CS182/282A
Spring 2020
Take Home Quiz 2

Name: Keisuke Watanabe

SID: 3034049880

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.

Rewards for t timesteps.

We can reduce this variance with discount factor which reduces the effect of rewards far in the future.

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?

It prefers the rewards in near future and ignores the rewards in far from now.

If $\gamma < 1$, it will converge in a finite number of iterations.

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.

(b) It looses long term strategy.

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.

If $\hat{Y} \perp\!\!\!\perp A$, $P(\hat{Y}=\hat{y} \mid A=1) = P(\hat{Y}=\hat{y} \mid A=0)$

WLOG, pick $A=0$. If we know $P(\hat{Y}=1, A=0)$ and $P(\hat{Y}=0, A=0)$,

$P(\hat{Y}=\hat{y}, A=0) = p(\hat{Y}=\hat{y} \mid A=0)P(A=0)$ implies

$$\begin{aligned} P(\hat{Y}=\hat{y} . A=1) &= P(\hat{Y}=\hat{y} \mid A=1) P(A=1) \\ &= p(\hat{Y}=\hat{y} \mid A=0)(1-P(A=0)) \\ &= (p(\hat{Y}=\hat{y} . A=0)/p(A=0))(1-P(A=0)) \end{aligned}$$

Hence, we have 3 equations and 3 unknowns:

$$\begin{cases} P(\hat{Y}=1, A=1) = P(\hat{Y}=1, A=0)\left(\frac{1}{P(A=0)} - 1\right) \\ P(\hat{Y}=0, A=1) = P(\hat{Y}=0, A=0)\left(\frac{1}{P(A=0)} - 1\right) \\ \sum_{\hat{Y}\in\{0,1\}} \sum_{A\in\{0,1\}} P(\hat{Y}=\hat{y}, A=a) = 1 \end{cases}$$

Hence, we can obtain $P(\hat{Y}=1, A=1)$, $P(\hat{Y}=0, A=1)$ and $P(A=0)$

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

No since $P(\hat{Y}=1 \mid A=0) = P(\hat{Y}=1 \mid A=1)$ and $P(\hat{Y}=0 \mid A=0) = P(\hat{Y}=0 \mid A=1)$,

$(\hat{Y}, A) = (1,0)$ and $(1,1)$ or $(\hat{Y}, A) = (0,0)$ and $(0,1)$

gives redundant information. Hence,

we need one probability associated with $\hat{Y}=1$

and one probability associated with $\hat{Y}=0$.

(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.

4

$P(\hat{Y}=1|A=1,Y=1) = P(\hat{Y}=1|A=0,Y=1)$

$P(\hat{Y}=1|A=1,Y=0) = P(\hat{Y}=1| A=0,Y=0)$

$P(\hat{Y}=0|A=1,Y=1) = P(\hat{Y}=0|A=0,Y=1)$

$P(\hat{Y}=0|A=1,Y=0) = P(\hat{Y}=0|A=0,Y=0)$

Each line gives redundant information.

Hence, 4 pairs are sufficient. To define $P(\hat{Y}=\hat{y}, A=a, Y=y)$

(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 $\hat{y}, a, y \in \{0, 1\}$

$P(\hat{Y}=1|A=a,Y=1) = P(\hat{Y}=1|A=b,Y=1) \quad \forall a,b \in \{0,1\}$

$P(\hat{Y}=1|A=a,Y=0) = P(\hat{Y}=1|A=b,Y=0) \quad \forall a.b \in \{0,1\}$

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

$$\begin{aligned} -H(\pi) + \mathbb{E}_\pi(C_I) &= -\mathbb{E}_\pi[-\log \pi] + \mathbb{E}_\pi[-\log \pi_c(a|s)] \\ &= \mathbb{E}_\pi[\log \pi - \log \pi_c(a|s)] \\ &= \mathbb{E}_\pi\left[\log \frac{\pi}{\pi_c}\right] \\ &= D(\pi \parallel \pi_c) \end{aligned}$$

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

π_c minimizes this divergence.

The minimum value of the divergence is 0

(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.

$$C_I(s,a) = \arg\max_c \left(\cancelto{0}{\min_\pi -H(\pi) + E_\pi[c(s,a)]}\right) - E_{\pi_E}[c(s,a)]$$

$$= \arg\min_c E_{\pi_E}[c(s,a)]$$

$$= \arg\min_c -\sum_{\pi_E} \pi_E \log \pi_c(a|s)$$

$$= \arg\min_c H(\pi_E, \pi_c)$$

(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).

$\pi_c = \pi_E$ minimizes the cross entropy above.

Hence, $\pi_I = \pi_c = \pi_E$

$\therefore \pi_I = \pi_E$ 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 over-estimation of Q-values. Describe *two* approaches to reducing this bias, or over-estimation.

① Double Q-learning reduces the overestimation by using two separate Q-value estimators which are unbiased estimate for the actions because each one is updated with a value from the other Q function which results in a different set of samples.

② We can also prevent overestimation by taking the average of a few previous Q values and the current Q value to compute the next Q value.

(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.

let B be the size of the replay buffer.

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?

Even if the environment has very sparse rewards, the agent can learn good exploration policies and generalizable skills. It can also learn good exploration policies and generalizable skills in higher dimensional space such as images.

(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.

ϕ becomes simple transformation from state to feature space including all information about the state that are both relevent and irrelevent to the action.

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 $O(\sqrt{n})$

let d be the distame reached after n steps.

let s be a step taken at each step $s \in \{-1,1\}$

$$\begin{aligned} \mathbb{E}[d^2] &= \mathbb{E}[(\sum_{i=1}^{n} s_i)^2] \\ &= \mathbb{E}[\sum_{i=j}^{n} s_i s_j + 2\sum_{i>j} s_i s_j] \\ &= \sum_{i=j}^{n} \cancelto{1}{\mathbb{E}[s_i^2]} + 2\sum_{i>j} \cancelto{0}{\mathbb{E}[s_i s_j]} \\ &= n \end{aligned}$$

Hence, $\mathbb{E}[|d|] = O(\sqrt{n})$

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

optimistic exploration

let Bonus $B(N(s)) = \frac{1}{N(s)}$

$s_0 = 0, \; N = 2$

$$r^{+}(s,a) = r(s,a) + B(N(s))$$

$$\text{let } r(s,a) = \begin{cases} 1, & \text{if } (s,a) = (1,\text{left}), (0,\text{left}), (-1,\text{left}) \\ 10 & \text{if } (s,a) = (2,\text{left}) \\ -0.5 & \text{otherwise} \end{cases}$$

	[-2	-1	0	1	2]
optimistic exploration :	→	→	→	→	←
greedy :	→	←	←	←	←