Imitating Careful Experts to Avoid Catastrophic Events

Jack Hanslope

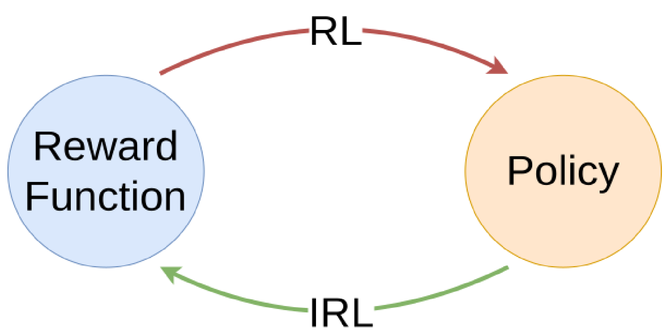
Laurence Aitchison

Abstract

Reinforcement learning (RL) is increasingly being used to control robotic systems that interact closely with humans. This interaction raises the problem of safe RL: how to ensure that an RL-controlled robotic system never, for instance, injures a human. This problem is especially challenging in rich, realistic settings where it is not even possible to clearly write down a reward function which incorporates these outcomes. In these circumstances, perhaps the only viable approach is based on inverse reinforcement learning (IRL), which infers rewards from human demonstrations. However, IRL is massively underdetermined as many different rewards can lead to the same optimal policies; we show that this makes it difficult to distinguish catastrophic outcomes (such as injuring a human) from merely undesirable outcomes. Our key insight is that humans do display different behaviour when catastrophic outcomes are possible: they become much more careful. We incorporate carefulness signals into IRL, and find that they do indeed allow IRL to disambiguate undesirable from catastrophic outcomes, which is critical to ensuring safety in future real-world human-robot interactions.

Inverse Reinforcement Learning

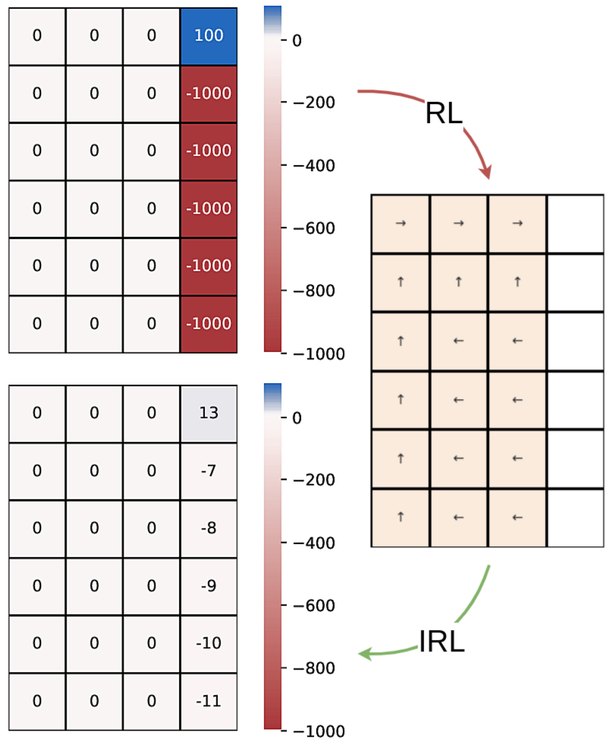
- Standard (or forwards) reinforcement learning takes a reward function as its input and outputs an optimal policy
- Inverse reinforcement learning takes an optimal policy as its input and outputs a reward function



The Problem with IRL

Often, many reward functions will be valid for a given optimal policy. IRL can fail to distinguish between events that are catastrophic and events that are undesirable.

Consider the gridworld example below with stochastic movement. The red states have a reward of –1000 and the blue state a reward of +100. We perform RL to get the optimal policy and then perform IRL on this optimal policy to obtain the reward function at the bottom. Whilst this reward captures the correct positive and negative states, it fails to appreciate the magnitude of the reward.



Carefulness

Humans and other animals display carefulness, especially when they perceive themselves to be at risk of catastrophe.

Other animals can learn to avoid catastrophic events without having to either observe or experience said event themselves.

IRL with Carefulness

We implement carefulness as follows

- Movement is stochastic
- There is a penalty for taking an action
- More careful means less stochastic movement and larger action penalty

We perform RL to obtain the optimal policy shown below before recovering a reward function. This function better captures the magnitude of the reward than when we had no carefulness.