Project Proposal

Alina Trinh, Samarth Goel

1 Which tasks or problems will you study? Where will you get your data or simulator (or real-world system)?

We want to study the problem of doing inverse RL using non-expert demonstrations when we don't have expert labeling of preferences. We can get the data ourselves by providing demonstrations on existing environments, such as gridworlds, OpenAI gym environments, or other open-source libraries.

2 What is the main research hypothesis your project will investigate? All projects should at least attempt to evaluate novel ideas that pertain to deep RL or its applications.

We hypothesize that an AI agent will be able to identify suboptimal or non-expert demonstrations by studying the future outcomes of certain trajectories. In this project, we assume that we face a bias where the expert is aware of the ideal behavior but is otherwise unable to act consistently towards that ideal. There are two possibilities as to why this is:

- The expert is incapable of acting consistently because they have high uncertainty about what to do in a particular state. For example, in navigating traffic safely: at an unprotected turn, should we just go for it? Or should we wait for that car that's kind of far away but not really to pass?
- The expert is incapable of acting consistently because they are indifferent about what to do in a particular state. Same traffic example: do we take the freeway exit from the rightmost lane or the right lane? Doesn't really matter.

Without expert labeling of preferences, we just have to infer when the demos are suboptimal so we can be more careful when crafting our reward function. One potential approach is to gather trajectory samples from the demonstrator without any knowledge if they are optimal or not, then compare the similarities between the states and actions. States that are very similar but have actions that are very different are those that denote expert inconsistency.

The assumption here is that in states with high uncertainty the resulting outcomes may be very different (e.g. accidentally cause an accident vs. waste time waiting), while in states with high indifference, the resulting outcomes would be the same either way, otherwise there wouldn't be indifference to start with. So for these inconsistent states, we can figure out whether they are due to uncertainty or indifference by looking at the state similarity between their trajectory outcomes, and for both outcomes we need to figure out what the ACTUAL optimal action is.

- For states that we've determined to have uncertainty, we can leave them and their actions out of our data, because once we run a typical IRL reward inference algorithm, we can figure out which action(s) in the uncertain state(s) give(s) us the highest reward.
- For states that we've determined to have indifference, we can simply use the action that occurs more often in inferring the reward.

3 How does the topic of your project relate to deep RL?

Deep RL can be used in learning the reward function as well as retrieving the optimal policy after we've estimated the reward function from the demonstrations.