Interactive Inverse Reinforcement Learning for Cooperative Games

Thomas Kleine Buening, Anne-Marie George, Christos Dimitrakakis

Overview

Overview

Motivation:
How to **collaborate** with a potentially suboptimal (human) partner **without** knowledge of or access to the **joint reward function**?

Overview

Motivation:
How to **collaborate** with a potentially suboptimal (human) partner **without** knowledge of or access to the **joint reward function**?

Research Question:
Can we achieve effective cooperation by learning about the reward function from **interactions** with the human partner?

Overview

Motivation:
How to **collaborate** with a potentially suboptimal (human) partner **without** knowledge of or access to the **joint reward function**?

Research Question:
Can we achieve effective cooperation by learning about the reward function from **interactions** with the human partner?

Insight:
Not only can we achieve effective cooperation, but we can infer the reward function **more precisely and with fewer samples** when **interacting with a human expert** compared to traditional IRL.

The Setting: Interactive IRL

The Setting: Interactive IRL

Episodic Cooperative 2-Agent MDP $(\mathcal{S}, A_1, A_2, \mathcal{P}, R, \gamma)$

The Setting: Interactive IRL

Episodic Cooperative 2-Agent MDP $(\mathcal{S}, A_1, A_2, \mathcal{P}, R, \gamma)$

- we control $\mathcal{A}_1$ (**the learner**), but not $\mathcal{A}_2$ (**the human**)

The Setting: Interactive IRL

Episodic Cooperative 2-Agent MDP $(\mathcal{S}, A_1, A_2, \mathcal{P}, R, \gamma)$

- we control $\mathcal{A}_1$ (**the learner**), but not $\mathcal{A}_2$ (**the human**)
- we don't know or observe the **joint** reward function R

The Setting: Interactive IRL

Episodic Cooperative 2-Agent MDP $(\mathcal{S}, A_1, A_2, \mathcal{P}, R, \gamma)$

- we control $\mathcal{A}_1$ (**the learner**), but not $\mathcal{A}_2$ (**the human**)
- we don't know or observe the **joint** reward function R
- we want to learn about R (in order to minimise regret)

The Setting: Interactive IRL

Episodic Cooperative 2-Agent MDP $(\mathcal{S}, A_1, A_2, \mathcal{P}, R, \gamma)$

- we control $\mathcal{A}_1$ (**the learner**), but not $\mathcal{A}_2$ (**the human**)
- we don't know or observe the **joint** reward function R
- we want to learn about R (in order to minimise regret)

Stackelberg Game:

The Setting: Interactive IRL

Episodic Cooperative 2-Agent MDP $(\mathcal{S}, A_1, A_2, \mathcal{P}, R, \gamma)$

- we control $\mathcal{A}_1$ (**the learner**), but not $\mathcal{A}_2$ (**the human**)
- we don't know or observe the **joint** reward function R
- we want to learn about R (in order to minimise regret)

Stackelberg Game:
Each episode t:

The Setting: Interactive IRL

Episodic Cooperative 2-Agent MDP $(\mathcal{S}, A_1, A_2, \mathcal{P}, R, \gamma)$

- we control $\mathcal{A}_1$ (**the learner**), but not $\mathcal{A}_2$ (**the human**)
- we don't know or observe the **joint** reward function R
- we want to learn about R (in order to minimise regret)

Stackelberg Game:
Each episode t:

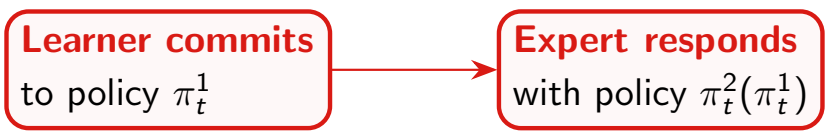
Learner commits
to policy π_t^1

The Setting: Interactive IRL

Episodic Cooperative 2-Agent MDP $(\mathcal{S}, A_1, A_2, \mathcal{P}, R, \gamma)$

- we control $\mathcal{A}_1$ (**the learner**), but not $\mathcal{A}_2$ (**the human**)
- we don't know or observe the **joint** reward function R
- we want to learn about R (in order to minimise regret)

Stackelberg Game:
Each episode t:



The Setting: Interactive IRL

Episodic Cooperative 2-Agent MDP $(\mathcal{S}, A_1, A_2, \mathcal{P}, R, \gamma)$

- we control $\mathcal{A}_1$ (**the learner**), but not $\mathcal{A}_2$ (**the human**)
- we don't know or observe the **joint** reward function R
- we want to learn about R (in order to minimise regret)

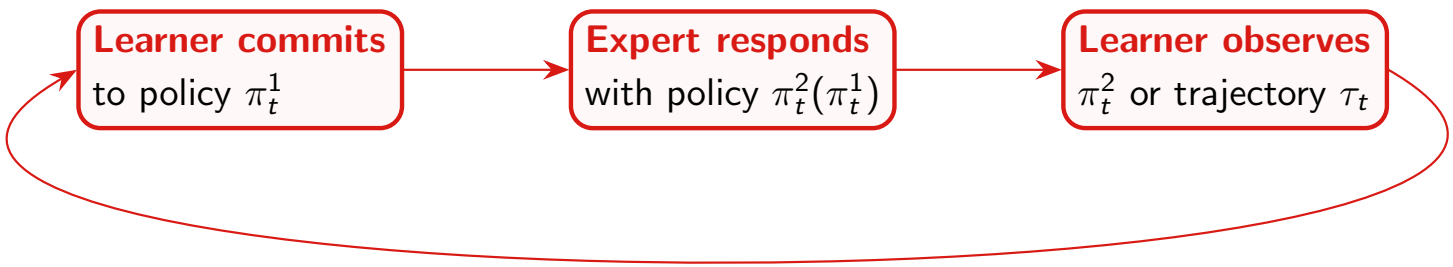
Stackelberg Game:
Each episode t:

The Setting: Interactive IRL

Episodic Cooperative 2-Agent MDP $(\mathcal{S}, A_1, A_2, \mathcal{P}, R, \gamma)$

- we control $\mathcal{A}_1$ (**the learner**), but not $\mathcal{A}_2$ (**the human**)
- we don't know or observe the **joint** reward function R
- we want to learn about R (in order to minimise regret)

Stackelberg Game:
Each episode t:



How to learn about rewards from interactions?
Main Idea: The Learner as an MDP Designer

How to learn about rewards from interactions?
Main Idea: The Learner as an MDP Designer

Learner commits to **policy** π_t^1 to which the expert responds

$\Updownarrow$

Learner chooses **environment** $\mathcal{P}_{\pi_t^1}$ in which the expert acts

How to learn about rewards from interactions?
Main Idea: The Learner as an MDP Designer

Learner commits to **policy** π_t^1 to which the expert responds

$\Updownarrow$

Learner chooses **environment** $\mathcal{P}_{\pi_t^1}$ in which the expert acts

This gives us a way to interpret the expert's actions:

How to learn about rewards from interactions?
Main Idea: The Learner as an MDP Designer

Learner commits to **policy** π_t^1 to which the expert responds

$\Updownarrow$

Learner chooses **environment** $\mathcal{P}_{\pi_t^1}$ in which the expert acts

This gives us a way to interpret the expert's actions:

expert's response $\pi_t^2(\pi_t^1) \widehat{=}$ expert policy π_t^2 in $(\mathcal{S}, A_2, \mathcal{P}_{\pi_t^1}, R, \gamma)$

How to learn about rewards from interactions?
Main Idea: The Learner as an MDP Designer

Learner commits to **policy** π_t^1 to which the expert responds

$\Updownarrow$

Learner chooses **environment** $\mathcal{P}_{\pi_t^1}$ in which the expert acts

This gives us a way to interpret the expert's actions:

expert's response $\pi_t^2(\pi_t^1) \widehat{=}$ expert policy π_t^2 in $(\mathcal{S}, A_2, \mathcal{P}_{\pi_t^1}, R, \gamma)$

We essentially get to observe the expert in different environments, environments that we (the learner) design.

Overview of Results

Overview of Results

Theoretical:

Overview of Results

Theoretical:

- online algorithm that is **no-regret**

Overview of Results

Theoretical:

- online algorithm that is **no-regret**
- existence of **optimal reward learning environments**

Overview of Results

Theoretical:

- online algorithm that is **no-regret**
- existence of **optimal reward learning environments**
- Stackelberg games with **suboptimal followers** are difficult

Overview of Results

Theoretical:

- online algorithm that is **no-regret**
- existence of **optimal reward learning environments**
- Stackelberg games with **suboptimal followers** are difficult

Experimental:

Overview of Results

Theoretical:

- online algorithm that is **no-regret**
- existence of **optimal reward learning environments**
- Stackelberg games with **suboptimal followers** are difficult

Experimental:

- IRL with interactions is more **sample-efficient** and **precise**

Overview of Results

Theoretical:

- online algorithm that is **no-regret**
- existence of **optimal reward learning environments**
- Stackelberg games with **suboptimal followers** are difficult

Experimental:

- IRL with interactions is more **sample-efficient** and **precise**