CS-GY 9223G / Fall 2020

Homework 6

Please upload your assignments on or before December 8, 2020.

- You are encouraged to discuss ideas with each other; but
- you **must acknowledge** your collaborator, and
- you **must compose your own** writeup and/or code independently.
- We **require** answers to theory questions to be written in LaTeX, and answers to coding questions in Python (Jupyter notebooks)
- Upload your answers in the form of a single PDF on Gradescope.

---

1. **(3 points)** *Policy gradients*. In class we derived a general form of policy gradients. Let us specialize it here under some simple parameterizations. Suppose the step size is η. We consider the so-called *bandit* setting where the trajectory does not matter; different actions a_i give rise to different rewards R_i.

   a. Define the mapping π such that $\pi(a_i) = \text{softmax}(\theta_i)$ for $i = 1, \ldots, k$, where k is the total number of actions and θ_i is a scalar parameter encoding the value of each action. Show that if action a_i is sampled, then the change in the parameters in REINFORCE is given by:

   $$\Delta\theta_i = \eta R_i(1 - \pi(a_i)).$$

   b. Intuitively explain the dynamics of the above gradient updates.

2. **(3 points)** *Designing rewards in Q-learning*. Suppose we are trying to solve a maze with a goal and a (stationary) monster in some location, and the goal is to reach the goal in the minimum number of moves. We are tasked with designing a suitable reward function for Q-learning. There are two options:

   a. We declare a reward of +2 for reaching the goal, -1 for running into a monster, and 0 for every other move.

   b. We declare a reward of +1.5 for reaching the goal, -1.5 for running into a monster, and -0.5 for every other move.

   Which of these reward functions might lead to better policies?

   (*Hint: For a general case, how does the expected discounted return change if a constant offset is added to all rewards?*)

3. **(4 points)** Open the (incomplete) Jupyter notebook provided as an attachment to this homework in Google Colab (or other environment of your choice) and complete the missing items. Save your finished notebook in PDF format and upload along with your answers to the above theory questions in a single PDF.