [2003]. To recover the feasible reward set and learn the expert policy with BC, we consider a scenario where the transition probabilities are deterministic. The Nash experts are learned via NashQ-Learning as proposed in Hu and Wellman [2003]. The resulting Nash Experts and more details on the environments can be found in H.2.

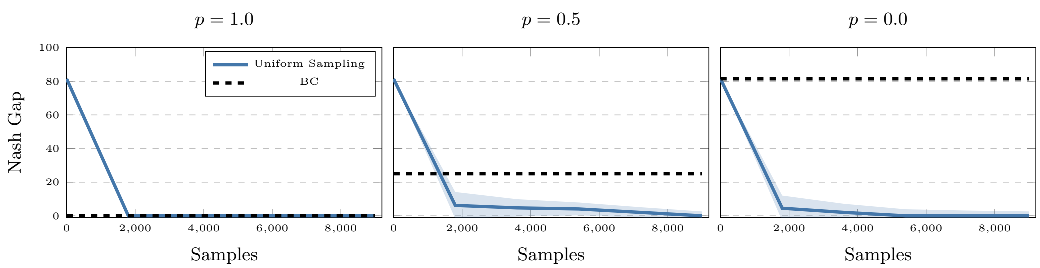

Figure 4: Nash Gap in Grid Games for different transition probabilities.

Using the Uniform Sampling algorithm to recover the entire feasible set, we then apply a Random Max Gap MAIRL algorithm to extract a reward function from the feasible set, similar to the approach introduced in Appendix C of Metelli et al. [2021]. A detailed description can be found in Appendix H.2. To test the transferability of the recovered reward function, we alter the transition probabilities so that in states $(0, \text{any})$ and $(\text{any}, 2)$, taking action "up" is only successful with a probability of 0.5; otherwise, the agent remains in the same state (as in Grid Game 2 in Hu and Wellman [2003]). In a second scenario, we introduce an obstacle into the environment, that prohibits the agent from passing through. While in the first scenario, the BC strategy still performs reasonably, the second altered environment leads to a failure to reach the goal state for agent 1, resulting in the maximal Nash Gap.

We observe that while BC may perform better in the original environment, for the first iterations, the Uniform Sampling Algorithm proves superior when transferring the reward function, especially as the number of samples increases and the environment changes.

Multi-agent Gridworld. In this section, we describe the environments used for the experiments. The environments are similar to the ones used in Hu and Wellman [2003]. We adjust them in such that for the random transition probabilities in the states $(0, \text{any})$ and $(2, \text{any})$ the environment still has different goals for each agent. Additionally, we introduce the scenario, where an obstacle is added in the middle of the environment, that bounces the agent back with probability 1. This results in a failure of reaching the goal for agent 1 in the BC case.
In the left column, we draw the learned Nash path for both agents in the deterministic environment, when applying the NashQ Learning algorithm to retrieve an expert policy.

Max Gap MAIRL. In this section, we describe the Max Gap MAIRL algorithm, an extension of the approach presented by Metelli et al. [2021] (see Appendix C in Metelli et al. [2021]) to the multi-agent setting. This algorithm is chosen due to the limited number of existing works that address the selection of feasible reward functions in general-sum Markov games with a Nash expert, particularly without imposing additional assumptions on the reward structure. Given the simplicity of the chosen environments, the Max Gap MAIRL procedure is a suitable choice.

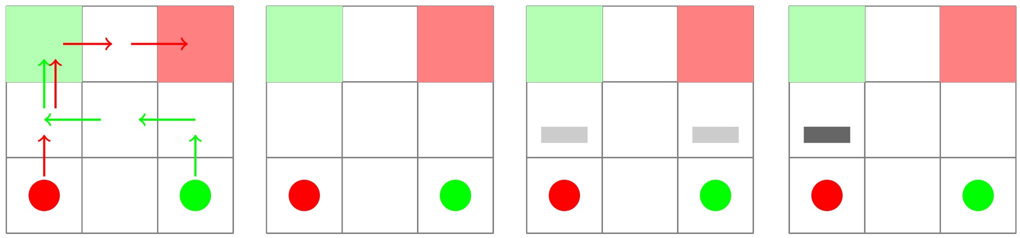

Figure 5: Multi-agent grid world environments with different transition probabilities and learned NE path

The algorithm operates as follows: for each agent $i \in [n]$, a random reward function $\tilde{R}^i$ is selected such that $\|\tilde{R}^i\| \leq R_{\max}$. The next step involves finding a reward function R^i that minimizes the squared 2-norm distance to the randomly chosen reward $\tilde{R}^i$, subject to two constraints: (1) R^i must belong to the recovered feasible set, and (2) it must maximize the maximal reward gap, thereby enforcing the feasible reward condition as introduced in D.4. This results in the following constrained quadratic optimization problem:

$$\begin{aligned}
&\max_{R^i \in \mathbb{R}^{\mathcal{S}}, A^{i,\boldsymbol{\pi}}} \quad A^{i,\boldsymbol{\pi}} \\
&\text{s.t.} \quad (\boldsymbol{\pi}^{\text{Nash}} - \tilde{\boldsymbol{\pi}})(I - \gamma \boldsymbol{P} \boldsymbol{\pi}^{\text{Nash}})^{-1} \boldsymbol{R}^i \geq A^{i,\boldsymbol{\pi}} \mathbf{1}_{\{\pi^{i,\text{Nash}}=0\}} \mathbf{1}_{\{\pi^{-i,\text{Nash}}>0\}} \mathbf{1}_{\mathcal{S}\times\mathcal{A}}, \\
&\qquad \|R^i\|_\infty \leq R_{\max}.
\end{aligned}$$

I Technical Results

In this section, we present results that were used throughout this work.

Theorem I.1 (compare Theorem 1 in Cao et al. [2021]). *For a fixed policy $\bar{\pi}(a|s) > 0$, discount factor $\gamma \in [0, 1)$, and an arbitrary choice of function $v : \mathcal{S} \to \mathbb{R}$, there is a unique corresponding reward function*

$$r(s,a) = \lambda \log \bar{\pi}(a|s) - \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) v(s') + v(s)$$

such that the MDP with reward r yields a value function $V_\lambda^* = v$ and entropy-regularized optimal policy $\pi_\lambda^* = \bar{\pi}$.

Proof. Fix r as in the statement of the theorem. Then the corresponding value function is given by

$$\begin{aligned}
V_\lambda^*(s) &= \lambda \log \sum_{a \in \mathcal{A}} \exp\left(\frac{1}{\lambda}\left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V_\lambda^*(s')\right)\right) \\
&= v(s) + \lambda \log \sum_{a \in \mathcal{A}} \bar{\pi}(a|s) \exp\left(\frac{\gamma}{\lambda} \sum_{s' \in \mathcal{S}} P(s'|s,a)(V_\lambda^*(s') - v(s'))\right),
\end{aligned}$$

which rearranges to give

$$\exp(g(s)) = \sum_{a \in \mathcal{A}} \bar{\pi}(a|s) \exp\left(\gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) g(s')\right) \tag{15}$$

with $g(s) = (V_\lambda^*(s) - v(s))/\lambda$. Applying Jensen's inequality, we can see that, for $\underline{s} \in \arg\min_{s \in \mathcal{S}} g(s)$,

$$\exp\left(\min_s g(s)\right) = \exp\left(g(\underline{s})\right) \geq \exp\left(\gamma \sum_{a \in \mathcal{A}, s' \in \mathcal{S}} \bar{\pi}(a|\underline{s}) P(s'|\underline{s}, a) g(s')\right).$$

However, the sum on the right is a weighted average of the values of g, so

$$\sum_{a \in \mathcal{A}, s' \in \mathcal{S}} \bar{\pi}(a|\underline{s}) P(s'|\underline{s}, a) g(s') \geq \min_s g(s).$$

Combining these inequalities, along with the fact $\gamma < 1$, we conclude that $g(s) \geq 0$ for all $s \in \mathcal{S}$.

Again applying Jensen's inequality to Eq. (15), for $\bar{s} \in \arg\max_{s \in \mathcal{S}} g(s)$ we have

$$\max_s \{\exp(g(s))\} = \exp(g(\bar{s})) \leq \sum_{a \in \mathcal{A}, s' \in \mathcal{S}} \bar{\pi}(a|\bar{s}) P(s'|\bar{s}, a) \exp(\gamma g(s')) .$$

As the sum on the right is a weighted average, we know

$$\sum_{a \in \mathcal{A}, s' \in \mathcal{S}} \bar{\pi}(a|\bar{s}) P(s'|\bar{s}, a) \exp(\gamma g(s')) \leq \max_s \{\exp(\gamma g(s))\} .$$

Hence, as $\gamma < 1$, we conclude that $g(s) \leq 0$ for all $s \in \mathcal{S}$.

Combining these results, we conclude that $g \equiv 0$, that is, $V_\lambda^* = v$. Finally, we substitute the definition of r and the value function v into (6) to see that the entropy-regularized optimal policy is $\pi_\lambda^* = \bar{\pi}$. □

The next lemma is an extension of Lemma 3 from Zanette et al. [2019] for the multi-agent setting, accounting that different Nash equilibria can have different values.

Lemma I.2 (Simulation Lemma)**.** *Let $i \in [n]$ be any agent. Then it holds true that*

$$\begin{aligned} &\hat{V}^{i,\boldsymbol{\pi}}(s) - V^{i,\boldsymbol{\pi}}(s) \\ &= \sum_{s,\boldsymbol{a}} \overline{w}_{s,\boldsymbol{a}}^{\boldsymbol{\pi}} \left(\hat{R}^i(s,\boldsymbol{a}) - R^i(s,\boldsymbol{a}) + \gamma \left(\sum_{s'} \left(\hat{P}(s' \mid s,\boldsymbol{a}) - P(s' \mid s,\boldsymbol{a}) \right) V^{i,\boldsymbol{\pi}}(s') \right) \right) \end{aligned}$$

Proof. Let the starting distribution be a dirac measure on some $s \in \mathcal{S}$. It then holds that

$$\begin{aligned} &\hat{V}^{i,\boldsymbol{\pi}}(s) - V^{i,\boldsymbol{\pi}}(s) \\ &= \hat{R}^i(s,\mathbf{a}) - R^i(s,\mathbf{a}) + \gamma \left(\sum_{s'} \hat{P}(s' \mid s,\mathbf{a}) \hat{V}^{i,\boldsymbol{\pi}}(s') - P(s' \mid s,\mathbf{a}) V^{i,\boldsymbol{\pi}}(s') \right) \\ &= \hat{R}^i(s,\mathbf{a}) - R^i(s,\mathbf{a}) + \gamma \left(\hat{P}(s' \mid s,\mathbf{a}) - P(s' \mid s,\mathbf{a}) \right) V^{i,\boldsymbol{\pi}}(s') \\ &\quad + \gamma \sum_{s'} \hat{P}(s' \mid s,\mathbf{a}) (\hat{V}^{i,\boldsymbol{\pi}}(s') - V^{i,\boldsymbol{\pi}}(s')) \end{aligned}$$

The proof follows by induction. □

Lemma I.3. *Let μ^*, ν^* be the QRE equilibrium expert policies, such that Assumption F.6 holds true and $\hat{\mu}^*, \hat{\nu}^*$ the respective empirical estimators with samples N. Then for $\delta \in (0,1)$ it holds true with probability $1-\delta$ if $\sqrt{\frac{\log(2/\delta)}{2N_k^+(s)}} \leq \frac{\Delta_{\min}}{2}$, that for $a \in \mathcal{A}, b \in \mathcal{B}$*

$$|\log(\mu^*(a \mid s)) - \log(\hat{\mu}^*(a \mid s))| \leq \frac{2}{\Delta_{\min}} \sqrt{\frac{\log(2/\delta)}{2N}}$$
$$|\log(\nu^*(a \mid s)) - \log(\hat{\nu}^*(a \mid s))| \leq \frac{2}{\Delta_{\min}} \sqrt{\frac{\log(2/\delta)}{2N}}$$

Proof. The difference in logarithms can be bounded using the inequality:

$$|\log(\mu(a \mid s)) - \log(\hat{\mu}(a \mid s))| \leq \frac{|\mu(a \mid s) - \hat{\mu}(a \mid s)|}{\min(\mu(a \mid s), \hat{\mu}(a \mid s))}.$$

This follows from the fact that the derivative of $\log(x)$ is $1/x$, so the difference in logarithms is controlled by the relative difference in probabilities.

By Hoeffding's inequality, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, with probability at least $1 - \delta/(SA)$, we have:

$$|\mu(a \mid s) - \hat{\mu}(a \mid s)| \leq \sqrt{\frac{\log(2/\delta)}{2N_k^+(s)}}.$$

Since $\mu(a \mid s) \geq \Delta_{\min}$, and assuming $\hat{\mu}(a \mid s)$ is close to $\mu(a \mid s)$, we have:

$$\hat{\mu}(a \mid s) \geq \mu(a \mid s) - \sqrt{\frac{\log(2/\delta)}{2N_k^+(s)}} \geq \Delta_{\min} - \sqrt{\frac{\log(2SA/\delta)}{2N_k^+(s)}}.$$

To ensure $\hat{\mu}(a \mid s) > 0$, we require:

$$\sqrt{\frac{\log(2/\delta)}{2N_k^+(s)}} < \Delta_{\min}.$$

Under this condition, the denominator satisfies:

$$\min(\mu(a \mid s), \hat{\mu}(a \mid s)) \geq \Delta_{\min} - \sqrt{\frac{\log(2/\delta)}{2N_k^+(s)}}.$$

Substitute the bounds for $|\mu(a \mid s) - \hat{\mu}(a \mid s)|$ and $\min(\mu(a \mid s), \hat{\mu}(a \mid s))$ into the inequality for the difference in logarithms:

$$|\log(\mu(a \mid s)) - \log(\hat{\mu}(a \mid s))| \leq \frac{\sqrt{\frac{\log(2/\delta)}{2N_k^+(s)}}}{\Delta_{\min} - \sqrt{\frac{\log(2/\delta)}{2N_k^+(s)}}}.$$

If $\sqrt{\frac{\log(2/\delta)}{2N_k^+(s)}} \leq \frac{\Delta_{\min}}{2}$, then:

$$|\log(\mu(a \mid s)) - \log(\hat{\mu}(a \mid s))| \leq \frac{2\sqrt{\frac{\log(2/\delta)}{2N_k^+(s)}}}{\Delta_{\min}}.$$

□

Lemma I.4. *Let p_i be a probability such that $p_i \geq \Delta_{\min} > 0$, and let $\hat{p}_i$ be the empirical estimate of p_i based on n independent samples. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, the following bound holds:*

$$\left|\frac{1}{p_i} - \frac{1}{\hat{p}_i}\right| \leq \frac{\sqrt{\frac{\log(2/\delta)}{2n}}}{\Delta_{\min}\left(\Delta_{\min} - \sqrt{\frac{\log(2/\delta)}{2n}}\right)}.$$

Furthermore, if $\sqrt{\frac{\log(2/\delta)}{2n}} \leq \frac{\Delta_{\min}}{2}$, the bound simplifies to:

$$\left|\frac{1}{p_i} - \frac{1}{\hat{p}_i}\right| \leq \frac{2\sqrt{\frac{\log(2/\delta)}{2n}}}{\Delta_{\min}^2}.$$

Proof. The difference can be rewritten as:

$$\left|\frac{1}{p_i} - \frac{1}{\hat{p}_i}\right| = \frac{|\hat{p}_i - p_i|}{p_i \hat{p}_i}.$$

By Hoeffding's inequality, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have:

$$|\hat{p}_i - p_i| \leq \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Let $\epsilon = \sqrt{\frac{\log(2/\delta)}{2n}}$. Then, with high probability:

$$|\hat{p}_i - p_i| \leq \epsilon.$$

Since $p_i \geq \Delta_{\min}$ and $|\hat{p}_i - p_i| \leq \epsilon$, we have:

$$\hat{p}_i \geq p_i - \epsilon \geq \Delta_{\min} - \epsilon.$$

To ensure $\hat{p}_i > 0$, we require $\epsilon < \Delta_{\min}$.

Using $p_i \geq \Delta_{\min}$ and $\hat{p}_i \geq \Delta_{\min} - \epsilon$, the denominator satisfies:

$$p_i \hat{p}_i \geq \Delta_{\min}(\Delta_{\min} - \epsilon).$$

Substitute the bounds for $|\hat{p}_i - p_i|$ and $p_i\hat{p}_i$ into the expression for the difference:

$$\left| \frac{1}{p_i} - \frac{1}{\hat{p}_i} \right| = \frac{|\hat{p}_i - p_i|}{p_i \hat{p}_i} \leq \frac{\epsilon}{\Delta_{\min}(\Delta_{\min} - \epsilon)}.$$

This gives the first part of the lemma.

If $\epsilon \leq \frac{\Delta_{\min}}{2}$, then $\Delta_{\min} - \epsilon \geq \frac{\Delta_{\min}}{2}$, and the bound simplifies to:

$$\left| \frac{1}{p_i} - \frac{1}{\hat{p}_i} \right| \leq \frac{\epsilon}{\Delta_{\min} \cdot \frac{\Delta_{\min}}{2}} = \frac{2\epsilon}{\Delta_{\min}^2}.$$

Substituting $\epsilon = \sqrt{\frac{\log(2/\delta)}{2n}}$, we obtain:

$$\left| \frac{1}{p_i} - \frac{1}{\hat{p}_i} \right| \leq \frac{2\sqrt{\frac{\log(2/\delta)}{2n}}}{\Delta_{\min}^2}.$$

This completes the proof of the lemma. $\square$

Lemma I.5 (Concentration Inequality for Total Variation Distance, see e.g. Thm 2.1 by Berend and Kontorovich [2012]). *Let $\mathcal{X} = \{1, 2, \cdots, |\mathcal{X}|\}$ be a finite set. Let P be a distribution on $\mathcal{X}$. Furthermore, let $\widehat{P}$ be the empirical distribution given m i.i.d. samples $x_1, x_2, \cdots, x_n$ from P, i.e.,*

$$\widehat{P}(j) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{x_i = j\}.$$

Then, with probability at least $1 - \delta$, we have that

$$\|P - \widehat{P}\|_1 := \sum_{x \in \mathcal{X}} |P(x) - \widehat{P}(x)| \leq \sqrt{\frac{2|\mathcal{X}| \log(1/\delta)}{n}}.$$

Proof. Define the function $f(x_1, \ldots, x_n) = \sum_{x \in \mathcal{X}} |\widehat{P}(x) - P(x)|$, where $\widehat{P}$ is the empirical distribution. Replacing one sample x_i can change f by at most $2/n$, since the empirical frequencies change by at most $1/n$ per coordinate and total variation sums these differences.

By McDiarmid's inequality, we have for any $\epsilon > 0$,

$$\Pr\left(f - \mathbb{E}[f] \geq \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2}\right).$$

Berend and Kontorovich (2013) show that $\mathbb{E}[f] \leq \sqrt{\frac{|\mathcal{X}|}{n}}$. Setting the failure probability to δ, we solve

$$\exp\left(-\frac{n\epsilon^2}{2}\right) = \delta \quad \Longrightarrow \quad \epsilon = \sqrt{\frac{2\log(1/\delta)}{n}}.$$

Therefore, with probability at least $1 - \delta$,

$$\|P - \widehat{P}\|_1 \leq \sqrt{\frac{|\mathcal{X}|}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}} \leq \sqrt{\frac{2|\mathcal{X}| \log(1/\delta)}{n}},$$

$\square$