University of Michigan, Ann Arbor
Department of Electrical Engineering and Computer Science

EECS 602 Winter 2024

Midterm Exam: Date: Aprl 4, 2024.

Instruction

- Time: 80 minutes
- Total points: 100.
- There are 5 questions in this exam.
- Open books and notes.
- You can use a calculator.
- You can use your laptop or iPad, but you should not access the Internet or use your cellphone. Your cellphone should be stored in your backpack.
- Print your name and UM ID neatly on every page.
- Please read every question carefully. If you have any doubt, do not hesitate to ask the instructor for a clarification.
- You have to show work to get full credit. Writing the correct answer with no work or with wrong work will earn no credit. Give a justification for every step of your solution. Excessive use of calculator to skip steps will result in partial credit.
- There is partial credit for the correct approach unless stated otherwise.
- RETURN the exam paper along with your answers.

1. **Honor Pledge** [10 points]

   Write and sign the honor pledge ("I have neither given nor received aid on this exam, nor have I concealed any honor code violations")

2. **SARSA [20 points]**

Consider a Markov decision process with three states $\{1, 2, 3\}$. In each state, we can choose one of the two possible actions $\{1, 2\}$. The transition probabilities and the mean rewards are unknown. We are interested in solving the following infinite-horizon discounted reinforcement learning problem

$$\max_{\mu} E\left[\sum_{k=0}^{\infty} 0.9^k r(x_k, \mu(x_k)) \middle| x_0\right],$$

where x_k is the state at time k, u_k is the action at time k, μ denotes a policy, and the discount factor is 0.9.

Consider the SARSA algorithm with $Q_0 = \begin{pmatrix} 0.1 & 0.9 \\ 0.2 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}$ and step size $\beta = 0.1$. Assume that under this SARSA algorithm, we have the following trace $(x_k, u_k, r(x_k, u_k))$:

$$(1, 2, 1) \to (2, 1, 2) \to (3, 2, 3) \to (2, 2, 0),$$

i.e.,

$$x_0 = 1, u_0 = 2, r(x_0, u_0) = 1, x_1 = 2, u_1 = 1, r(x_1, u_1) = 2,$$
$$x_2 = 3, u_2 = 2, r(x_2, u_2) = 3, x_3 = 2, u_3 = 2, r(x_3, u_3) = 0.$$

Please calculate the sequence of Q-values under this SARSA algorithm.

Solution: Given $(x_k, u_k, r(x_k, u_k), x_{k+1}, u_{k+1})$, the update formula of SARSA is as follows:

$$Q_{t+1}(x_k, u_k) = (1 - \beta) Q_t(x_k, u_k) + \beta \left[r(x_k, u_k) + 0.9 Q_t(x_{k+1}, u_{k+1}) \right].$$

Given the trace, the sequence of Q-values under this SARSA algorithm is as follows:

$$Q_1(1,2) = (1-\beta) * Q_0(1,2) + \beta * (r(1,2) + 0.9 * Q_0(2,1)) = 0.928; Q_1 = \begin{pmatrix} 0.1 & 0.928 \\ 0.2 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}$$

$$Q_2(2,1) = (1-\beta) * Q_1(2,1) + \beta * (r(2,1) + 0.9 * Q_1(3,2)) = 0.425; Q_2 = \begin{pmatrix} 0.1 & 0.928 \\ 0.425 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}$$

$$Q_3(3,2) = (1-\beta) * Q_2(3,2) + \beta * (r(3,2) + 0.9 * Q_2(2,2)) = 0.777; Q_3 = \begin{pmatrix} 0.1 & 0.928 \\ 0.425 & 0.3 \\ 0.3 & 0.777 \end{pmatrix}$$

3. **UCB Algorithm for Multi-Armed Bandit [20 points]**

Consider a multi-armed bandit problem with two arms, arm 1 and arm 2. Assume that the rewards are Bernoulli. Consider the upper confidence bound (UCB) algorithm in class with the UCB bonus term $\sqrt{\frac{4\log_2 K}{N_i(k)}}$, where K is the total number of time steps and $N_i(k)$ is the number of times arm i has been played by time step k (not including k). Let $K = 1024$.

Suppose that at the beginning of time step $k = 411$, $N_1(k) = 250$, $N_2(k) = 160$, the empirical mean of arm 1's reward is 0.6, and the empirical mean of arm 2's reward is 0.4.

(1) Please determine which arm the UCB algorithm will pull at this time step. Please include the detailed steps that lead to your answer.

(2) Suppose that we obtain a reward of 0 after pulling the arm in this time step. Please update the empirical mean of arm 1 and arm 2. (Round to four decimal points.)

(3) Suppose that the true mean of arm 1's reward is 0.55 and the true mean of arm 2's reward is 0.45. Suppose that the algorithm pulled a sequence of arms as follows:

$$1, 1, 2, 2, 1, 1, 2, 1, 2, 1.$$

If we consider only these 10 time steps, what is the expected regret?

Solution:

(1) Given empirical means $\widehat{\mu}_1 = 0.6$ and $\widehat{\mu}_2 = 0.4$, define

$$\text{UCB}(i) = \widehat{\mu}_i + \sqrt{\frac{4\log K}{N_i(k)}}$$

Then we want to compute

$$i_k = \arg\max_i \text{UCB}(i)$$

Note that

$$\text{UCB}(1) = 0.6 + \sqrt{\frac{4\log 1024}{250}} = 1$$
$$\text{UCB}(2) = 0.4 + \sqrt{\frac{4\log 1024}{160}} = 0.9$$

Thus, $i_k = 1$.

(2) Since arm 1 is pulled, this is the only arm that will be updated. The empirical mean will become

$$\widehat{\mu}_i = \frac{250 * 0.6 + 0}{251} = 0.5976$$

Thus, the new empirical means are $\widehat{\mu}_1 = 0.5976$ and $\widehat{\mu}_2 = 0.4$

(3) Recall that the regret is defined as

$$\begin{aligned} R(K) &= \mu^* K - \mathbb{E}\left[\sum_{i=1}^{K} \mu_{a(i)}\right] \\ &= 10(0.55) - [6(0.55) + 4(0.45)] \\ &= 4(0.55 - 0.45) \\ &= 0.4 \end{aligned}$$

4. **Deterministic Policy Gradient [25 points]**

In this problem, we are interested in solving the infinite horizon discounted-reward problem

$$\max_{\mu} E\left[\sum_{k=0}^{\infty} 0.9^k r(x_k, \mu(x_k))\middle| x_0\right]$$

for an MDP that has both continuous state space $\mathcal{S} = \mathbb{R}$ and action space $\mathcal{A} = [0, 1]$. We now have a parametrized actor

$$\mu_\theta(x) = \frac{1}{1 + e^{-\theta x}}$$

and Q function

$$Q_w(x, u) = \left(\frac{1}{2}(x + wu) - 1\right)^2$$

where the current value of $\theta = 0.5$ and $w = 1$. Given the following trajectory of state transitions (x_t):

$$(x_0 = 2.0) \rightarrow (x_1 = 4.0) \rightarrow (x_2 = 2.0)$$

Assuming no updates are made to w, please calculate the deterministic policy gradient

$$\nabla_\theta Q(x_0, \mu_\theta(x_0)).$$

for $x_0 = 2.0$ with respect to this trajectory.

Please use the approximation $e = 3$ in this problem. You can round to four decimal points.

Solution: From DPG theorem, we want to compute a monte-carlo estimate of

$$\mathbb{E}_x\left[\sum_{k=0}^{\infty} \alpha^k \nabla_u Q_w(x, u)|_{\mu_\theta(x_k)} \nabla_\theta \mu_\theta(x_k)\right]$$

For this trajectory, this yields

$$\nabla_\theta Q(x_0, \mu_\theta(x_0)) = \sum_{k=0}^{2} \alpha^k \nabla_u Q_w(x, u)|_{\mu_\theta(x_k)} \nabla_\theta \mu_\theta(x_k)$$

Therefore, we must compute each of the necessary quantities. First, we will compute both the necessary gradients

$$\nabla_u Q_w(x, u) = \frac{w}{2}(x + wu) - w$$

and

$$\nabla_\theta \mu_\theta(x) = \frac{x e^{-\theta x}}{(1 + e^{-\theta x})^2}$$

For $\theta = 0.5$ and $w = 1$, we have that

$$\nabla_u Q_w(x,u) = \frac{1}{2}(x+u) - 1 \quad \text{and} \quad \nabla_\theta \mu_\theta(x) = \frac{xe^{-x/2}}{(1+e^{-x/2})^2}$$

To evaluate the gradient of Q, we need the action taken by our deterministic policy i.e.

$$\mu_\theta(x_0 = 2.0) = \frac{1}{1+3^{-1}} = \frac{3}{4}$$
$$\mu_\theta(x_0 = 4.0) = \frac{1}{1+3^{-2}} = \frac{9}{10}$$

Then

$$\nabla_u Q_w(x_0,u)|_{\mu_\theta(x_0)} = \frac{1}{2}\left(2 + \frac{3}{4}\right) - 1 = \frac{3}{8}$$
$$\nabla_u Q_w(x_1,u)|_{\mu_\theta(x_1)} = \frac{1}{2}\left(4 + \frac{9}{10}\right) - 1 = 1 + \frac{9}{20}$$

and

$$\nabla_\theta \mu_\theta(x_0 = 2.0) = \frac{2e^{-1}}{(1+e^{-1})^2} = \frac{2/3}{16/9} = \frac{3}{8}$$
$$\nabla_\theta \mu_\theta(x_0 = 4.0) = \frac{4e^{-2}}{(1+e^{-2})^2} = \frac{4/9}{100/81} = \frac{36}{100}$$

Thus,

$$\begin{aligned}\nabla_\theta Q(x_0, \mu_\theta(x_0)) &= \sum_{k=0}^{2} \alpha^k \nabla_u Q_w(x,u)|_{\mu_\theta(x_k)} \nabla_\theta \mu_\theta(x_k) \\ &= \left(\frac{3}{8}\right)^2 + (0.9)\left(1 + \frac{9}{20}\right)\left(\frac{36}{100}\right) + (0.9)^2\left(\frac{3}{8}\right)^2 = 0.7243\end{aligned}$$

5. **Monte Carlo Tree Search in AlphaZero [25 points]**

Suppose that we are using the AlphaZero algorithm for a simplified Go game (a two-player game). During a single simulation of the Monte Carlo Tree Search using self-play, we obtained the following data samples:

$$(s_1, a_1) \rightarrow (s_2, a_2) \rightarrow (s_3),$$
$$\text{Player 1} \quad \text{Player 2} \quad \text{Player 1}$$

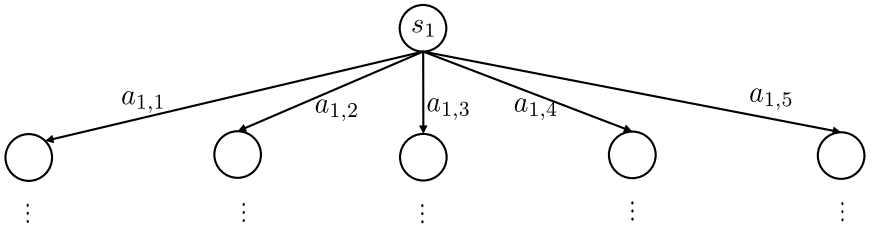
where s_3 is a terminal state (the game ends) so the simulation terminates. Suppose Player 2 won. The reward is 1 when a player wins the game and -1 when a player loses the game. Suppose that the visit counts and the Q-value for each (state, action) pair before this simulation is as shown in the following table.

	(s_1, a_1)	(s_2, a_2)
$N(s, a)$	12	4
$Q(s, a)$	-0.5	0.8

(1) Please calculate the updated Q-values and visit counts after this one simulation and fill in the blanks in the following table. Please include the detailed steps that lead to your answer.

	(s_1, a_1)	(s_2, a_2)
$N(s, a)$		
$Q(s, a)$		

(2) Suppose the visit counts after many simulations starting from s_1 are shown in the following figure and table. Please calculate the probabilities of taking different actions in state s_1, i.e., $\pi(a_{1,1}|s_1), \pi(a_{1,2}|s_1), \pi(a_{1,3}|s_1), \pi(a_{1,4}|s_1), \pi(a_{1,5}|s_1)$.



	$(s_1, a_{1,1})$	$(s_1, a_{1,2})$	$(s_1, a_{1,3})$	$(s_1, a_{1,4})$	$(s_1, a_{1,5})$
$N(s, a)$	13	14	10	8	5

Solution:

(1) We update the $Q(s,a)$ and $N(s,a)$ as follows:

$$\begin{aligned} Q(s_2,a_2) \leftarrow & \frac{N(s_2,a_2)Q(s_2,a_2)+1}{N(s_2,a_2)+1} = \frac{4\times 0.8+1}{4+1} = 0.84 \\ N(s_2,a_2) \leftarrow & N(s_2,a_2)+1 = 4+1 = 5 \end{aligned}$$

$$\begin{aligned} Q(s_1,a_1) \leftarrow & \frac{N(s_1,a_1)Q(s_1,a_1)-1}{N(s_1,a_1)+1} = \frac{12\times(-0.5)-1}{12+1} = -\frac{7}{13} \\ N(s_1,a_1) \leftarrow & N(s_1,a_1)+1 = 12+1 = 13 \end{aligned}$$

The table is as follows:

(2) Under MCTS algorithm, we know that the probabilities of taking action a in state s_1 is

$$\pi(a|s_1) = \frac{N(s_1,a)}{\sum_{a'} N(s_1,a')}.$$

Note that $\sum_{a'} N(s_1,a') = 13+14+10+8+5 = 50$. Hence, we have

$$\begin{aligned} \pi(a_{1,1}|s_1) = & \frac{13}{50} = 0.26 \\ \pi(a_{1,2}|s_1) = & \frac{14}{50} = 0.28 \\ \pi(a_{1,3}|s_1) = & \frac{10}{50} = 0.20 \\ \pi(a_{1,4}|s_1) = & \frac{8}{50} = 0.16 \\ \pi(a_{1,5}|s_1) = & \frac{5}{50} = 0.10. \end{aligned}$$