CS 277 (W22): Control and Reinforcement Learning

Quiz 2: Imitation and TD Learning

Due date: Wednesday, January 19, 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 Check all that hold in Imitation Learning:

- ☐ With enough data, BC can learn an optimal policy, even if the optimal demonstrator is perturbed by ϵ-greedy behavior (i.e. has probability ϵ to take a uniformly random action).
- ☐ It may be impossible, with any amount of data, to successfully imitate a demonstrator with a different state observability (different sensors) than the learner.
- ☐ Both DAgger and DART can overcome inconsistent demonstrations more easily than BC.
- ☐ DART tends to outperform DAgger if teacher actions tend to get worse the less likely a state is to appear in (noiseless) teacher demonstrations.

Question 2 Value Iteration in finite state and action spaces (check all that hold):

- ☐ Converges regardless of how it is initialized.
- ☐ Can be computed in $O(|S|^2|A|)$ time per iteration.
- ☐ Finds the optimal value function in a finite number of iterations.

Question 3 Reinforcement learning with MC policy evaluation (check all that hold):

- ☐ Always converges in finite state and action spaces, if it samples enough data in each iteration.
- ☐ Can benefit from a replay buffer, due to the data diversity it provides.
- ☐ Can benefit from using an ϵ-greedy interaction policy, compared with greedy.
- ☐ If using ϵ-greedy, can benefit from gradually taking ϵ to 0, compared with constant ϵ.

Question 4 We discussed Fitted Value-Iteration (FVI), Fitted Q-Iteration (FQI), and Sampling-based Fitted Q-Iteration, but not Sampling-based Fitted Value-Iteration (using V). Is such an algorithm possible? **Yes / No**.

Briefly justify:

Question 5 In Deep Q-Learning (check all that hold):

- □ Representing the Q function with a network that outputs a size $|A|$ vector enables taking its maximum.
- □ Using a replay buffer stabilizes the training process.
- □ Gradually taking the ϵ (of ϵ-greedy exploration) to 0 throughout learning lessens the train–test distribution mismatch.
- □ Using a target network is useful in diversifying the target values to effectively consider more experience.