CS 277 (W22): Control and Reinforcement Learning

Quiz 4: Exploration and Optimal Control

Due date: Friday, February 11, 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 In Multi-Armed Bandits, the regret grows sub-linearly (check all that hold):

- □ If and only if the probability of taking an optimal action converges to 1.
- □ If and only if every action is almost surely (with probability 1) taken infinitely many times.
- □ If in each time step we take the action that is most likely to be optimal.
- □ With ϵ-greedy exploration, if and only if ϵ converges to 0.

Question 2 In a noiseless linear control system (A, B) (check all that hold):

- □ If the system is reachable (i.e., the controllability matrix has full column rank) then any state x_t can be reached from any initial state x_0, for any $t > 0$.
- □ If a state x' can be reached from x in exactly t steps, then x can also be reached from x' in exactly t steps.
- □ If the system is reachable then it is stabilizable.
- □ If the system is stabilizable then it is reachable.

Question 3 The LQR solution we saw (lecture 8, slide 12) follows the same principle as Value Iteration (lecture 3, slide 32). If we instead tried to apply Policy Iteration (lecture 3, slide 31) to noiseless linear control systems, would the policy $u_t = \pi(x_t)$ still be linear in the state x_t? **Yes / No**.

Briefly justify:

Question 4 The Linear–Quadratic–Gaussian (LQG) problem is *separable* into estimation (LQE) and control (LQR), in the sense that optimality is guaranteed if the two parts are solved separately and then combined. This separability property is lost if (check all that hold):

- □ The system is not time-invariant.
- □ The system has lower-order terms (e.g. the process noise is Gaussian with non-0 mean, or the cost is quadratic with linear terms).
- □ The system is not stabilizable.
- □ The initial state distribution is not Gaussian.
- □ The cost $c(x_t, u_t)$ is not quadratic in (x_t, u_t).
- □ The observation has Gaussian conditional distribution $p(y_t|x_t)$ with nonlinear mean.
- □ The horizon is infinite.