CS 277 (W22): Control and Reinforcement Learning

Quiz 3: Policy-Gradient Methods

Due date: Friday, January 28, 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 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 Generalized Advantage Estimation (check all that hold):

- □ Has lower variance the larger λ is.
- □ Has lower bias the larger λ is.
- □ Can be run faster (in wall-time, not total compute time) by computing multiple gradients in parallel and using them for gradient steps on a centralized parameter server.

Question 4 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 5 In continuous action spaces (check all that hold):

□ The policy-based methods we've seen in lecture 5 can be readily used by representing the policy as a function from a state to the parameters of a continuous action distribution.

□ The actor–critic policy-gradient methods we've seen in lecture 5 require further tricks in order to find maximal critic values.

□ DDPG is prone to local optima when the critic is multi-modal.

Question 6 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 is a source of variance in PG methods.

□ Use the importance-sampling term $\frac{\pi_\theta(a|s)}{\pi_{\bar{\theta}}(a|s)}$, which reduces the estimation variance.

□ Have an unbiased objective, assuming an accurate critic, in the limit of a vanishing learning rate.