CS 277 (W24): Control and Reinforcement Learning

Quiz 4: Policy-Gradient Methods

Due date: Monday, February 5, 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 variance of the gradient estimator in REINFORCE (check all that hold):

□ Poses less of a problem in environments where all rewards are very small.

□ Can be reduced by sampling multiple trajectories and averaging the resulting gradients.

□ Can be reduced by sampling multiple trajectories and concatenating them into a longer one.

□ Can be reduced by segmenting each trajectory into shorter ones and considering them as separate trajectories.

Question 2 Using a critic instead of empirical returns in a policy-gradient method (check all that hold):

□ Reduces the variance of the gradient estimator.

□ Can add significant bias to a method that would otherwise only have a slight bias.

□ Can make the method off-policy by using a Q_ϕ critic trained with TD-learning.

□ Requires separately learning two sets of perceptual features, for the actor and the critic.

Question 3 In continuous action spaces, some methods use deterministic policies and perform deterministic policy gradient. Generally, however, policy-gradient methods use stochastic policies. Can we use deterministic policies in policy-gradient methods in discrete action spaces? **Yes / No**.

Briefly justify:

Question 4 The trust-region methods TRPO and PPO (check all that hold):

- □ Can use GAE(λ) for their advantage estimation.
- □ Avoid the policy-gradient term $\nabla_\theta \log \pi_\theta(a|s)$ which in other PG methods is a major source of variance.
- □ Use the importance-sampling weight $\frac{\pi_\theta(a|s)}{\pi_{\bar{\theta}}(a|s)}$, which reduces the gradient estimation variance compared to the mathematically correct weight.
- □ Have an unbiased objective, assuming an accurate critic, in the limit of a vanishing learning rate.