CS 277 (W24): Control and Reinforcement Learning

Quiz 3: TD Learning

Due date: Monday, January 29, 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 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.

□ Typically improves the policy faster than Policy Iteration when the state space is large.

Question 2 Reinforcement learning with MC policy evaluation and greedy policy improvement (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 a buffer 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 3 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 4 In Deep Q-Learning (check all that hold):

□ Representing the Q function with a network that outputs a size $|\mathcal{A}|$ 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.