CS 277 (W22): Control and Reinforcement Learning

Quiz 6: Inverse RL and Bounded RL

Due date: Monday, March 7, 2022 (Pacific Time)

Roy Fox
https://royf.org/crs/W22/CS277

Instructions: please solve the quiz in the marked spaces and submit this PDF to Gradescope.

Question 1 The Inverse RL (IRL) algorithms we saw also find a good policy. Comparing IRL to Imitation Learning (IL) (check all that hold):

- □ Learning both a reward function and a policy can be an easier problem than only learning a policy.
- □ IL methods that also learn a reward function are typically more robust to suboptimal demonstrations than those that don't.
- □ IL methods that also learn a reward function are typically more robust to conflicting or multi-modal demonstrations than those that don't.
- □ Pre-training with IRL in one environment can provide a good starting point for IL in another environment with similar but different dynamics, such as in sim2real.
- □ Pre-training with IRL in one task can provide a good starting point for IL in a completely different task with the same environment dynamics.

Question 2 Generative Adversarial Imitation Learning (GAIL) was formulated in terms of entropy-regularized RL with discriminator-based rewards; see lecture 14, slide 16, last line of the algorithm. If another RL algorithm is used in GAIL, is the justification to use discriminator-based rewards, as presented in slide 15, still correct? **Yes / No**.

Briefly justify:

Question 3 In Soft Q-Learning (SQL) (check all that hold):

- □ As $\beta \to 0$, the algorithm learns a value function Q_{π_0} that evaluates π_0.
- □ In large action spaces, we can obtain an unbiased estimate of the target value $r + \frac{\gamma}{\beta} \log \mathbb{E}_{(a'|s') \sim \pi_0}[\exp \beta Q(s', a')]$ by replacing the expectation with a sample $(a'|s') \sim \pi_0$.
- □ The soft-optimal policy can also be used for exploration.
- □ When π_0 is uniform and β is finite, $Q(s, a)$ penalizes actions that lead to future states in which some actions are much better than others.