Name: Alfred Cueva
Student No.: 2020-12788

Exercise 7

Data Science & Reinforcement Learning

Spring 2023

1. When is a policy gradient a better choice than a value-based method? When is a value-based method more preferred than a policy gradient? Please justify your answer.

- Policy gradient better when we have stochastic policies (optimal) and/or continous action and smoother changes.
- Value-based better for deterministic policies and/or action-space discrete. Computational efficient, stable

2. Explain what benefits Trust Region Policy Optimization (TRPO) has over simple linesearch-based policy gradient methods.

because learning rate in line searches should be estimated by tunning.

because we don't overshoot we estimating the function, in TRPO we use the surrogate function and the TR about the policy (always improves in TR and smooth changes in f). Also more convenient when the variables are too large, also don't need to evaluate the constraint for all actions but rather focus on the trust region. Policy dist. restricted to change rapidly too. Always stable and robust too.

3. Why is REINFORCE with Baseline not an actor-critic method?

Because 因为 its State-value function is only used as baseline, not acritic. State-value function isn't used for updating values of a state with estimates of subsequent states. Baseline doesn't play a role in policy update.

4. The policy gradient theorem for the episodic case establishes that

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s,\theta).$$

Prove the above statement.

$\nabla v_\pi(s) = \nabla\left[\sum_a \pi(a|s)\, q_\pi(s,a)\right]$ for all $s \in S$

$= \sum_a \left[\nabla \pi(a|s)\, q_\pi(s,a) + \pi(a|s) \nabla q_\pi(s,a)\right]$

$= \sum_a \left[\nabla \pi(a|s)\, q_\pi(s,a) + \pi(a|s) \nabla \sum_{s',r} p(s',r|s,a)(r + v_\pi(s'))\right]$

$= \sum_a \left[\nabla \pi(a|s)\, q_\pi(s,a) + \pi(a|s) \sum_{s'} p(s'|s,a) \nabla v_\pi(s')\right]$

$= \sum_a \left[\nabla \pi(a|s)\, q_\pi(s,a) + \pi(a|s) \sum_{s'} p(s'|s,a) \sum_{a'}\left[\nabla \pi(a'|s')\, q_\pi(s',a') + \pi(a'|s') \sum_{s''} p(s''|s',a') \nabla v_\pi(s'')\right]\right]$

$= \sum_{x \in S} \sum_{k=0}^{\infty} \Pr(s \to x, k, \pi) \sum_a \nabla \pi(a|x)\, q_\pi(x,a)$

then $\nabla J(\theta) = \nabla v_\pi(s_0) = \sum_s \left(\sum_{k=0}^{\infty} \Pr(s_0 \to s, k, \pi)\right) \sum_a \nabla \pi(a|s)\, q_\pi(s,a) = \sum_s \eta(s) \sum_a \nabla \pi(a|s)\, q_\pi(s,a)$

$= \sum_{s'} \eta(s') \sum_s \frac{\eta(s)}{\sum_{s'} \eta(s')} \sum_a \nabla \pi(a|s)\, q_\pi(s,a) = \sum_{s'} \eta(s') \sum_s \mu(s) \sum_a \nabla \pi(a|s)\, q_\pi(s,a)$

$\propto \sum_s \mu(s) \sum_a \nabla \pi(a|s)\, q_\pi(s,a)$ ∎