Optimising Misspecified Reward Functions

This project relates to the Theoretical Foundations of Reward Learning research agenda. You can find more information about this research agenda here (link), including a broader description of the aims and background assumptions behind this research. This project would be mainly theoretical/mathematical.

Overview:

The aim of this project is to develop methods for optimising a given reward function R in a reinforcement learning environment, which gives us some formal guarantees even under the assumption that R is misspecified (i.e., fails to capture everything we care about).

Background:

In practice, we should not expect to be able to create reward functions that perfectly capture our preferences. This then raises the question of how we should optimise reward functions, in light of this fact. Can we create some special "conservative" optimisation methods that yield provable guarantees? There is a lot of existing work on this problem, including e.g., quantilizers, or the work on side-effect minimisation. However, there are several promising directions for improving this work. In particular, much of the existing work on conservative optimisation makes no particular assumptions about how the reward function is misspecified. For example, in Karwowski et al. (2023), we derive a (tight) criterion that describes how a policy may be optimised according to some proxy reward while still ensuring that the true reward does not decrease, based on the STARC-distance (Skalse et al., 2024) between the proxy reward and the true reward. Other specific assumptions about *how* the reward function is misspecified may similarly produce other conservative optimisation methods.

Methodology:

There are multiple ways to approach this problem. One interesting case of conservative optimisation is the case where we have multiple reward functions produced from different sources. If we have several sources of information about a latent variable (such as a reward function), then we can normally simply combine the evidence using Bayes' theorem. However, if the observation models are misspecified, then the posterior distributions will conflict, and the straightforward way to combine them is likely to lead to nonsensical results (e.g., Krasheninnikov et al., 2021). This applies to the problem of inferring human preferences – there are several sources of information we can rely on, but they can often lead to conflicting conclusions. For example, if you ask people what kind of chocolate they prefer, they will typically say that they like dark chocolate, but if you let people choose which type of chocolate to eat, they will typically pick light chocolate. If we take these two pieces of information at face value, they lead to incompatible conclusions. The reason for this is misspecification – humans

sometimes misreport their preferences, and sometimes do things that are against their preferences. This means that we cannot combine this information in the straightforward way, and that special methods are required. We can set up this problem as follows; suppose we have n reward learning methods, and that they have produced n distributions over reward functions $\Delta_1 \dots \Delta_n$ We know that any (or all) of these may be subject to some degree of misspecification, and we wish to derive a policy π that is reasonably good in the light of what we know. There are several considerations that we might want π to satisfy:

1. We might want π to get reasonably high reward according to each of the reward function distributions $\Delta_1 \dots \Delta_n$.

2. We might want π to preserve option value, in case more information can be obtained later.

3. We might want to supply π with the ability to make active queries for resolving ambiguity. In this case, we want π to make intelligent use of this ability, by making relevant queries when they are needed.

There are several different ways to approach this problem. A simple option might be to let π maximise $\min_{i \in \{1 \dots n\}} J_i(\pi)$. Alternatively, one could explore solution concepts from social choice theory (where each Δ_i is considered to be a voter). A good solution to this problem would provide methods for aggregating misspecified (but informative) preferences in a useful but conservative way.

References

Karwowski, Jacek, Oliver Hayman, Xingjian Bai, Klaus Kiendlhofer, Charlie Griffin, and Joar Skalse (2023). *Goodhart's Law in Reinforcement Learning*. arXiv: 2310.09144 [cs.LG]. URL: https://arxiv.org/abs/2310.09144.

Krasheninnikov, Dmitrii, Rohin Shah, and Herke van Hoof (2021). *Combining Reward Information from Multiple Sources*. arXiv: 2103.12142 [cs.LG]. URL: https://arxiv.org/abs/2103.12142.

Skalse, Joar, Lucy Farnik, Sumeet Ramesh Motwani, Erik Jenner, Adam Gleave, and Alessandro Abate (2024). *STARC: A General Framework For Quantifying Differences Between Reward Functions*. arXiv: 2309.15257 [cs.LG]. URL: https://arxiv.org/abs/2309.15257.