CS182/282A
Spring 2020
Take Home Quiz 2

Name: _____

SID: _____

This exam contains 10 pages (including this cover page) and 8 questions. The total number of points available is 80. This exam is **Open Book** and **Open Internet**. The exam is designed to take about 80 minutes to complete.

---

1 Short Answer

1. (4 points) Policy gradient methods estimate the gradient of the expected reward with respect to the policy parameters. What is the main source of variance in this estimate? Explain one approach to reducing this variance.

2. (4 points) Q-learning typically includes a discount factor $\gamma < 1$. What is the effect of γ on the reward that is optimized, and on the convergence of Q-iteration?

3. (4 points) Consider the Nature DQN architecture. The output of the final convolution layer is flattened before applying the dense layers for the Q-function estimates. If one were to mean pool (across spatial dimensions) the final convolution layer's output instead of flattening, would you expect the agent to work (a) better, (b) worse or (c) more or less the same? Explain your choice.

2 Long Answer

4. (16 points) (Fairness) Suppose you have a prediction task with a protected attribute A, admissible attribute X, output Y and prediction $\hat{Y}$. Assume all the variables are binary.

   (a) (6 points) Consider the variables A and $\hat{Y}$. Their joint distribution can be specified by four probabilities $\Pr(\hat{Y} = \hat{y}, A = a)$ with $a, \hat{y} \in \{0, 1\}$. Show that if $\hat{Y} \perp\!\!\!\perp A$ (demographic parity), then two of the four probabilities are sufficient to specify the joint distribution.

   (b) (2 points) Are *any* two of the four probabilities sufficient? Explain.

(c) (4 points) The separation criterion for fairness requires that $(\hat{Y} \perp\!\!\!\perp A) \mid Y$. Assuming separation holds, how many of the 8 probabilities $\Pr(\hat{Y} = \hat{y}, A = a, Y = y)$ with $\hat{y}, a, y \in \{0, 1\}$ are sufficient to define the joint distribution of $\hat{Y}, A, Y$? Explain.

(d) (4 points) Express the separation criterion as two equations involving conditional probabilities of the form $\Pr(\hat{Y} = \hat{y} \mid A = a, Y = y)$ with $\hat{y}, a, y \in \{0, 1\}$

5. (18 points) Given an expert policy π_E, GAIL finds a cost function $c_I(s,a)$ satisfying

$$c_I(s,a) = \arg\max_c \left(\min_\pi -H(\pi) + \mathbb{E}_\pi[c(s,a)] \right) - \mathbb{E}_{\pi_E}[c(s,a)] \tag{1}$$

and then an imitation policy π_I that satisfies:

$$\pi_I = \arg\min_\pi -H(\pi) + \mathbb{E}_\pi[c_I(s,a)] \tag{2}$$

where $H(\pi) = \mathbb{E}_\pi[-\log \pi]$. Expected values over policies mean expected values over trajectories generated by those policies.

(a) (6 points) Assume without loss of generality that the cost function $c_I(s,a)$ can be written as the negative log of a policy π_c,

$$c_I(s,a) = -\log \pi_c(a|s)$$

Show that the expression minimized during policy optimization, i.e. $-H(\pi)+\mathbb{E}_\pi[c_I]$, is a divergence between π and π_c

(b) (4 points) What value of π minimizes this divergence, and what is the minimum value of the divergence?

(c) (4 points) Now assume $-H(\pi)+\mathbb{E}_{\pi}[c_I]$ takes its minimum value, show that equation (1) simplifies to maximizing the negative of a cross-entropy between the distributions of two policies. In other words (1) minimizes the cross-entropy between two policies.

(d) (4 points) Finally conclude what π_c should be to optimize (1), i.e. to minimize the cross entropy in (c) above, and therefore what π_I should be to optimize (2).

6. (12 points) Q-Learning. The basic Q-Learning update value is:

$$Q'(s_t, a_t) = r(s_t, a_t) + \gamma \max_a Q(s_{t+1}, a) \tag{3}$$

and then Q-learning updates Q-values using a learning rate α as follows:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(Q'(s_t, a_t) - Q(s_t, a_t)) \tag{4}$$

(a) (8 points) Applying Q-learning directly leads to several sources of bias and overestimation of Q-values. Describe *two* approaches to reducing this bias, or overestimation.

(b) (4 points) Suppose all game epochs take N steps. How large should a reply buffer be to ensure that samples are equally likely to be drawn from step $1, \ldots, N$ of a game? Explain your reasoning.

7. (8 points) Structure of the Intrinsic Curiosity Module (ICM):

   (a) (4 points) The reward from curiosity is defined as $r_t^i = ||\hat{\phi}(s_{t+1}) - \phi(s_{t+1})||$, where ϕ is a featurization of s and $\hat{\phi}$ are forward predicted features. Suppose we instead predict the future state directly, and let $r_t^i = ||\hat{s}_{t+1} - s_{t+1}||$. What are the advantages of using the featurized prediction error instead of the state-based prediction error?

   (b) (4 points) The ICM also trains an inverse model to predict the action $\hat{a}_t$ from state representations $\phi(s_t)$ and $\phi(s_{t+1})$. What form would you expect the featurizing function ϕ if the action prediction loss were *not* used.

8. (14 points) Exploration in RL

   (a) (6 points) An epsilon-greedy exploration strategy with $\epsilon = 1$ is a random walk. Consider a discrete random walk where the actions are to move one step left or right (with equal probability) on an integer line $\mathbb{Z}$, starting from zero. Prove that after n steps, the expected distance reached from the start point is $O(\sqrt{n})$

   (b) (8 points) Give an example of an exploration strategy from lecture which would more efficiently explore a finite discrete space $[-N \dots N]$. Demonstrate how it explores. You can use a small value of N such as 2 or 3.