European Workshop on Reinforcement Learning 15 (2022) September 2022, Milan, Italy.

Interactive Inverse Reinforcement Learning

Thomas Kleine Buening thomkl@uio.no
Department of Informatics
University of Oslo

Anne-Marie George annemage@ifi.uio.no
Department of Informatics
University of Oslo

Christos Dimitrakakis christos.dimitrakakis@gmail.com
Department of Computer Science
University of Neuchâtel

Abstract

We study the problem of designing autonomous agents that can learn to cooperate effectively with a potentially suboptimal partner while having no access to the joint reward function. This problem is modeled as a cooperative episodic two-agent Markov decision process. We assume control over only the first of the two agents in a Stackelberg formulation of the game, where the second agent is acting so as to maximise expected utility given the first agent's policy. How should the first agent act in order to learn the joint reward function as quickly as possible and so that the joint policy is as close to optimal as possible? We analyse how knowledge about the reward function can be gained in this interactive two-agent scenario. We show that when the learning agent's policies have a significant effect on the transition function, the reward function can be learned efficiently.

Keywords: Inverse Reinforcement Learning, Human-AI Collaboration

1. Introduction

Recent applications of autonomous systems in our daily lives show that autonomous agents are no longer deployed in isolation only, but in situations where they are in close interaction with humans. To facilitate successful and safe cooperation between autonomous systems and humans, we need to design agents that can learn about human preferences as well as adapt to suboptimal human behaviour. We focus on the situation where the autonomous agent and the human simultaneously act in the same environment. As a result, observed human behaviour, which could be used to infer preferences, depends on the learning agent's actions. This leads to the problem of learning preferences and intentions from interactions. Learning in these interactive scenarios brings its own challenges, but also significant benefits as we will see in the following.

In this paper, we consider the problem of learning to cooperate with a potentially suboptimal partner while having no access to the joint reward function. This problem is modeled as a cooperative episodic Markov Decision Process (MDP) between two agents $\mathcal{A}_1$ and $\mathcal{A}_2$. While agent $\mathcal{A}_2$ (the human) knows the joint reward function, we take the perspective of agent $\mathcal{A}_1$ (the learner) that has to cooperate with $\mathcal{A}_2$ without knowing or observing the rewards. As an example, consider a maze in which the human tries to reach a target while the learning agent can unlock doors to help the human move, but without knowing the precise target location. We focus on the Stackelberg formulation of the game, in which at the beginning of each episode the learner commits to a policy before the human does. This allows us to view the learning agent as a *designer of environments* that the human operates in. For instance, when the learning agent's actions correspond to unlocking doors in a grid world, then, in the Stackelberg game, we can interpret the learner's policy as choosing a maze layout, which is communicated to the human at the beginning of the episode and in which she operates.

Inverse Reinforcement Learning (IRL) (Russell, 1998) can be used to infer the reward function of an agent from observations of that agent's behaviour, which is assumed to be (near-)optimal. In our case, the learner also obtains observations of the human's behaviour through interactions, which could then be used to infer the joint reward function.

However, the human's actions, e.g. the path taken in a maze, depend on the learner's policy, e.g. the maze layout, so that in contrast to the standard IRL formulation the learner now *actively influences* the demonstrations of the human expert. This leads to an interesting Interactive IRL setting, where the learner can actively seek information about the joint reward function by playing specific policies. In this paper, we analyse how to infer the unknown (joint) reward function from interactions with the expert and how the learner should choose its policy so that the two agents collaborate efficiently over both the short and long term. We lay an emphasis on the role of the learner as the designer of environments and investigate what environments allow the learning agent to infer the reward function quickly while achieving high levels of cooperation.

Outline and Contribution. We discuss related work in Section 2 and formally introduce the setting in Section 3. Section 4 considers the case where $\mathcal{A}_2$ plays *optimally*. We show how to learn about the reward function from interactions with $\mathcal{A}_2$ and prove the existence of ideal reward learning environments. We then construct an algorithm based on linear programming that is no-regret under mild assumptions. Section 5 considers the case where $\mathcal{A}_2$ responds *suboptimally*. In Section 5.1, we adapt conventional Bayesian IRL methods for estimating the reward function to our setting. We then analyse optimal commitment strategies for cooperating with suboptimal followers in Section 5.2. Section 6 describes the experiments, which we perform on random MDPs and specially constructed maze problems. Our experiments support our theoretical results and show that the interactive nature of our setting allows the learning agent to obtain a much better estimate of the reward function (compared to the standard IRL setting). We thus achieve better cooperation by intelligently probing the human's responses. Future work is discussed in Section 7. Finally, omitted proofs, experimental details and algorithms can be found in the Appendix.

2. Related Work

Since our setting requires (a) inferring the joint reward function, as in IRL, and (b) collaborating with a potentially suboptimal agent, in this section we present related work in those two domains.

Inverse Reinforcement Learning. IRL (Russell, 1998) aims to find a reward function that explains observed behaviour of an agent. We face the same problem, with the main difference being that *two* agents act in the environment simultaneously, one of which (the human) knows the reward function and the other (the learner) does not. Our algorithm for the case when $\mathcal{A}_2$ is optimal is based on a characterisation of reward functions consistent with an optimal policy, similarly to Ng and Russell (2000). Ramachandran and Amir (2007) adopt a Bayesian perspective to the IRL problem as it provides a principled way to reason under uncertainty. The Bayesian formulation of the IRL problem can naturally account for suboptimal demonstrations as well as partial information and we will show how to translate the Bayesian approach to our interactive IRL setting.

Hadfield-Menell et al. (2016) introduce the problem of cooperative IRL in which a robot must cooperate with a human but does not initially know the reward function. Their work focuses on apprenticeship learning, where the robot and the human *take turns* demonstrating and performing a task. In particular, they examine the problem of calculating optimal human demonstrations for the robot to observe. Instead, we consider the situation when the agents *interact* by simultaneously acting in the same environment. Our setting also notably differs from apprenticeship learning (Abbeel and Ng, 2004) and imitation learning (Ratliff et al., 2006) more generally in that our goal is *not* to mimic the behaviour of $\mathcal{A}_2$, as effective cooperation between $\mathcal{A}_1$ and $\mathcal{A}_2$ may require both agents to perform entirely different tasks. Nikolaidis and Shah (2013) consider a cross-training approach in which a human expert and a robot repeatedly switch roles. In the first of two phases, the expert operates in an environment, which is influenced by the robot. The learner then observes the expert and updates its estimates of the reward function. In the second phase, the robot then demonstrates the learned policy while the expert influences the transitions. Crucially, in this approach the human steers the learning of the robot similar to teaching approaches for IRL (Brown and Niekum, 2019; Parameswaran et al., 2019). In contrast, we consider the situation where the learner actively seeks information from the human over whom we have no control. Natarajan et al. (2010) consider a multi-agent extension of IRL in which the learner observes multiple experts maximising a joint reward function. Similarly, Lin et al. (2019) address the problem of multi-agent IRL in certain general-sum games. In contrast to their work, we consider the case where the learner is *not* a passive observer, but interacts with the other agent and thereby influences what observations it collects.

Zhang and Parkes (2008) and Zhang et al. (2009) consider the problem of *environment design*: how to modify an environment so as to influence an agent's decisions. They analyse how to construct *reward incentives* to induce a particular policy when the reward function of the acting agent is unknown. In our setting, we can also view the learner as a designer of environments that the human operates in, however, with the difference that the learner influences transitions, but not the underlying reward function. Moreover, our goal is generally not to steer the human towards certain behaviour, but rather to learn from and cooperate with a human expert.

Cooperating with suboptimal partners. In the context of human-AI collaboration, there have been recent efforts addressing the problem of cooperating with a potentially suboptimal partner *when the reward function is known.* In particular, Dimitrakakis et al. (2017) and Radanovic et al. (2019) consider a setting where the human responds suboptimally to the learning agent's policy. The former focuses on a single-stage Stackelberg game, while the latter on an online learning variant of the problem. However, in both cases the learning agent knows the human's reward function. Our work also has some links to the problem of optimal commitment in Stackelberg games (Conitzer and Sandholm, 2006; Letchford et al., 2012). While prior work assumes optimal responses and a potentially competitive game, we focus on finding optimal commitment strategies when playing with a *suboptimal* follower in a strictly *cooperative* setting.

3. Setting

We model the problem as a cooperative two-agent MDP $(\mathcal{S}, A_1, A_2, \mathcal{P}, r, \gamma)$ between agents $\mathcal{A}_1$ and $\mathcal{A}_2$, where $\mathcal{S}$ denotes a finite state space, A_i the finite action space of agent $\mathcal{A}_i$, $\mathcal{P} : \mathcal{S} \times A_1 \times A_2 \to \Delta(\mathcal{S})$ the transition function, $r : \mathcal{S} \to \mathbb{R}$ the *joint* reward function and $\gamma \in [0, 1)$ the discount factor. We will take the perspective of agent $\mathcal{A}_1$ that, without knowing or observing the joint reward function, aims to cooperate with its partner $\mathcal{A}_2$. We assume that the interaction between the two agents and the environment takes place in a sequence of episodes, where at the beginning of each episode, $\mathcal{A}_1$ commits to a policy π^1 first. Agent $\mathcal{A}_2$ then responds with a policy π^2 and the joint policy is executed until the end of the episode.[1] We assume that agents $\mathcal{A}_1$ and $\mathcal{A}_2$ know the transition dynamics.

Interaction. The repeated interaction of both agents can be specified as the following *Stackelberg game*. In episode t:

1) $\mathcal{A}_1$ commits to policy π^1_t,
2) $\mathcal{A}_2$ observes π^1_t and responds with policy π^2_t,
3a) $\mathcal{A}_1$ observes the *fully specified policy* π^2_t, *or*
3b) $\mathcal{A}_1$ observes a *trajectory* τ_t of (random) length $H + 1$, where $\tau_t = (s_0, a_0, b_0, \dots, s_H, a_H, b_H)$.

Alternative 3a) describes the *full information* setting in which the complete policy π^2_t is available to the learner at the end of each episode. This could, for instance, be the case when interaction takes place for a sufficiently long time in each episode, or the same policy is committed by $\mathcal{A}_1$ several times and thus $\mathcal{A}_2$'s response is explored exhaustively. Alternative 3b) constitutes the case of *partial information* in which $\mathcal{A}_1$ interacts with $\mathcal{A}_2$ in a series of $H + 1$ time steps and observes the generated trajectory only.

3.1 Preliminaries

By a slight abuse of notation, we sometimes refer to functions $f : \mathcal{S} \to \mathbb{R}$ as vectors $f \in \mathbb{R}^{|\mathcal{S}|}$. For instance, when convenient, we treat reward functions $r : \mathcal{S} \to \mathbb{R}$ as vectors $r \in \mathbb{R}^{|\mathcal{S}|}$. Let V_{π^1,π^2} denote the value function under joint policy (π^1, π^2). The value function satisfies the Bellman equation, which we can concisely express in matrix-form as

$$V_{\pi^1,\pi^2} = (I - \gamma \mathcal{P}_{\pi^1,\pi^2})^{-1} r,$$

where V_{π^1,π^2} and r are column vectors and $\mathcal{P}_{\pi^1,\pi^2}$ is the transition matrix obtained from $\mathcal{P}$ by marginalising over policy (π^1, π^2). Let $Q_{\pi^1,\pi^2}(s, a, b)$ denote the value of taking joint action (a, b) in state s under policy (π^1, π^2). When $\mathcal{A}_1$ commits to a policy π^1 first, agent $\mathcal{A}_2$ gets to plan under marginalised transitions $\mathcal{P}_{\pi^1} : \mathcal{S} \times A_2 \to \Delta(\mathcal{S})$ given by $\mathcal{P}_{\pi^1}(s'|s, b) = \mathbb{E}_{a \sim \pi^1}[\mathcal{P}(s'|s, a, b)]$. The Q-values for $\mathcal{A}_2$ under transitions $\mathcal{P}_{\pi^1}$ then equal $Q_{\pi^1,\pi^2}(s, b) = \mathbb{E}_{a \sim \pi^1}[Q_{\pi^1,\pi^2}(s, a, b)]$ and we denote the optimal Q-value under $\mathcal{P}_{\pi^1}$ by $Q^*_{\pi^1}(s, b) = \max_{\pi^2} Q_{\pi^1,\pi^2}(s, b)$.

1. Even in MDPs without termination condition, discounting corresponds to episodes that end w.p. $1 - \gamma$ each time step.

Behavioural Models for $\mathcal{A}_2$. A typical assumption about the behaviour of a partner (or opponent) in game theory (Nisan et al., 2007) and IRL (Ng and Russell, 2000) is that of optimal behaviour, sometimes referred to as fully rational behaviour. In our case, this means that in episode t, agent $\mathcal{A}_2$ plays an optimal response $\pi_t^2(\pi_t^1)$ to the policy π_t^1 committed by agent $\mathcal{A}_1$. Note that we will simply write π_t^2 when the dependence on π_t^1 is clear from the context.

We are also interested in the case when $\mathcal{A}_2$ is suboptimal. A common decision-model for suboptimal human behaviour in IRL (Jeon et al., 2020), economics (Luce, 1959), and cognitive science (Baker et al., 2009) are Boltzmann-rational policies for which the probability of choosing an action is exponentially dependent on its expected value:

$$\pi^2(b \mid s, \pi^1) \propto \exp\big(\beta Q^*_{\pi^1}(s,b)\big).$$

Here, $\beta \geq 0$ is called the inverse temperature of the distribution and indicates the irrationality of $\mathcal{A}_2$. For $\beta = 0$, $\mathcal{A}_2$ acts uniformly at random, and for $\beta \to \infty$, $\mathcal{A}_2$ acts perfectly rational, i.e. optimally in response to $\mathcal{A}_1$'s commitment.

Objective and Regret. Agent $\mathcal{A}_1$ aims to maximise the expected sum of discounted rewards by learning about the joint reward function and cooperating with $\mathcal{A}_2$. In general, due to the possibly suboptimal nature of $\mathcal{A}_2$, we have that $\max_{\pi^1} V_{\pi^1,\pi^2(\pi^1)} \preceq \max_{\pi^1,\pi^2} V_{\pi^1,\pi^2}$, i.e. the value of the game under $\mathcal{A}_2$'s behavioural model is bounded by the value of the joint optimal policy. For an initial state distribution D, we define the value of the optimal commitment strategy as

$$V^* = \max_{\pi^1} \mathbb{E}_{s_0 \sim D}\big[V_{\pi^1,\pi^2(\pi^1)}(s_0)\big],$$

where $\pi^2(\pi^1)$ denotes the response of $\mathcal{A}_2$ to policy π^1. Note that the optimal value V^* may only be well-defined with respect to a specific initial state distribution as a dominating commitment strategy may fail to exist when $\mathcal{A}_2$ responds suboptimally (see Section 5.2). We define the (per-episode) regret of playing policy π^1 as the difference $\mathscr{L}(\pi^1) = V^* - \mathbb{E}_{s_0\sim D}[V_{\pi^1,\pi^2(\pi^1)}(s_0)]$. Similarly, we define the (online) regret of playing policies $\pi_1^1, \dots, \pi_T^1$ as the sum $\mathscr{L}(\pi_1^1, \dots, \pi_T^1) = \sum_{t=1}^T \mathscr{L}(\pi_t^1)$.

3.2 Interactive IRL

In the classical IRL problem, the learning agent observes an expert performing a task. The observations are then interpreted as demonstrations of (near-)optimal behaviour in a *fixed* single-agent MDP with unknown reward function. Our setting is substantially different, as two agents must collaborate in the same two-agent MDP, with the first agent not knowing the common reward function. As a result, the second agent's demonstrations depend on the first agent's policy and so become *context-dependent*. In addition, learning must take place in an *online* fashion, as the first agent must adapt its policy to extract information and to better collaborate.

$\mathcal{A}_1$ as an MDP Designer. When the first agent $\mathcal{A}_1$ commits to a policy π^1 at the beginning of an episode, then – with knowledge of π^1 – the second agent $\mathcal{A}_2$ can be seen as planning in a single-agent MDP with transition function $\mathcal{P}_{\pi^1}$. Consequently, from the perspective of the learner, choosing a policy π^1 is equivalent to designing single-agent MDPs for the human expert to act in. While the state space, $\mathcal{A}_2$'s action space, the (unknown) reward function as well as the discount factor remain the same across these simplified MDPs, agent $\mathcal{A}_2$ may face different environment dynamics $\mathcal{P}_{\pi^1}$ depending on $\mathcal{A}_1$'s policy. This is in contrast to the standard IRL setting in which demonstrations always take place in the same fixed MDP. An abstract example where $\mathcal{A}_1$ creates different environments for $\mathcal{A}_2$ to operate in is illustrated in Figure 1(a).

Context-Dependent Responses. Agent $\mathcal{A}_1$ can now interpret $\mathcal{A}_2$'s response to a policy π^1 as a demonstration in the single-agent MDP $(\mathcal{S}, A_2, \mathcal{P}_{\pi^1}, r^*, \gamma)$, where r^* is the true reward function that is unknown and unobserved by $\mathcal{A}_1$. Since $\mathcal{A}_2$ faces possibly different environment dynamics across episodes, we can also expect $\mathcal{A}_2$'s behaviour to vary between episodes. In Figure 1(a), for instance, $\mathcal{A}_2$ adapts their policy to the specific maze layout created by the learner. Therefore, $\mathcal{A}_2$'s responses (and thus demonstrations) become context-dependent in the sense that they always depend on $\mathcal{A}_1$'s policy, i.e. the environment that is implicitly generated by $\mathcal{A}_1$. In particular, we see that even though the underlying reward function remains the same, the results of IRL methods vary depending on the environment in which demonstrations were provided.

Online Learning. As the game progresses, the learner interacts with the expert in a series of episodes, thereby collecting a stream of observations. Then, in order to extract more information as well as to improve cooperation in

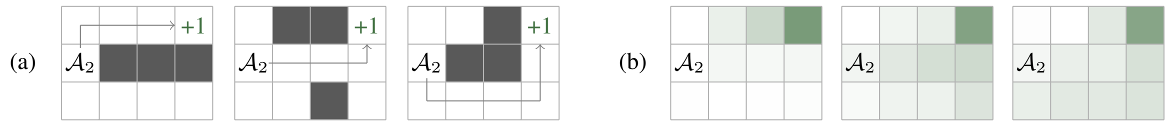


Figure 1: (a) $\mathcal{A}_1$ designs a maze for $\mathcal{A}_2$ to navigate in and collect a reward in the top right corner. $\mathcal{A}_2$ behaves differently, i.e. chooses a different path, depending on the maze created by $\mathcal{A}_1$. (b) The mean reward function computed using Bayesian IRL (Ramachandran and Amir, 2007) when observing $\mathcal{A}_2$ navigate in each of the three mazes. Dark colours denote higher estimated rewards.

the next episode, $\mathcal{A}_1$ may want to leverage the observations up to episode t to learn about the joint reward function and to inform its decisions in episode $t+1$. Naturally, since the learner *actively influences* the demonstrations by the expert, we ask ourselves whether demonstrations under some environment dynamics $\mathcal{P}_{\pi^1}$ are more informative than others. In particular, how much more information (if any) can be gained from demonstrations in unseen environments? In the following, we will address these questions both theoretically and empirically.

4. Cooperating with Optimal Agents

Here we consider the case when $\mathcal{A}_2$ responds optimally to the commitment of $\mathcal{A}_1$. In Section 4.1, we characterise the set of *feasible* reward functions, i.e. those that are consistent with observed responses, and prove the existence of ideal (reward) learning environments. We then describe an algorithm that is no-regret under an assumption on the identifiability of suboptimal behaviour in Section 4.2. The omitted proofs can be found in Appendix A.

4.1 Learning from Optimal Responses

For our theoretical analysis, we focus on the full information setting in which $\mathcal{A}_1$ observes the fully specified policy played by $\mathcal{A}_2$ at the end of each episode. In a first step, we define a *feasible* reward function under (π^1, π^2) as a reward function for which $\mathcal{A}_2$'s response to the commitment of $\mathcal{A}_1$ is optimal.

Definition 1 *We say that a reward function r is* feasible *when observing policy π^2 in response to π^1 if π^2 is optimal in the single-agent MDP $(\mathcal{S}, A_2, \mathcal{P}_{\pi^1}, r, \gamma)$.*

We now adapt the standard result by Ng and Russell (2000) to obtain a characterisation of the set of feasible reward functions under policies π^1 and π^2. Here, we let $\succeq$ denote element-wise inequality.

Theorem 1 (Ng and Russell (2000)) *Let there be an MDP without reward function $(\mathcal{S}, A_1, A_2, \mathcal{P}, \gamma)$. A reward function r is feasible under policies π^1 and π^2 if and only if*

$$\left(\mathcal{P}_{\pi^1,\pi^2} - \mathcal{P}_{\pi^1,b}\right)\left(I - \gamma \mathcal{P}_{\pi^1,\pi^2}\right)^{-1} r \succeq 0 \quad \forall b \in A_2,$$

where $\mathcal{P}_{\pi^1,b}$ is the one-step transition matrix under policy π^1 and action $b \in A_2$.

Since $\mathcal{A}_1$ and $\mathcal{A}_2$ repeatedly interact in a series of episodes, a reward function is feasible after t episodes if and only if it is feasible under all policies $\pi^1_1, \dots, \pi^1_t$ and corresponding responses $\pi^2_1, \dots, \pi^2_t$. As an immediate consequence of Theorem 1, we then obtain the following characterisation of rewards that are feasible under multiple observations.

Corollary 1 *Let there be an MDP without reward function $(\mathcal{S}, A_1, A_2, \mathcal{P}, \gamma)$. A reward function r is feasible when observing policies $(\pi^1_1, \pi^2_1), \dots, (\pi^1_t, \pi^2_t)$ if and only if*

$$\left(\mathcal{P}_{\pi^1_1,\pi^2_1} - \mathcal{P}_{\pi^1_1,b}\right)\left(I - \gamma \mathcal{P}_{\pi^1_1,\pi^2_1}\right)^{-1} r \succeq 0 \quad \forall b \in A_2,$$
$$\dots$$
$$\left(\mathcal{P}_{\pi^1_t,\pi^2_t} - \mathcal{P}_{\pi^1_t,b}\right)\left(I - \gamma \mathcal{P}_{\pi^1_t,\pi^2_t}\right)^{-1} r \succeq 0 \quad \forall b \in A_2.$$

We denote the set of reward functions that satisfy these constraints by $\mathcal{R}_t = \mathcal{R}((\pi^1_1, \pi^2_1), \dots, (\pi^1_t, \pi^2_t))$. The IRL problem is an inherently ill-posed problem as degenerate solutions such as constant reward functions explain any observed behaviour. In fact, we see that any reward function $r \in \mathbb{R}^{|\mathcal{S}|}$ is indistinguishable from its positive affine transformations $\mathrm{Aff}(r) = \{\lambda_1 r + \lambda_2 \mathbf{1} \colon \lambda_1 \geq 0, \lambda_2 \in \mathbb{R}\}$.

Lemma 1 *If $\mathcal{A}_2$ responds optimally to the commitment of $\mathcal{A}_1$, any reward function r is indistinguishable from its positive affine transformations, i.e. r is feasible iff every $\bar{r} \in \mathrm{Aff}(r)$ is feasible.*

In particular, Lemma 1 states that all positive affine transformations of the true reward function r^* are always feasible.[2] However, since any reward function in $\mathrm{Aff}(r^*)$ induces the same optimal (joint) policy, finding it is sufficient for optimally solving the IRL problem. Perhaps surprisingly, we find that if $\mathcal{A}_1$'s policies can induce any transition matrix for $\mathcal{A}_2$, then there exists a policy π^1 such that its optimal response $\pi^2(\pi^1)$ can only be explained by positive affine transformations of the true reward function r^*.

Theorem 2 *(A) If $\mathcal{A}_2$ responds optimally and (B) if for all $\mathcal{T} : \mathcal{S} \times A_2 \to \Delta(\mathcal{S})$ there exists π^1 such that $\mathcal{P}_{\pi^1} \equiv \mathcal{T}$, then there exists a policy π^1 with optimal response π^2 such that the feasible set of reward functions under (π^1, π^2) is given by $\mathrm{Aff}(r^*)$, i.e. $\mathcal{R}((\pi^1, \pi^2)) = \mathrm{Aff}(r^*)$.*

This leads to the following corollary, which shows that it is possible to check in a single episode whether any given reward function is an affine transformation of r^*.

Corollary 2 *Under Assumptions (A) and (B) of Theorem 2, the learner can verify in any episode whether a reward function r is a positive affine transformation of the unknown and unobserved reward function r^*.*

We have shown that for any reward function r^* there exists an environment $\mathcal{T} : \mathcal{S} \times A_2 \to \Delta(\mathcal{S})$ such that the optimal policy with respect to $\mathcal{T}$ and r^* characterises r^* up to positive affine transformations (Theorem 2). This implied that the learner, without knowledge of r^*, can verify whether a reward function is element in $\mathrm{Aff}(r^*)$ by playing a specific policy (Corollary 2). However, the assumption that $\mathcal{A}_1$ can create any environment dynamics is very strong and we notice that while retrieving the set $\mathrm{Aff}(r^*)$ is clearly desirable, it is generally not necessary in order to cooperate optimally as other reward functions may also induce optimal behaviour. Thus, milder assumptions may be sufficient to learn about the reward function so that $\mathcal{A}_1$ is an optimal partner to $\mathcal{A}_2$. In the following, we propose an algorithm that learns about the reward function by adaptively designing environments and that is no-regret under mild assumptions.

4.2 An Algorithm for Interactive IRL

We now present an online algorithm for learning from and cooperating with an optimally responding agent $\mathcal{A}_2$ when agent $\mathcal{A}_1$ gets to observe the fully specified policy of $\mathcal{A}_2$ at the end of each episode. Note that we can always restrict the space of reward functions to the $|\mathcal{S}|$-dimensional unit simplex $\Delta(\mathcal{S})$ as any positive affine transformation of $r \in \Delta(\mathcal{S})$ is equivalent to r in the sense that they are feasible under the same observations and induce the same optimal (joint) policies (Lemma 1). Now, as the constraints characterising the feasible set $\mathcal{R}_t = \mathcal{R}((\pi^1_1, \pi^2_1), \dots, (\pi^1_t, \pi^2_t))$ are linear in the reward function (Corollary 1), we can use a Linear Program (LP) to find a reward function in $\mathcal{R}_t \cap \Delta(\mathcal{S})$. Let $\mathcal{C}((\pi^1_1, \pi^2_1), \dots, (\pi^1_t, \pi^2_t))$ denote the set of constraints induced by $(\pi^1_1, \pi^2_1), \dots, (\pi^1_t, \pi^2_t)$. In episode $t+1$, we then sample an $|\mathcal{S}|$-dimensional objective function c uniformly at random and solve the following LP:

$$\max_{r \in \Delta^{|S|}} c^\top r \text{ subject to } \mathcal{C}((\pi^1_1, \pi^2_1), \dots, (\pi^1_t, \pi^2_t)). \tag{1}$$

In the unlikely event that the LP computes the constant reward function in $\Delta(\mathcal{S})$, we resample the objective c and solve the LP again. Given a prospective reward function r, we then want to compute an optimal commitment strategy in $(\mathcal{S}, A_1, A_2, \mathcal{P}, r, \gamma)$. We see that if $\mathcal{A}_2$ responds optimally, it suffices to find an optimal joint policy as it yields an optimal commitment strategy for $\mathcal{A}_1$.

Lemma 2 *Let $(\bar{\pi}^1, \bar{\pi}^2)$ be an optimal joint policy. If agent $\mathcal{A}_2$ responds optimally to the commitment of $\mathcal{A}_1$, then $V_{\bar{\pi}^1, \pi^2(\bar{\pi}^1)} = V_{\bar{\pi}^1, \bar{\pi}^2}$. In particular, this entails that $\max_{\pi^1} V_{\pi^1, \pi^2(\pi^1)} = \max_{\pi^1, \pi^2} V_{\pi^1, \pi^2}$.*

2. We generally denote the true underlying reward function by r^*. Note that r^* is unknown to and unobserved by $\mathcal{A}_1$.

Algorithm 1 Interactive IRL via Linear Programming

input: $(\mathcal{S}, A_1, A_2, \mathcal{P}, \gamma)$, initial policy π_1^1
for $t = 1, 2, \dots$ **do**
  commit to policy π_t^1
  observe response π_t^2
  get constraints $\mathcal{C}_t = \mathcal{C}((\pi_1^1, \pi_1^2), \dots, (\pi_t^1, \pi_t^2))$
  sample objective vector c uniformly at random
  find solution $r_t \in \mathcal{R}_t$ of LP (1) for $\mathcal{C}_t$ and c
  compute $\pi_{t+1}^1 \in \Pi_1^{\mathrm{opt}}(r_t)$
end for

Note that an optimal joint policy and thus an optimal commitment strategy for $\mathcal{A}_1$ can be computed in time polynomial in the number of states and actions. In episode $t+1$, the algorithm then commits to a policy $\pi_{t+1}^1 \in \Pi_1^{\mathrm{opt}}(r)$, where r is the solution of the LP (1) and $\Pi_1^{\mathrm{opt}}(r)$ is the set of optimal commitment strategies under r. A description of this approach is given by Algorithm 1. In fact, we can show that Algorithm 1 is *no-regret* under the assumption that rewards that induce suboptimal joint policies are identifiable in the sense that these also induce suboptimal responses.

Proposition 1 *Suppose that for any non-constant reward function $r \in \Delta(\mathcal{S})$ it holds that if an optimal joint policy (π^1, π^2) under r is suboptimal under r^*, then in return there exists an optimal response $\pi^2(\pi^1)$ under r^* that is suboptimal under r. Moreover, assume that $\mathcal{A}_2$ responds optimally and breaks ties between equally good policies uniformly at random. Then, the average regret suffered by Algorithm 1 converges to zero almost surely.*

5. Cooperating with Suboptimal Agents

We now consider the case when $\mathcal{A}_2$ responds suboptimally according to some behavioural model such as Boltzmann-rational policies. Section 5.1 extends the Bayesian IRL formulation to our setting and Section 5.2 analyses the problem of computing optimal commitment strategies when $\mathcal{A}_2$ is playing suboptimally.

5.1 Learning from Suboptimal Responses

When demonstrations are possibly suboptimal, it is natural to take a Bayesian perspective (Ramachandran and Amir, 2007) as it provides a principled way to reason under uncertainty. Moreover, the Bayesian approach naturally extends to the partial information setting, where only trajectories generated by both agents' policies are available for learning. We assume that $\mathcal{A}_2$ responds with Boltzmann-rational policies with *unknown* inverse temperature β[3] and adapt the Bayesian IRL formulation to our setting. Suppose that $\mathcal{A}_1$ has observed $(\pi_1^1, \tau_1), \dots, (\pi_t^1, \tau_t)$, where τ_i is the trajectory generated by $\mathcal{A}_1$'s policy π_i^1 and $\mathcal{A}_2$'s response $\pi_i^2(\pi_i^1)$ for $i \in [t]$.[4] Bayesian IRL aims to estimate the posterior

$$\mathbb{P}(r, \beta \mid (\pi_1^1, \tau_1), \dots, (\pi_t^1, \tau_t)) = \frac{\mathbb{P}((\pi_1^1, \tau_1), \dots, (\pi_t^1, \tau_t) \mid r, \beta)\mathbb{P}(r)\mathbb{P}(\beta)}{\mathbb{P}((\pi_1^1, \tau_1), \dots, (\pi_t^1, \tau_t))},$$

given priors $\mathbb{P}(r)$ and $\mathbb{P}(\beta)$ over reward functions and inverse temperatures, respectively. We notice that the observations $(\pi_1^1, \tau_1), \dots, (\pi_t^1, \tau_t)$ are conditionally independent under measure $\mathbb{P}(\cdot \mid r, \beta)$. As a result, we can express their likelihood as $\mathbb{P}((\pi_1^1, \tau_1), \dots, (\pi_t^1, \tau_t) \mid r, \beta) = \prod_{i=1}^t \mathbb{P}((\pi_i^1, \tau_i) \mid r, \beta)$. The likelihood for each observation (π_i^1, τ_i) can then be computed as

$$\mathbb{P}((\pi_i^1, \tau_i) \mid r, \beta) = \prod_{h=0}^{H} \pi^2(b_{i,h} \mid s_{i,h}, \pi_i^1, r, \beta) \propto \exp\big(\beta \sum_{h=0}^{H} Q^*_{\pi_i^1}(s_{i,h}, b_{i,h}, r)\big).$$

3. Note that any other parameterised behavioural model could also be modeled by this Bayesian formulation.
4. For notational conciseness, we assume here that the length of a trajectory is fixed across all episodes.

Algorithm 2 Approximate Value Iteration for Boltzmann-Rational Responses

initialise V and $\hat{V}$
repeat until V converges:
for $s \in \mathcal{S}$ **do**
  for $(a, b) \in A_1 \times A_2$ **do**
    $\hat{Q}(s, a, b) = r(s) + \gamma \sum_{s'} \mathcal{P}(s'|s, a, b)\hat{V}(s')$
    $\pi^2(b \mid s, a) = \exp(\beta \hat{Q}(s, a, b))/Z$
  end for
  $\pi^1(s) = \arg\max_a \sum_{s'} \mathbb{E}_{b \sim \pi^2}[\mathcal{P}(s'|s, a, b)]V(s')$
  $V(s) = r(s) + \gamma \sum_{s'} \mathbb{E}_{b \sim \pi^2}[\mathcal{P}(s'|s, \pi^1(s), b)]V(s')$
  $\hat{V}(s) = \max_b \hat{Q}(s, \pi^1(s), b)$
end for

The Bayesian method we employ generates samples from the posterior via Markov Chain Monte Carlo (MCMC), similarly to (Ramachandran and Amir, 2007; Rothkopf and Dimitrakakis, 2011). At a high level, we employ a Metropolis-Hastings algorithm on the reward simplex, with a uniform prior distribution on the reward function and an exponential prior on the inverse temperature (see Algorithm 4 in Appendix D).

5.2 Planning with Suboptimal Agents

Prior work on computing optimal commitment strategies in stochastic games typically assumes that the other agent is responding optimally (Letchford et al., 2012). In this section, we analyse optimal commitment strategies for the cooperative Stackelberg game from Section 3 when $\mathcal{A}_2$ responds suboptimally, e.g. with Boltzmann-rational policies or ε-greedy policies. For this, the concept of dominating policies play a crucial role.

Definition 2 *A policy π^1 is* dominating *if $V_{\pi^1, \pi^2(\pi^1)}(s) \geq V_{\bar{\pi}^1, \pi^2(\bar{\pi}^1)}(s)$ for all policies $\bar{\pi}^1$ and states $s \in \mathcal{S}$.*

The existence of dominating policies is closely linked to our capacity to compute an optimal commitment strategy efficiently as it is a key requirement for dynamic programming. We show that if $\mathcal{A}_2$ plays proportionally with respect to the expected value of taking an action, a dominating policy for $\mathcal{A}_1$ may fail to exist.

Theorem 3 *If $\pi^2(b \mid s) \propto f(Q^*_{\pi^1}(s, b))$ for any strictly increasing function $f : [0, \infty) \to [0, \infty)$, then a dominating commitment strategy for agent $\mathcal{A}_1$ may not exist.*

In particular, this means that if $\mathcal{A}_2$ plays Boltzmann-rational policies, a dominating commitment strategy may fail to exist. Note that Theorem 3 generally only holds for *strictly* increasing functions f, as, for instance, there always exists a dominating commitment strategy when $\mathcal{A}_2$ plays uniformly at random. However, even for behavioural models as simple as ε-greedy, we see that a dominating commitment strategy does not necessarily exist.

Lemma 3 *If $\mathcal{A}_2$ plays ε-greedy, a dominating commitment strategy for $\mathcal{A}_1$ may not exist.*

Despite these difficulties, we provide algorithms to approximate optimal commitment strategies for the case of Boltzmann-rational responses (Algorithm 2) and ε-greedy responses (Algorithm 3 in Appendix C). The proposed methods can be viewed as approximate value iteration algorithms that keep track of two value functions, each modelling one agent.

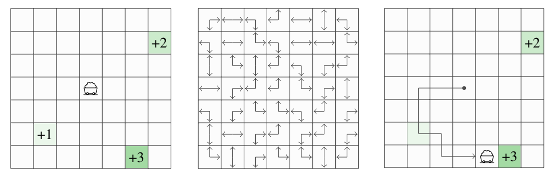
While Theorem 3 states that a dominating commitment strategy may not exist when $\mathcal{A}_2$ responds with Boltzmann-rational policies, the approximate value iteration algorithm for Boltzmann-rational responses described in Algorithm 2 acts as if a dominating commitment strategy does exist and could therefore converge to suboptimal solutions. However, it aims to account for the suboptimality of agent $\mathcal{A}_2$ and keeps track of two value functions, where V corresponds to what $\mathcal{A}_1$ believes to be the actual value given that $\mathcal{A}_2$ plays Boltzmann-rational policies, and $\hat{V}$ aims to approximate the belief of agent $\mathcal{A}_2$ about the value of the game. We include an empirical evaluation of the proposed algorithms in Appendix C.2, which demonstrates that accounting for the suboptimal nature of $\mathcal{A}_2$ reliably improves performance.

6. Experiments

In our experiments, we investigate how much the learner benefits from repeatedly interacting with the expert. To address this question and emphasise the potential benefit of demonstrations in different environments, we include the situation where $\mathcal{A}_1$ only observes the response of $\mathcal{A}_2$ to the initial policy π_1^1 played by $\mathcal{A}_1$. This resembles the standard IRL setting where the expert demonstrates a task in a fixed environment $(\mathcal{S}, A_2, \mathcal{P}_{\pi_1^1}, r, \gamma)$. Here, the initial policy π_1^1 is chosen uniformly at random. We model the standard IRL setting by repeatedly generating responses of $\mathcal{A}_2$ with respect to π_1^1, i.e. in the implied environment $\mathcal{P}_{\pi_1^1}$. Using these observations, we then estimate the reward function using standard IRL, compute the optimal policy with respect to the estimated rewards and evaluate the regret of this policy. In contrast, in the Interactive IRL setting, the learner gets to choose a different policy in subsequent episodes. In this case, we report the online regret of the actually played policies, i.e. the actual regret of the learner. More details are provided in Appendix D.

6.1 Environments

Maze-Maker. In this environment, agents $\mathcal{A}_1$ and $\mathcal{A}_2$ jointly control a cart in a 7×7 grid world. In this grid world, the doors leading from one cell to the neighbouring ones are locked. However, $\mathcal{A}_1$ can unlock exactly two doors at any time step before they fall shut again. Agent $\mathcal{A}_2$ can attempt to move the cart through a door to a neighbouring cell. However, when the door is locked, the cart stays where it was. We assume that any attempted move of the cart succeeds with probability 0.8 and that with probability 0.2 the cart moves to a random neighbouring cell. The agents are tasked with collecting three rewards of different value (+1, +2, +3), which disappear once collected. While the expert $\mathcal{A}_2$ knows where the rewards are placed, the helper $\mathcal{A}_1$ does not know their location. We model this environment as a two-agent MDP with 392 states (49×8) and discount factor $\gamma = 0.9$, where $\mathcal{A}_1$ has six actions (unlocking two out of four doors) and $\mathcal{A}_2$ four actions (moving the cart North, East, South, West). An illustration of the environment is given in Figure 2.



(a) (b) (c)

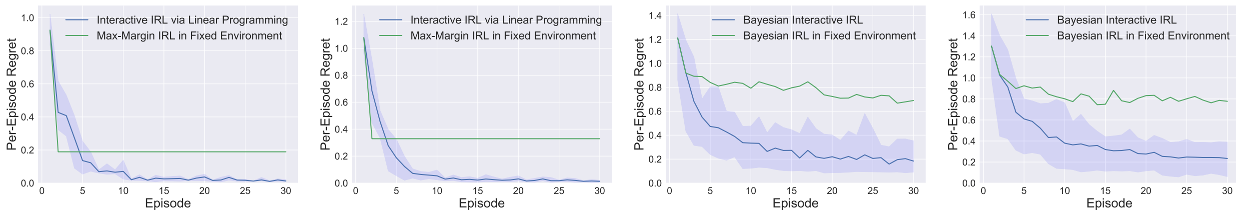
Figure 2: Maze-Maker Environment: (a) The initial game setup with initial position in the center and three rewards of different value scattered across the grid. (b) When $\mathcal{A}_1$ commits to a policy it implicitly creates a maze for $\mathcal{A}_2$ to navigate the cart in. (c) An exemplary path taken by $\mathcal{A}_2$ in the maze implied by $\mathcal{A}_1$'s policy.

Random MDPs. We also randomly generated MDPs with 200 states and four actions for each agent. We randomly draw the transition dynamics from a Dirichlet distribution, with restrictions on the influence of each agent on the transitions, and the rewards from an i.i.d. Beta distribution. The discount factor is set to $\gamma = 0.9$.

6.2 Results

Optimal Responses and Full Information: In Figure 3(a) and 3(b), we observe that the per-episode regret suffered by Algorithm 1 in both environments decreases notably with the number of episodes played. In particular, we see that after only a few episodes the per-episode regret of Algorithm 1 is significantly lower than for maximum-margin IRL (Ng and Russell, 2000) when $\mathcal{A}_1$ only observes the response to the initial policy π_1^1. This roughly corresponds to the standard IRL setting in which demonstrations are obtained in a single environment only. We thus find that the learner significantly benefits from observing $\mathcal{A}_2$'s behaviour in new and different environments, i.e. with respect to different policies of $\mathcal{A}_1$. In particular, it appears to be necessary to observe the expert's response to different policies of $\mathcal{A}_1$ in order to infer a (near-)optimal reward function. The results in both environments are averaged over 5 runs.

Suboptimal Responses and Partial Information Figure 3(c) and 3(d) show that Bayesian Interactive IRL (Algorithm 4) reliably improves its estimate of the true reward function with the number of episodes played and that the learner again substantially benefits from observing $\mathcal{A}_2$ act in different environments. While obtaining an increasing amount of trajectories in the same environment improves the estimate of the reward function as well, we see that trajectories generated in new environments, i.e. with respect to different policies of $\mathcal{A}_1$, yield much more information and thus allow for a better estimate of the unknown reward function. The results are averaged over 10 runs.



(a) Maze-Maker (b) Random MDPs (c) Maze-Maker (d) Random MDPs

Figure 3: (a) and (b): Optimal Responses and Full Information. Blue lines show the per-episode regret $\mathscr{L}(\pi_t^1)$ of Algorithm 1. Green lines show the regret of max-margin IRL (Ng and Russell, 2000) for observation (π_1^1, π_1^2) only. (c) and (d): Suboptimal Responses and Partial Information. Blue lines describe the per-episode regret of Bayesian Interactive IRL (Algorithm 4 in Appendix D). Green lines refer to Bayesian IRL in the fixed standard IRL setting.

7. Discussion and Future Work

We considered an interactive cooperation problem when the objective is unknown to one of the agents. This can be seen as a two-agent version of the IRL problem, where one agent is actively trying to infer the preferences of the other in order to cooperate. While the classical IRL problem is generally ill-posed, the interactive version that we study here can indeed be solved if the learning agent has sufficient power to affect the transitions. This is supported by both our experimental and theoretical results. In particular, the experiments clearly show that we can more accurately estimate the reward function (and hence collaborate more effectively) if we intelligently probe the other agent's responses.

An open theoretical question is whether upper and lower problem-dependent bounds on the episodic regret could be obtained in this setting. We presume that such bounds would involve a characterisation of $\mathcal{A}_1$'s power to affect the transitions. A natural extension of our setting would be the case where $\mathcal{A}_1$ does not reveal its policy to $\mathcal{A}_2$, but instead the latter simply observes the former's actions. In future work, it will also be interesting to construct interactive IRL algorithms that scale to large state spaces (or continuous domains) and test these in real-world applications.

Our observation that reward learning benefits from demonstrations under different environment dynamics also opens up a new and interesting perspective on IRL more generally. While current IRL methods still struggle to learn satisfactory reward functions in certain domains (even with abundant data), it could be promising to infer the reward function from demonstrations in slight variations of the target environment (when possible). Moreover, our results suggest that obtaining samples under new environment dynamics is generally more valuable than collecting additional samples from the same environment. Thus, such an approach could be useful in domains where resources are limited and samples expensive.

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first International Conference on Machine learning*, page 1, 2004.

Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113 (3):329–349, 2009.

Daniel S Brown and Scott Niekum. Machine teaching for inverse reinforcement learning: Algorithms and applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2019.

Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 82–90, 2006.

Christos Dimitrakakis, David C Parkes, Goran Radanovic, and Paul Tylkin. Multi-view decision processes: The helper-ai problem. In *Advances in Neural Information Processing Systems*, pages 5449–5458, 2017.

Ahana Ghosh, Sebastian Tschiatschek, Hamed Mahdavi, and Adish Singla. Towards deployment of robust ai agents for human-machine partnerships. *arXiv preprint arXiv:1910.02330*, 2019.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 3909–3917, 2016.

Hong Jun Jeon, Smitha Milli, and Anca Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. In *Advances in Neural Information Processing Systems*, pages 4415–4426, 2020.

Joshua Letchford, Liam MacDermed, Vincent Conitzer, Ronald Parr, and Charles L Isbell. Computing optimal strategies to commit to in stochastic games. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 1380–1386, 2012.

Xiaomin Lin, Stephen C Adams, and Peter A Beling. Multi-agent inverse reinforcement learning for certain general-sum stochastic games. *Journal of Artificial Intelligence Research*, 66:473–502, 2019.

R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 1959.

Sriraam Natarajan, Gautam Kunapuli, Kshitij Judah, Prasad Tadepalli, Kristian Kersting, and Jude Shavlik. Multi-agent inverse reinforcement learning. In *2010 Ninth International Conference on Machine Learning and Applications*, pages 395–400, 2010.

Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, page 2, 2000.

Stefanos Nikolaidis and Julie Shah. Human-robot cross-training: Computational formulation, modeling and evaluation of a human team training strategy. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '13, page 33–40, 2013.

Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay Vazirani. *Algorithmic Game Theory*. Cambridge University Press, 2007.

Kamalaruban Parameswaran, Devidze Rati, Volkan Cevher, and Singla Adish. Interactive teaching algorithms for inverse reinforcement learning. In *28th International Joint Conference on Artificial Intelligence, 2019.*, number CONF, 2019.

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

Goran Radanovic, Rati Devidze, David Parkes, and Adish Singla. Learning to collaborate in markov decision processes. In *International Conference on Machine Learning*, pages 5261–5270, 2019.

Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 2586–2591, 2007.

Nathan D Ratliff, J Andrew Bagnell, and Martin A Zinkevich. Maximum margin planning. In *Proceedings of the 23rd International Conference on Machine learning*, pages 729–736, 2006.

Constantin A Rothkopf and Christos Dimitrakakis. Preference elicitation and inverse reinforcement learning. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 34–48, 2011.

Stuart Russell. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on computational learning theory*, pages 101–103, 1998.

Haoqi Zhang and David Parkes. Enabling environment design via active indirect elicitation. In *4th Multidisciplinary Workshop on Advances in Preference Handling*, 2008.

Haoqi Zhang, David C. Parkes, and Yiling Chen. Policy teaching through reward function learning. EC '09, page 295–304, New York, NY, USA, 2009. Association for Computing Machinery.

Appendix A. Proofs for Section 4

A.1 Proof of Theorem 1

Theorem 1 *Let there be some MDP without reward function $(\mathcal{S}, A_1, A_2, \mathcal{P}, \gamma)$. A reward function r is feasible under policies π^1 and π^2 if and only if*

$$\left(\mathcal{P}_{\pi^1,\pi^2} - \mathcal{P}_{\pi^1,b}\right)\left(I - \gamma \mathcal{P}_{\pi^1,\pi^2}\right)^{-1} r \succeq 0 \quad \forall b \in A_2,$$

where $\mathcal{P}_{\pi^1,b}$ is the one-step transition matrix under policy π^1 and action $b \in A_2$.

Proof [Proof of Theorem 1] Substituting transition matrix $\mathcal{P}$ by $\mathcal{P}_{\pi^1}$ in the proof by Ng and Russell (2000) readily implies Theorem 1. Note that if $\pi^2(s) = \bar{b}$ for all $s \in \mathcal{S}$, the inequality vacuously holds for $b = \bar{b}$. Thus, in general we obtain $|A_2| - 1$ many of the above vector inequalities. ■

A.2 Proof of Lemma 1

Lemma 1 *If $\mathcal{A}_2$ responds optimally to the commitment of $\mathcal{A}_1$, any reward function r is indistinguishable from its positive affine transformations, i.e. r is feasible iff every $\bar{r} \in \mathcal{A}(r)$ is feasible.*

Proof [Proof of Lemma 1] We write $V_{\pi^1,\pi^2}(r)$ for the value function under joint policy (π^1, π^2) and reward function r. The Bellman equation tells us that the value function under (π^1, π^2) and reward function $\lambda_1 r + \lambda_2 \mathbf{1} \in \mathrm{Aff}(r)$ is given by

$$V_{\pi^1,\pi^2}(\lambda_1 r + \lambda_2 \mathbf{1}) = (I - \gamma \mathcal{P}_{\pi^1,\pi^2})^{-1}(\lambda_1 r + \lambda_2 \mathbf{1}).$$

Now, since $\mathcal{P}_{\pi^1,\pi^2}$ is a stochastic matrix, it is easy to check that $(I - \gamma \mathcal{P}_{\pi^1,\pi^2})^{-1}\mathbf{1} = (1-\gamma)^{-1}\mathbf{1}$. It then follows that

$$V_{\pi^1,\pi^2}(\lambda_1 r + \lambda_2 \mathbf{1}) = \lambda_1 V_{\pi^1,\pi^2}(r) + K,$$

where $K = \lambda_2(1-\gamma)^{-1}\mathbf{1}$. Hence, we find that any policy π^2 that maximises $V_{\pi^1,\pi^2}(r)$ also maximises $V_{\pi^1,\pi^2}(\lambda_1 r + \lambda_2 \mathbf{1})$ for $\lambda_1 \geq 0$ and $\lambda_2 \in \mathbb{R}$, and vice versa. This means that r is feasible if and only if every $\bar{r} \in \mathrm{Aff}(r)$ is feasible. ■

A.3 Proof of Theorem 2

Theorem 2 *(A) If $\mathcal{A}_2$ responds optimally and (B) if for all $\mathcal{T} : \mathcal{S} \times A_2 \to \Delta(\mathcal{S})$ there exists π^1 such that $\mathcal{P}_{\pi^1} \equiv \mathcal{T}$, then there exists a policy π^1 with optimal response π^2 such that the feasible set of reward functions under (π^1, π^2) is given by $\mathrm{Aff}(r^*)$, i.e. $\mathcal{R}((\pi^1, \pi^2)) = \mathrm{Aff}(r^*)$.*

For the proof of Theorem 2, we will need the following technical lemma.

Lemma A.1 *Any (two-dimensional) plane $\mathcal{R} \subseteq \mathbb{R}^N$ can be uniquely characterized by the intersection of $N-1$ many half-spaces $H_i = \{x \in \mathbb{R}^N : \varphi_i^\top x \geq 0\}$, where $\varphi_1, \ldots, \varphi_{N-1} \in \mathbb{R}^N$ are vectors orthogonal to $\mathcal{R}$.*

Proof [Proof of Lemma A.1] W.lo.g. let $\mathcal{R}$ be some plane in $\mathbb{R}^N$ through the origin. Let the vectors v_1 and v_2 denote an orthogonal basis of $\mathcal{R}$, i.e. $\mathcal{R} = \{\lambda_1 v_1 + \lambda_2 v_2 \colon \lambda_1, \lambda_2 \in \mathbb{R}\}$ and $v_1^\top v_2 = 0$. We can then find vectors $\varphi_1, \ldots, \varphi_{N-2}$ such that $\{\varphi_1, \ldots, \varphi_{N-2}, v_1, v_2\}$ forms an orthogonal basis of $\mathbb{R}^N$. In particular, we then have $\varphi_i^\top x = 0$ for all $x \in \mathcal{R}$ and $i \in [N-2]$. Moreover, we define the vector

$$\varphi_{N-1} = -(\varphi_1 + \cdots + \varphi_{N-2})$$

and note that φ_{N-1} is orthogonal to $\mathcal{R}$ as well. Let the half-spaces induced by vectors $\varphi_1, \dots, \varphi_{N-1}$ be given by $H_i = \{x \in \mathbb{R}^N : \varphi_i^\top x \geq 0\}$ for $i \in [N-1]$. We now show that $H_1 \cap \dots \cap H_{N-1} = \mathcal{R}$.

We begin by verifying that $H_1 \cap \dots \cap H_{N-1} \subseteq \mathcal{R}$. Suppose this is not true and there exists a vector $w \notin \mathcal{R}$ such that $\varphi_i^\top w \geq 0$ for all $i \in [N-1]$, i.e. $w \in H_1 \cap \dots \cap H_{N-1}$. Then, we must have $\varphi_j^\top w > 0$ for some $j \in [N-2]$ as the orthogonal complement of $\mathrm{span}(\varphi_1, \dots, \varphi_{N-2})$ is given by $\mathcal{R}$ and we assumed $w \notin \mathcal{R}$. By definition of φ_{N-1}, we have $\varphi_1 + \dots + \varphi_{N-1} = 0$ and thus, $(\varphi_1 + \dots + \varphi_{N-1})^\top w = 0$. However, it also holds that

$$\varphi_1^\top w + \dots + \varphi_{N-1}^\top w > 0,$$

since $\varphi_i^\top w \geq 0$ for $i \in [N-1]$ and $\varphi_j^\top w > 0$ for some $j \in [N-2]$. Thus, such w cannot exist and we have shown that $H_1 \cap \dots \cap H_{N-1} \subseteq \mathcal{R}$. Finally, the relation $\mathcal{R} \subseteq H_1 \cap \dots \cap H_{N-1}$ also holds as $\varphi_1, \dots, \varphi_{N-1}$ are chosen orthogonal to $\mathcal{R}$ and thus, $\varphi_i^\top x = 0$ for all $i \in [N-1]$ and $x \in \mathcal{R}$.

Note that we can analogously prove that any line $\mathcal{C} = \{\lambda v : \lambda \in \mathbb{R}\}$ in $\mathbb{R}^N$ can be uniquely characterised by N half-spaces. In this case, we can find an orthogonal basis $\{\varphi_1, \dots, \varphi_{N-1}, v\}$ and define $\varphi_N = -(\varphi_1 + \dots + \varphi_{N-1})$. The remainder of the proof then follows the same line of argument as before. ■

Proof [Proof of Theorem 2] Let $N = |\mathcal{S}|$. We will now show that under the assumptions of Theorem 2, there exists a policy π^1 with optimal response π^2 so that only positive affine transformations of r^* are feasible under observation (π^1, π^2), i.e. $\mathcal{R}((\pi^1, \pi^2) = \mathrm{Aff}(r^*)$.

First we observe that we can w.l.o.g. assume only two actions for $\mathcal{A}_2$, i.e. $|A_2| = 2$. To see this suppose that $|A_2| > 2$ and consider an action space $A_2' \subset A_2$ with $|A_2'| \geq 2$ and transition kernel $\mathcal{P}'_{\pi^1} : \mathcal{S} \times A_2' \to \Delta(\mathcal{S})$ defined as $\mathcal{P}'_{\pi^1}(\cdot \mid s, b) = \mathcal{P}_{\pi^1}(\cdot \mid s, b)$ for $b \in A_2'$. If $\pi^2(s) \in A_2'$ for all $s \in \mathcal{S}$, then the feasible set under action space A_2 is subset of the feasible set under action space A_2'. Thus, we can assume w.l.o.g. that $A_2 = \{b_1, b_2\}$. From hereon out, we assume that the true reward function r^* is non-constant. The special case of a constant true reward function is addressed at the end.

We first construct an orthogonal basis $\{\varphi_1, \dots, \varphi_N\}$ such that the corresponding half-spaces characterise $\mathrm{Aff}(r^*$ and then show that there exists π_1 such that

$$(\mathcal{P}_{\pi^1,b_1} - \mathcal{P}_{\pi^1,b_2})(I - \gamma \mathcal{P}_{\pi^1,b_1})^{-1} = (\varphi_1, \dots, \varphi_N)^\top.$$

For non-constant r^* we have that $\mathcal{R} \triangleq \mathrm{span}(r^*, \mathbf{1})$ describes a plane in $\mathbb{R}^N$ and $\mathrm{Aff}(r^*) \subset \mathcal{R}$. By Lemma A.1, there exist vectors $\varphi_1, \dots, \varphi_{N-1} \in \mathbb{R}^N$ such that $\varphi_i^\top x = 0$ for all $x \in \mathcal{R}$ and $H_1 \cap \dots \cap H_{N-1} = \mathcal{R}$ with $H_i = \{x \in \mathbb{R}^N : \varphi_i^\top x \geq 0\}$. In particular, it holds that $\varphi_i^\top \mathbf{1} = 0$, i.e. $\|\varphi_i\|_1 = 0$ for all $i \in [N-1]$.

Now, let us consider the orthogonal projection of r^* given by $r^* = \alpha \mathbf{1} + w$ for $\alpha \in \mathbb{R}$ and $w \in \mathbb{R}^N$ with $w^\top \mathbf{1} = 0$. It follows that $w^\top r^* = w^\top(\alpha \mathbf{1} + w) = w^\top w > 0$, since r^* is non-constant and thus, $w \neq \mathbf{0}$. Let us define $\varphi_N = \eta w$ for some scalar $\eta > 0$. Then, we have $\varphi_N^\top x \geq 0$ for all $x \in \{\lambda_1 r^* + \lambda_2 \mathbf{1} : \lambda \geq 0, \lambda_2 \in \mathbb{R}\}$, since $w^\top r^* > 0$ and $w^\top \mathbf{1} = 0$. Similarly, we have $\varphi_N^\top \hat{x} < 0$ for all $\hat{x} \in \{\lambda_1 r^* + \lambda_2 \mathbf{1} : \lambda_1 < 0, \lambda_2 \in \mathbb{R}\}$. It then follows that

$$H_1 \cap \dots \cap H_N = \mathcal{R} \cap H_N = \mathrm{Aff}(r^*),$$

where $H_N = \{x \in \mathbb{R}^N : \varphi_N^\top x \geq 0\}$. Note that every φ_i with $i \in [N]$ satisfies $\|\varphi_i\|_1 = 0$ and that the half-spaces H_i are invariant under positive linear transformation of φ_i. We can therefore assume that $\varphi_1, \dots, \varphi_N$ take values in $[\frac{1}{N} - 1, \frac{1}{N}]$. We denote with $\Phi = (\varphi_1, \dots, \varphi_N)^\top$ the matrix with rows $\varphi_1, \dots, \varphi_N$.

Recall that $A_2 = \{b_1, b_2\}$. We will now show that there exists a policy π_1 such that

$$(\mathcal{P}_{\pi^1,b_1} - \mathcal{P}_{\pi^1,b_2})(I - \gamma \mathcal{P}_{\pi^1,b_1})^{-1} = \Phi.$$

By assumption, there exists a π^1 such that $\mathcal{P}_{\pi^1,b_1} \equiv B_1$ and $\mathcal{P}_{\pi^1,b_2} \equiv B_2$ for any two stochastic matrices B_1 and B_2. We set $\mathcal{P}_{\pi^1,b_1}(s' \mid s) = \frac{1}{N}$ for all $s, s' \in \mathcal{S}$, which yields

$$\Phi(I - \gamma \mathcal{P}_{\pi^1,b_1}) = \Phi - \gamma \Phi \mathcal{P}_{\pi^1,b_1} = \Phi, \tag{2}$$

since $\|\varphi_i\|_1 = 0$ for all $i \in [N]$ and $\mathcal{P}_{\pi^1,b_1}$ is a constant matrix. Now, set $\mathcal{P}_{\pi^1,b_2} \equiv \mathcal{P}_{\pi^1,b_1} - \Phi$ and note that since $\|\varphi_i\|_1 = 0$ for all $i \in [N]$, the matrix $\mathcal{P}_{\pi^1,b_2}$ is indeed stochastic. It then follows that

$$(\mathcal{P}_{\pi^1,b_1} - \mathcal{P}_{\pi^1,b_2})(I - \gamma \mathcal{P}_{\pi^1,b_1})^{-1} = \Phi(I - \gamma \mathcal{P}_{\pi^1,b_1})^{-1} = \Phi,$$

by equation (2). Note that this means that indeed action b_1 is the optimal response to policy π^1 as $\Phi r^* \succeq 0$ by construction of Φ.[5] Therefore, from Theorem 1 it follows that any feasible reward function r must satisfy

$$(\mathcal{P}_{\pi^1,b_1} - \mathcal{P}_{\pi^1,b_2})(I - \gamma \mathcal{P}_{\pi^1,b_1})^{-1} r = \Phi r \succeq 0,$$

i.e. $\varphi_i^\top r \geq 0$ for all $i \in [N]$. Hence, any feasible reward function must be in $H_1 \cap \cdots \cap H_N$ and thus element in $\mathrm{Aff}(r^*)$. So, we have shown that the feasible set of reward functions under π^1 with response $\pi^2 \equiv b_1$ is given by $\mathrm{Aff}(r^*)$.

In the special case of the constant reward function r^*, we have that the set $\mathrm{Aff}(r^*) = \{\lambda \mathbf{1} : \lambda \in \mathbb{R}\}$ becomes not a plane, but a line in $\mathbb{R}^N$. The proof for this case then progresses similarly to the proof above with the difference that we describe $\mathrm{Aff}(r^*)$ by N many half-spaces and that there is no need to consider the orthogonal projection of r^* as done before.

■

A.4 Proof of Corollary 2

Corollary 2 *Under Assumptions (A) and (B) of Theorem 2, the learner can verify in any episode whether a reward function r is a positive affine transformation of the actual and unknown reward function r^*.*

Proof Recall that it follows from Lemma 1 that $\mathrm{Aff}(r^*) \subseteq \mathcal{R}((\pi^1, \pi^2))$ for any policy π^1 with optimal response π^2. In other words, the positive affine transformations of the unknown reward function r^* are always feasible as r^* is always feasible. Now, let $r \in \mathbb{R}^{|\mathcal{S}|}$ be some reward function and suppose that $\mathcal{A}_1$ plays the "ideal" policy π^1 with respect to r as it is constructed in the proof of Theorem 2. Let π^2 be an optimal response to π^1. It follows from the combination of Lemma 1 and Theorem 2 that $\mathcal{R}((\pi^1, \pi^2)) = \mathrm{Aff}(r)$ if and only if $r \in \mathrm{Aff}(r^*)$. Now, using linear programming, we can check whether $\mathcal{R}((\pi^1, \pi^2)) = \mathrm{Aff}(r)$ holds true. If $\mathcal{R}((\pi^1, \pi^2)) = \mathrm{Aff}(r)$, we know that r must be a positive affine transformation of r^*. On the other hand, if we observe $\mathcal{R}((\pi^1, \pi^2)) \neq \mathrm{Aff}(r)$, then r cannot be element in $\mathrm{Aff}(r^*)$.

■

A.5 Proof of Lemma 2

Lemma 2 *Let $(\bar{\pi}^1, \bar{\pi}^2)$ be an optimal joint policy. If $\mathcal{A}_2$ responds optimally to the commitment of $\mathcal{A}_1$, then $V_{\bar{\pi}^1, \pi^2(\bar{\pi}^1)} = V_{\bar{\pi}^1, \bar{\pi}^2}$. In particular, this entails that $\max_{\pi^1} V_{\pi^1, \pi^2(\pi^1)} = \max_{\pi^1, \pi^2} V_{\pi^1, \pi^2}$.*

Proof [Proof of Lemma 2] Let $(\bar{\pi}^1, \bar{\pi}^2) \in \arg\max_{\pi^1, \pi^2} V_{\pi^1, \pi^2}$. Suppose $\mathcal{A}_1$ commits to $\bar{\pi}^1$. Then, $\mathcal{A}_2$ responds with $\pi^2(\bar{\pi}^1)$ such that $V_{\bar{\pi}^1, \pi^2(\bar{\pi}^1)} \succeq V_{\bar{\pi}^1, \pi^2}$ for all π^2 by optimality of $\mathcal{A}_2$. Now, since $V_{\bar{\pi}^1, \bar{\pi}^2} \succeq \max_{\pi^1} V_{\pi^1, \pi^2(\pi^1)}$ always, we also have

$$\max_{\pi^1} V_{\pi^1, \pi^2(\pi^1)} \succeq V_{\bar{\pi}^1, \pi^2(\bar{\pi}^1)} \succeq V_{\bar{\pi}^1, \bar{\pi}^2} \succeq \max_{\pi^1} V_{\pi^1, \pi^2(\pi^1)}.$$

Thus, $\max_{\pi^1} V_{\pi^1, \pi^2(\pi^1)} = V_{\bar{\pi}^1, \bar{\pi}^2} = \max_{\pi^1, \pi^2} V_{\pi^1, \pi^2}$. In other words, Lemma 2 states that the optimal joint policy yields an optimal commitment strategy for $\mathcal{A}_1$ when $\mathcal{A}_2$ responds optimally. ■

A.6 Proof of Proposition 1

Proposition 1 *Suppose that for any non-constant reward function $r \in \Delta(\mathcal{S})$ it holds that if an optimal joint policy (π^1, π^2) under r is suboptimal under r^*, then in return there exists an optimal response $\pi^2(\pi^1)$ under r^* that is*

5. This can, for instance, be verified using Theorem 1.

suboptimal under r. Moreover, assume that $\mathcal{A}_2$ responds optimally and breaks ties between equally good policies uniformly at random. Then, the average regret suffered by Algorithm 1 converges to zero almost surely.

For the proof of Proposition 1, we will need the following sets: Let $\Pi^{\text{opt}}(r)$ denote the set of optimal joint policies under reward function r, i.e. the set of optimal joint policies in the MDP $(\mathcal{S}, A_1, A_2, \mathcal{P}, r, \gamma)$. Further, we denote the set of optimal responses under policy π^1 and reward function r by $\Pi_2^{\text{opt}}(r, \pi^1)$. A key object of interest is the following set of reward functions. Let $\mathcal{O}$ be the set of reward functions in $\Delta(\mathcal{S})$ that always induce an optimal joint policy, i.e.

$$\mathcal{O} = \{r \in \Delta(\mathcal{S}) : \Pi^{\text{opt}}(r) \subseteq \Pi^{\text{opt}}(r^*)\}.$$

Note that by Lemma 2 any optimal joint policy yields an optimal commitment strategy for agent $\mathcal{A}_1$, i.e. any $r \in \mathcal{O}$ induces an optimal commitment strategy. We can easily check that $\mathcal{O}$ is a convex set.

Lemma A.2 *The set $\mathcal{O}$ is convex.*

Proof [Proof of Lemma A.2] Let $r_1, r_2 \in \mathcal{O}$. We show that $\lambda r_1 + (1-\lambda) r_2 \in \mathcal{O}$ for any $\lambda \in [0,1]$. Recall that the value function $V_\pi(r) = (I - \gamma \mathcal{P}_\pi)^{-1} r$ is linear in r and we therefore have $V_\pi(\lambda r_1 + (1-\lambda) r_2) = \lambda V_\pi(r_1) + (1-\lambda) V_\pi(r_2)$. In a first step, we prove $\Pi^{\text{opt}}(\lambda r_1 + (1-\lambda) r_2) \subseteq \Pi^{\text{opt}}(r^*)$. Let $\pi \in \Pi^{\text{opt}}(\lambda r_1 + (1-\lambda) r_2)$. Then, for all policies ν it must hold that

$$V_\pi(\lambda r_1 + (1-\lambda) r_2) \succeq V_\nu(\lambda r_1 + (1-\lambda) r_2), \tag{3}$$

where $\succ$ denotes element-wise inequality. Now, suppose that $\pi \notin \Pi^{\text{opt}}(r^*)$. It follows that $V_\pi(r_1) \preceq V_\nu(r_1)$ and $V_\pi(r_2) \preceq V_\nu(r_2)$ for some $\nu \in \Pi^{\text{opt}}(r^*) = \Pi^{\text{opt}}(r_1) = \Pi^{\text{opt}}(r_2)$ with strict inequality for at least one $s \in \mathcal{S}$. This contradicts equation (3) and it follows that $\Pi^{\text{opt}}(\lambda r_1 + (1-\lambda) r_2) \subseteq \Pi^{\text{opt}}(r^*)$. We will now verify the relation $\Pi^{\text{opt}}(r^*) \subseteq \Pi^{\text{opt}}(\lambda r_1 + (1-\lambda) r_2)$. For any $\pi \in \Pi^{\text{opt}}(r^*)$, we have $V_\pi(r_1) \succeq V_\nu(r_1)$ and $V_\pi(r_2) \succeq V_\nu(r_2)$ for all policies ν. It then directly follows that $\pi \in \Pi^{\text{opt}}(\lambda r_1 + (1-\lambda) r_2)$ and thus, $\Pi^{\text{opt}}(r^*) \subseteq \Pi^{\text{opt}}(\lambda r_1 + (1-\lambda) r_2)$, i.e. $\lambda r_1 + (1-\lambda) r_2 \in \mathcal{O}$. ■

Interestingly, Lemma A.2 implies that the set of reward functions that induce an optimal commitment strategy is a connected set. We will now prove Proposition 1.

Proof [Proof of Proposition 1] As Algorithm 1 only considers reward functions in the simplex $\Delta(\mathcal{S})$, we will simply write $\mathcal{R}_t$ instead of $\mathcal{R}_t \cap \Delta(\mathcal{S})$ for notational convenience.

In episode t, Algorithm 1 chooses a vertex of the set of feasible solutions of the linear program, i.e. a reward function $r_t \in \mathcal{R}_t$. Note that by construction of Algorithm 1 we never select the constant reward function in $\Delta(\mathcal{S})$. For any $r_t \in \mathcal{R}_t$ obtained from the LP (1) with uniformly random objective function c there are two possible cases: $r_t \in \mathcal{O}$ or $r_t \notin \mathcal{O}$. If $r_t \in \mathcal{O}$, then r_t induces an optimal joint policy, i.e. an optimal commitment strategy by Lemma 2. Accordingly, Algorithm 1 commits to an optimal commitment strategy and thus suffers zero regret in episode $t+1$. We want to highlight that the proof does not require that the objective function in Algorithm 1 is being chosen in a randomised fashion. However, randomising the choice of the objective improved exploration in our experiments.

In the following, we show that for the case of $r_t \notin \mathcal{O}$, Algorithm 1 strictly decreases the set of feasible reward functions with positive probability. In order to show this, we first construct a finite cover of $\Delta(\mathcal{S})$. Let Π_1 and Π_2 denote the sets of deterministic policies for $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively.[6] Note that both Π_1 and Π_2 are finite as we assumed finite action spaces A_1 and A_2. Let 2^{Π_2} denote the power set of Π_2. For $\pi^1 \in \Pi_1$ and $\bar{\Pi}_2 \in 2^{\Pi_2}$, we define

$$B(\pi^1, \bar{\Pi}_2) = \{r \in \Delta(\mathcal{S}) : \bar{\Pi}_2 = \Pi_2^{\text{opt}}(r, \pi^1)\}.$$

The set $B(\pi^1, \bar{\Pi}_2)$ thus describes the reward functions that make the policies in $\bar{\Pi}_2$ optimal in response to π^1. Indeed, for any fixed $\pi^1 \in \Pi_1$, the collection $\mathcal{B}(\pi^1) = \{B(\pi^1, \bar{\Pi}_2) : \bar{\Pi}_2 \in 2^{\Pi_2}\}$ forms a finite partition of $\Delta(\mathcal{S})$

$$\bigcup\nolimits_{\bar{\Pi}_2 \in 2^{\Pi_2}} B(\pi^1, \bar{\Pi}_2) = \Delta(\mathcal{S}),$$

6. We assume here that $\mathcal{A}_2$ responds with deterministic policies in order to keep the proof as comprehensible as possible. However, this assumption can be dropped as we can still give a finite partition of $\Delta(\mathcal{S})$ when $\mathcal{A}_2$ also responds with optimal stochastic policies.

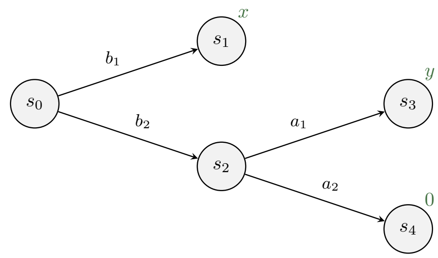


Figure 4: Counterexample. All transitions are deterministic. The action of $\mathcal{A}_2$ alone determines the transitions from state s_0 to states s_1 and s_2, whereas in state s_2 only the action of $\mathcal{A}_1$ affects transitions. The green x, y and 0 denote the rewards obtained in states s_1, s_3, and s_4, respectively. States s_0 and s_2 yield zero reward.

as for any $r \in \Delta(\mathcal{S})$ there always exists at least one deterministic optimal policy in the MDP $(\mathcal{S}, A_2, \mathcal{P}_{\pi^1}, r, \gamma)$ (Puterman, 2014). In other words, for any $\pi^1 \in \Pi_1$, we partition $\Delta(\mathcal{S})$ into sets that induce the same set of optimal responses to π^1. Naturally, due to $\mathcal{B}(\pi^1)$ being a *finite* partition of $\Delta(\mathcal{S})$ for any π^1, the Lebesgue-measure for all but finitely many $B(\pi^1, \bar{\Pi}_2)$ must be larger than some constant $\varepsilon > 0$.

We now show that if $r_t \notin \mathcal{O}$, then with positive probability the set of feasible solutions is decreased by at least ε. If $r_t \notin \mathcal{O}$, then Algorithm 1 computes an optimal commitment strategy $\pi^1_{t+1} \in \Pi_1^{\text{opt}}(r_t)$ (by computing the optimal joint policy under r_t, see Lemma 2), which may be suboptimal under r^*, i.e. $\pi^1_{t+1} \notin \Pi_1^{\text{opt}}(r^*)$. Now, if π^1_{t+1} is suboptimal under r^*, then by assumption[7] there exists an optimal response $\pi^2_{t+1} \in \Pi_2^{\text{opt}}(r^*, \pi^1_{t+1})$ that is suboptimal under r_t, i.e. $\pi^2_{t+1} \notin \Pi_2^{\text{opt}}(r_t, \pi^1_{t+1})$. Recall that by our assumption $\mathcal{A}_2$ selects its response uniformly at random from $\Pi_2^{\text{opt}}(r^*, \pi^1_{t+1})$. Since $\Pi_2^{\text{opt}}(r^*, \pi^1_{t+1})$ is finite, $\mathcal{A}_2$ will respond with $\pi^2_{t+1} \notin \Pi_2^{\text{opt}}(r_t, \pi^1_{t+1})$ with positive probability.

In that case, after observing π^2_{t+1} the reward function r_t cannot be feasible anymore, i.e. $r_t \notin \mathcal{R}_{t+1}$. In addition, we then also have that $B(\pi^1_{t+1}, \Pi_2^{\text{opt}}(r_t, \pi^1_{t+1})) \cap \mathcal{R}_{t+1} = \emptyset$, as all reward functions in $B(\pi^1_{t+1}, \Pi_2^{\text{opt}}(r_t, \pi^1_{t+1}))$ induce the same optimal responses $\Pi_2^{\text{opt}}(r_t, \pi^1_{t+1})$ and π^2_{t+1} is not in $\Pi_2^{\text{opt}}(r_t, \pi^1_{t+1})$. In other words, any $r \in B(\pi^1_{t+1}, \Pi_2^{\text{opt}}(r_t, \pi^1_{t+1}))$ cannot satisfy the constraints of Corollary 1.

As seen before, for all but finitely many $\bar{\Pi}_2 \in 2^{\Pi_2}$ we have $\lambda(B(\pi^1, \bar{\Pi}_2)) > \varepsilon$, where λ is the Lebesgue-measure. As a consequence, if $r_t \notin \mathcal{O}$, then we have for all but finitely many cases that $\lambda(\mathcal{R}_{t+1}) \leq \lambda(\mathcal{R} \backslash B(\pi^1_{t+1}, \Pi_2^{\text{opt}}(r_t, \pi^1_{t+1})) \leq \lambda(\mathcal{R}_t) - \varepsilon$.

Therefore, every time when Algorithm 1 chooses a reward function $r_t \notin \mathcal{O}$[8] inducing a suboptimal commitment strategy, (with positive probability) r_t will not be feasible anymore and (except for finitely many times) we reduce the size of the feasible set by at least the constant amount ε. As a result, the feasible set of reward function $\mathcal{R}_t$ will eventually become smaller than or equal to $\mathcal{O}$, i.e. $\mathcal{R}_t \subseteq \mathcal{O}$. Consequently, Algorithm 1 will almost surely converge to choosing only reward function in $\mathcal{O}$ and will thus only play optimal commitment strategies. ■

Appendix B. Proofs for Section 5

B.1 Proof of Theorem 3

Theorem 3 *If $\pi^2(b \mid s) \propto f(Q^*_{\pi^1}(s, b))$ for any strictly increasing function $f : [0, \infty) \to [0, \infty)$, then a dominating commitment strategy for agent $\mathcal{A}_1$ may not exist.*

7. Note that if π^1 is a suboptimal commitment strategy, then the joint policy (π^1, π^2) is suboptimal for any π^2.
8. Recall that the special case of the constant reward function (which is not in $\mathcal{O}$) can be ignored.

Proof [Proof of Theorem 3] We provide a problem instance for which there exists no dominating policy for any strictly increasing function $f : [0, \infty) \to [0, \infty)$. Consider the two-agent MDP in Figure 4. We omitted consecutive transitions in Figure 4, but assume that states s_1, s_3, and s_4 lead to the same (terminal) state with probability one.

We will show that the strictly optimal policy when in state s_0 is strictly suboptimal when in state s_2 for specific choices of $x > 0$ and $y > 0$. For simplicity, we omit the discount factor γ in the following.

$\mathcal{A}_1$ only influences transitions in state s_2 and thus there are essentially only two deterministic policies for $\mathcal{A}_1$, namely π^1 with $\pi^1(s_2) = a_1$ and $\bar{\pi}^1$ with $\bar{\pi}^1(s_2) = a_2$. Since $y > 0$, action a_1 is optimal in state s_2 and so π^1 is the optimal policy in state s_2. We now show that there exists $x, y > 0$ such that $V_{\pi^1,\pi^2(\pi^1)}(s_0) < V_{\bar{\pi}^1,\pi^2(\bar{\pi}^1)}(s_0)$, i.e. $\bar{\pi}^1$ is strictly better than π^1 when in state s_0.

Omitting the discount factor, we have $Q^*_{\pi^1}(s_0, b_1) = x$ and $Q^*_{\pi^1}(s_0, b_2) = y$ as well as $Q^*_{\bar{\pi}^1}(s_0, b_1) = x$ and $Q^*_{\bar{\pi}^1}(s_1, b_2) = 0$. We therefore want to show that there exist $x, y > 0$ such that

$$
\begin{aligned}
V_{\pi^1,\pi^2(\pi^1)}(s_1) &= x\,\frac{f(x)}{f(x)+f(y)} + y\,\frac{f(y)}{f(x)+f(y)} \\
&< x\,\frac{f(x)}{f(x)+f(0)} = V_{\bar{\pi}^1,\pi^2(\bar{\pi}^1)}(s_1).
\end{aligned}
$$

Suppose the contrary is true. Then, for all $x, y > 0$ it must hold that

$$
\begin{aligned}
& x\,\frac{f(x)}{f(x)+f(y)} + y\,\frac{f(y)}{f(x)+f(y)} \geq x\,\frac{f(x)}{f(x)+f(0)} \\
\Leftrightarrow\;& x\left(\frac{f(x)}{f(x)+f(0)} - \frac{f(x)}{f(x)+f(y)}\right) \leq y\,\frac{f(y)}{f(x)+f(y)} \\
\Leftrightarrow\;& xf(x)\left(\frac{f(x)+f(y)}{f(x)+f(0)} - 1\right) \leq yf(y) \\
\Leftrightarrow\;& xf(x)\,\frac{f(y)-f(0)}{f(x)+f(0)} \leq yf(y). \qquad (4)
\end{aligned}
$$

Note that $f(y) - f(0) > 0$, since f is strictly increasing. Now, for any fixed $y > 0$, we have that $f(x)\frac{f(y)-f(0)}{f(x)+f(0)} \to 1$ as $x \to \infty$, and the expression is therefore bounded from below by some positive value for x sufficiently large. Hence, for any fixed y there exists an $x > 0$ such that (4) does not hold. This shows that in fact for any $y > 0$ there exists $x > 0$ such that $V_{\pi^1,\pi^2(\pi^1)}(s_0) < V_{\bar{\pi}^1,\pi^2(\bar{\pi}^1)}(s_0)$, whereas we have seen before that $V_{\pi^1,\pi^2(\pi^1)}(s_2) > V_{\bar{\pi}^1,\pi^2(\bar{\pi}^1)}(s_2)$. Hence, no dominating commitment strategy exists for the MDP depicted in Figure 4. ■

B.2 Proof of Lemma 3

Lemma 3 *If $\mathcal{A}_2$ plays ε-greedy responses, a dominating commitment strategy for $\mathcal{A}_1$ may not exist.*

We define an ε-greedy response to a policy π^1 as the policy

$$
\pi^2_\varepsilon(s, \pi^1) = \begin{cases} \pi^2_*(s, \pi^1) & w.p.\ 1-\varepsilon \\ \mathcal{U}(A_2) & w.p.\ \varepsilon, \end{cases}
$$

where $\varepsilon \in [0, 1]$, $\pi^2_*(\pi^1)$ is an optimal response to π^1, and $\mathcal{U}(A_2)$ the uniform distribution over A_2.

Proof [Proof of Lemma 3] We prove Lemma 3 by means of the counterexample shown in Figure 5. For convenience, we omit the discount factor here and assume that states s_1, s_3, s_4, and s_5 lead to some terminal state with probability one. There are two (deterministic) policies $\mathcal{A}_1$ can commit to: $\pi^1(s_2) = a_1$ and $\bar{\pi}^1(s_2) = a_2$.

For notational convenience, we write $V_{a_1}(s) \triangleq V_{\pi^1,\pi^2_\varepsilon(\pi^1)}(s)$ and $V_{a_2}(s) \triangleq V_{\bar{\pi}^1,\pi^2_\varepsilon(\bar{\pi}^1)}(s)$. Note that if $\mathcal{A}_1$ commits to π^1, the optimal action for $\mathcal{A}_2$ in state s_0 is to play b_2 followed by b_1 in state s_2. Recall that $\mathcal{A}_2$ is assumed to play

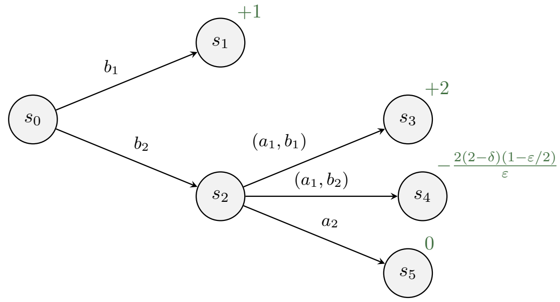


Figure 5: Counterexample for ε-greedy responses. All transitions are deterministic. The actions from agent $\mathcal{A}_2$ alone determine the transitions from state s_0 to states s_1 and s_2. The green numbers denote the rewards obtained in the respective states. States s_0 and s_2 yield zero reward.

ϵ-greedy, i.e. in any state, $\mathcal{A}_2$ plays the optimal response with probability $(1-\epsilon)$ and with probability ϵ selects an action uniformly at random. As a result, we have

$$V_{a_1}(s_2) = 2(1-\varepsilon/2) - (2-\delta)(1-\varepsilon/2) = \delta(1-\varepsilon/2) > 0$$
$$V_{a_1}(s_0) = \delta(1-\varepsilon/2)^2 + \varepsilon/2.$$

On the other hand, if $\mathcal{A}_1$ commits to $\bar{\pi}^1$, it is optimal for $\mathcal{A}_2$ to play b_1 in state s_0, i.e. $V_{a_2}(s_0) = (1-\varepsilon/2)$. We observe that in state s_2, playing a_1 is optimal as $V_{a_1}(s_2) > V_{a_2}(s_2) = 0$. However, we also have $V_{a_1}(s_0) - V_{a_2}(s_0) = \varepsilon + \delta(1-\varepsilon/2)^2 - 1$. As we can choose δ arbitrarily close to 0, we then have $V_{a_1}(s_0) < V_{a_2}(s_0)$ for some $\delta > 0$. Thus, π^1 is strictly optimal in state s_2, whereas $\bar{\pi}^1$ is strictly optimal in state s_0. Therefore, there exists no dominating commitment strategy for the MDP in Figure 5.

■

Appendix C. Approximate Algorithms for Planning with Suboptimal Partners

In this section, in addition to the approximate value iteration algorithm for Boltzmann-rational policies (Algorithm 2), we describe an analogous algorithm for ε-greedy policies. We then evaluate both algorithms in the Maze-Maker and Random MDP environment for different levels of rationality (i.e. optimality) of agent $\mathcal{A}_2$.

C.1 $\mathcal{A}_2$ responds with ε-greedy policies

The problem of planning with an agent that responds with ε-greedy policies is similar to the setting considered by Dimitrakakis et al. (2017) in the sense that $\mathcal{A}_2$ plans with the original transition kernel $\mathcal{P}$ (by computing an optimal response $\pi^2_*(\pi^1)$), whereas $\mathcal{A}_1$ plans (or should plan) with the "correct" transition kernel

$$\mathcal{P}_\varepsilon(\cdot \mid s, a, b) \equiv \varepsilon\mathcal{P}(\cdot \mid s, a, \mathcal{U}(A_2)) + (1-\varepsilon)\mathcal{P}(\cdot \mid s, a, b).$$

In particular, note that $\varepsilon\mathcal{P}(s' \mid s, a, \mathcal{U}(A_2))$ is independent of the choice of b. Algorithm 3 approximately solves the planning problem. While Lemma 3 states that a dominating commitment policy need not exist, Algorithm 3 simply acts as if one exists. Similarly to Algorithm 2, the idea is to maintain two value functions, one representing the value from the perspective of $\mathcal{A}_1$ and the other the value from the perspective of $\mathcal{A}_2$.

Algorithm 3 Approximate Value Iteration for ε-Greedy Responses

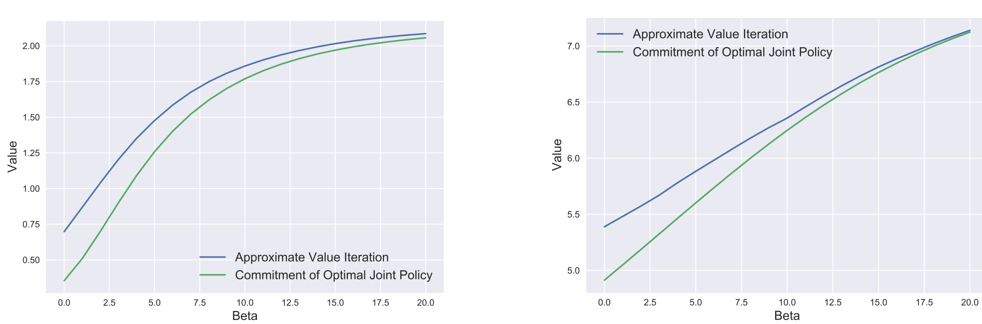
initialise V and $\hat{V}$
repeat until V converges:
for $s \in \mathcal{S}$ **do**
  for $a \in A_1$ **do**
    $\pi^2(s,a) = \arg\max_b \sum_{s'} \mathcal{P}(s'|s,a,b)\hat{V}(s')$
  end for
  $\pi^1(s) = \arg\max_a \sum_{s'} \mathbb{E}_{b\sim\pi^2}[\mathcal{P}_\varepsilon(s'|s,a,b)]V(s')$
  $V(s) = r(s) + \gamma \sum_{s'} \mathbb{E}_{b\sim\pi^2}[\mathcal{P}_\varepsilon(s'|s,\pi^1(s),b)]V(s')$
  $\hat{V}(s) = r(s) + \gamma \sum_{s'} \mathbb{E}_{b\sim\pi^2}[\mathcal{P}(s'|s,\pi^1(s),b)]\hat{V}(s')$
end for

C.2 Evaluation of Algorithm 2 and Algorithm 3

In this section, we empirically evaluate our approximate value iteration algorithms for Boltzmann-rational responses (Algorithm 2) and ε-greedy responses (Algorithm 3). We compare Algorithm 2 and Algorithm 3 in the Maze-Maker and Random MDP environment against committing $\mathcal{A}_1$'s part of the optimal joint policy. Note that by Lemma 2, committing $\mathcal{A}_1$'s part of an optimal joint policy is optimal when $\mathcal{A}_2$ responds optimally.

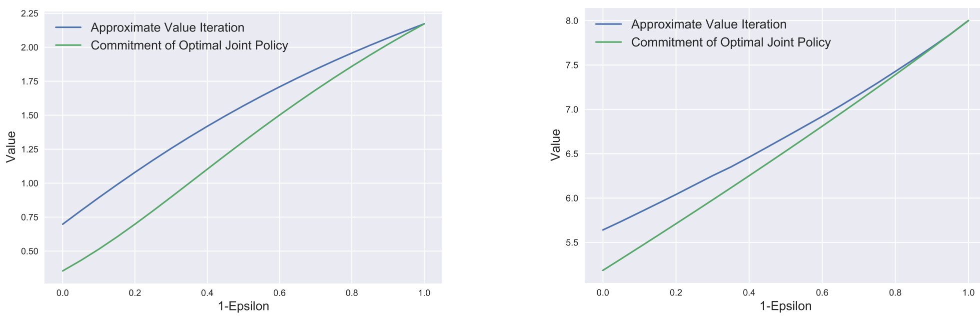
In both environments, we test the performance of our algorithms for different levels of rationality of $\mathcal{A}_2$. For the case of Boltzmann-rational responses (Figure 6), we increase the inverse temperature of agent $\mathcal{A}_2$, which corresponds to the rationality (i.e. optimality) of $\mathcal{A}_2$. We see in Figure 6 that Algorithm 2 consistently outperforms playing $\mathcal{A}_1's$ part of the optimal joint policy. In particular, the more suboptimal $\mathcal{A}_2$ is playing (lower values of β), the larger the advantage of Algorithm 2 is compared to playing $\mathcal{A}_1$'s part of the optimal joint policy. If $\mathcal{A}_2$ responds almost optimally ($\beta = 20$), the performance of both approaches is almost identical as expected.

For the case of ε-greedy responses (Figure 7), we increase the rationality of $\mathcal{A}_2$ by decreasing the probability ε of random actions. Figure 7 shows that Algorithm 3 outperforms playing the optimal joint policy for all values of ε in both environments. In particular, for $\varepsilon = 0$ agent $\mathcal{A}_2$ responds optimally and both approaches play an optimal commitment strategy.



(a) Maze-Maker (b) Random MDPs

Figure 6: Evaluation of Approximate Value Iteration for Boltzmann-Rational Responses (Algorithm 2) in the Maze-Maker and Random MDP environment for increasing values of β. The green line describes the return of playing $\mathcal{A}_1$'s part of an optimal joint policy.



(a) Maze-Maker (b) Random MDPs

Figure 7: Evaluation of Approximate Value Iteration for ε-Greedy Responses (Algorithm 3) in the Maze-Maker and Random MDP environment for decreasing values of ε. The green line describes the return of playing $\mathcal{A}_1$'s part of an optimal joint policy.

Appendix D. Experimental Details

The experiments were carried out on a virtual machine with 32 CPUs, 60GB RAM, and CentOS Linux 8 operating system. The experiments were implemented in Python 3.7 and the libraries matplotlib 3.2.1, numpy 1.20.1, and scipy 1.6.2 (for the linear program) were used. The code is available at https://github.com/InteractiveIRL/src.

Setup. The initial policy π_1^1 is chosen uniformly at random. We model the standard IRL setting by repeatedly generating responses of $\mathcal{A}_2$ with respect to π_1^1, i.e. in the implied environment $\mathcal{P}_{\pi_1^1}$. Using these observations, we then estimate the reward function using standard IRL, compute the optimal policy with respect to the estimated rewards, and evaluate the regret of this policy. In contrast, in the Interactive IRL setting, the learner gets to choose a different policy in subsequent episodes. In this case, we report the online regret of the actually played policies, i.e. the actual regret of the learner.

For the case of suboptimal responses and partial information, we assume that $\mathcal{A}_2$ responds with Boltzmann-rational policies with inverse temperature $\beta = 10$ in both environments. We assume that the inverse temperature, that is, the optimality of the second agent, is *unknown* to the learner and must therefore be inferred. We simulate the partial information setting by generating trajectories according to policies π_t^1 and π_t^2 in episode t, where the length of the episode is random. More precisely, we let an episode end with probability $1 - \gamma = 0.1$ each time step.[9]

D.1 Bayesian Interactive IRL

We employ a Bayesian approach using the Metropolis-Hastings algorithm to sample from the posterior, with a uniform prior on the reward function and an exponential prior on the inverse temperature. Our approach is specified in Algorithm 4. As a proposal distribution for the reward function, we consider a discretisation of the $|\mathcal{S}|$-dimensional unit simplex $\Delta(\mathcal{S})$ with step size δ, similarly to (Ramachandran and Amir, 2007). The Metropolis-Hastings algorithm then generates a Markov chain on the discretised simplex. To sample from the posterior given the last candidate r_{k-1}^t then means to choose a neighbour in the discretised simplex. This type of proposal distribution, which we refer to as Simplex Walk, proved to be a more efficient and robust sampling strategy as other proposal distributions (e.g. Dirichlet distributions). For the inverse temperature, we use a Gamma proposal distribution. Similarly to Algorithm 1, we play greedily with respect to our current estimate of the true reward function. After sampling K times from the posterior, we take the empirical means $\bar{r}_t$ and $\bar{\beta}_t$ and compute an approximately optimal commitment strategy under $\bar{r}_t$ and $\bar{\beta}_t$ by means of Algorithm 2. As a natural burn-in we use the last sampled reward and inverse temperature from episode t as the first candidate in episode $t + 1$.

9. We impose a minimal trajectory length of 2 time steps to prevent vacuous episodes.

Algorithm 4 Bayesian Interactive IRL via Simplex Walk

input: $(\mathcal{S}, A_1, A_2, \mathcal{P}, \gamma)$, priors $\mathbb{P}(r)$, $\mathbb{P}(\beta)$, proposal distributions g_1, g_2, sample size K
initialise: choose π_1^1 uniformly at random, sample $r_0^0 \sim \mathbb{P}(r)$ and $\beta_0^0 \sim \mathbb{P}(\beta)$
for $t = 1, 2, \dots$ **do**
  commit to policy π_t^1
  observe trajectory τ_t
  // `sample from posterior via Metropolis-Hastings`
  for $k = 1, \dots, K$ **do**
    sample $r \sim g_1(\cdot \mid r_{k-1}^t)$
    sample $\beta \sim g_2(\cdot \mid \beta_{k-1}^t)$
    compute $p = \frac{\mathbb{P}((\pi_1^1,\tau_1),\dots,(\pi_t^1,\tau_t)|r,\beta)\mathbb{P}(r)\mathbb{P}(\beta)}{g_1(r|r_{k-1}^t)g_2(\beta|\beta_{k-1}^t)}$
    w.p. $\min\{1, \frac{p}{p_{k-1}}\}$**:** $r_k^t = r$, $\beta_k^t = \beta$, $p_k^t = p$
    else: $r_k^t = r_{k-1}$, $\beta_k^t = \beta_{k-1}$, $p_k^t = p_{k-1}^t$
  end for
  set $r_0^{t+1} = r_K^t$, $\beta_0^{t+1} = \beta_K^t$, $p_0^{t+1} = p_K^t$
  calculate mean reward function $\bar{r}_t$ and beta $\bar{\beta}_t$
  compute π_{t+1}^1 under $\bar{r}_t$ and $\bar{\beta}_t$ via Algorithm 2
end for

D.2 Environments: Maze-Maker

In the Maze-Maker environment, agents $\mathcal{A}_1$ and $\mathcal{A}_2$ jointly control a cart in a 7×7 grid world. In this grid world, the doors leading from one cell to the neighbouring ones are locked. However, $\mathcal{A}_1$ can unlock exactly two doors at any time step before they fall shut again. $\mathcal{A}_2$ can attempt to move the cart through a door to a neighbouring cell. However, when the door is locked, the cart stays where it was. We assume that any attempted move of the cart succeeds with probability 0.8 and that with probability 0.2 the cart moves to a random neighbouring cell. Agents $\mathcal{A}_1$ and $\mathcal{A}_2$ are tasked with collecting three rewards of different value (+1, +2, +3), which are scattered in the grid world and disappear once collected. While $\mathcal{A}_2$ knows where the rewards are placed, $\mathcal{A}_1$ does not know their location. An illustration of the environment is given by Figure 2. We model this environment as a two-agent MDP with 392 states (49×8) and discount factor $\gamma = 0.9$, where $\mathcal{A}_1$ has six actions (unlocking two out of four doors) and $\mathcal{A}_2$ four actions (attempting to move the cart North, East, South, West). As we consider a Stackelberg game, $\mathcal{A}_2$ knows beforehand which doors $\mathcal{A}_1$ will unlock. Therefore, $\mathcal{A}_1$ essentially selects a maze layout, which is communicated to $\mathcal{A}_2$ and through which $\mathcal{A}_2$ can move the cart.

D.3 Details on Figure 1

In Figure 1b, we assumed that $\mathcal{A}_2$ plays a Boltzmann-rational policy with inverse temperature $\beta = 10$. For simplicity and proper comparison, we assume that we can observe the fully specified Boltzmann policy played by $\mathcal{A}_2$ in each of the mazes. We use an adaption if Bayesian IRL (Ramachandran and Amir, 2007) and display the mean reward function in Figure 1b, where the colour scale, i.e. colour transparency, is obtained from the mean reward function in a given cell. More precisely, we use the Metropolis-Hastings algorithm with uniform prior and a Dirichlet proposal to sample from the posterior distribution $\mathbb{P}(r \mid (\pi^1, \pi^2))$, where π^1 describes the maze layout.

Appendix E. Influence

Prior work on two-agent cooperation has considered measurements of how much one agent can influence the transition probabilities. Dimitrakakis et al. (2017) define the influence of agent $\mathcal{A}_1$ (analogously for $\mathcal{A}_2$) on the transition probabilities as

$$\mathcal{I}(\mathcal{A}_1) = \max_s \max_{a_1, a_2, b} \|\mathcal{P}(\cdot \mid s, a_1, b) - \mathcal{P}(\cdot \mid s, a_2, b)\|_1,$$

which has also been adopted by Radanovic et al. (2019) and Ghosh et al. (2019). They use this definition of influence to bound the performance gap when the beliefs or the behaviour of the two agents are misaligned. In our setting, however, the influence of an agent also relates to the IRL problem and our capacity to solve it. In particular, if $\mathcal{I}(\mathcal{A}_1) = 0$, agent $\mathcal{A}_1$ does not influence the transition probabilities and it is therefore irrelevant what actions $\mathcal{A}_1$ takes. In terms of the IRL problem, we are then in the typical single-agent setting as $\mathcal{A}_2$ can ignore the presence of agent $\mathcal{A}_1$. On the other hand, if $\mathcal{I}(\mathcal{A}_2) = 0$, then $\mathcal{A}_2$ does not influence transitions at all and the IRL problem becomes intractable as $\mathcal{A}_2$'s actions yield no information about the underlying reward function.