CS 277 (W24): Control and Reinforcement Learning

Quiz 8: Inverse RL and Bounded RL

Due date: Wednesday, March 13, 2024 (Pacific Time)

Roy Fox
https://royf.org/crs/CS277/W24

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, in which sense they are like Imitation Learning (IL). Comparing IRL to 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 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 $r + \frac{\gamma}{\beta} \log \mathbb{E}_{(a'|s')\sim\pi_0}[\exp \beta Q_\beta(s', a')]$ 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, $Q_\beta(s, a)$ penalizes actions that lead to future states in which some actions are much better than others.

Question 3 In Soft Actor–Critic (SAC) (check all that hold):

- ☐ Both SAC and TRPO can be effective for Offline RL because they can both constrain the KL-divergence from the data distribution.
- ☐ As $\beta \to \infty$ and $\pi_\theta \to$ deterministic, the SAC loss approaches the DDPG loss.
- ☐ SAC can be applied in large and continuous action spaces because the actor maintains an explicit policy, unlike SQL which is value-based.
- ☐ The SAC actor imitates the soft-greedy policy suggested by the critic, $\frac{\pi_0(a|s)\exp\beta Q_\beta(s,a)}{\exp\beta V(s)}$. Since the normalizer needs to integrate the numerator, which is infeasible in large or continuous action spaces, SAC needs to also maintain a V network.