CAP 6629 Reinforcement Learning

Exam

Thursday, November 09, 2023
11:00am-12:45pm
(90mins for the exam + 15mins for uploading your results on Canvas).

- The exam will be open book, open notes with the detailed information below. **Any internet search tool or AI generating/search tool are NOT ALLOWED. Violations will result in zero for the exam.**
    - Book: Andrew Barto and Richard S. Sutton, Reinforcement Learning: An Introduction, MIT Press.
    - Notes: Lecture notes and lecture slides for this course.
- Work on yourself. DO NOT DISCUSS WITH OTHERS! **Any identical or highly similar solving process will not be graded.**
- It is your responsibility to make sure the scanned version **is clear and readable**.
- The solving process and steps are important and need to be provided.
- Submit through Canvas in a SINGLE copy.
- You are not allowed to share and/or upload the exam to any other platform.

Name (Print): Matthew ACS

GOOD LUCK!!!

Problem 1: (25 points)

Consider the continuing MDP shown in Fig.1. The only decision to be made is that in the middle state (state B), where two actions are available, *left* and *right*. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, π_{left} and π_{right}.

Determine the optimal policy (starting at B) when
(1) $\gamma = 0$
(2) $\gamma = 0.5$
(3) $\gamma = 0.75$
(4) $\gamma = 0.95$
where γ is the discount factor. Please show the solving steps.

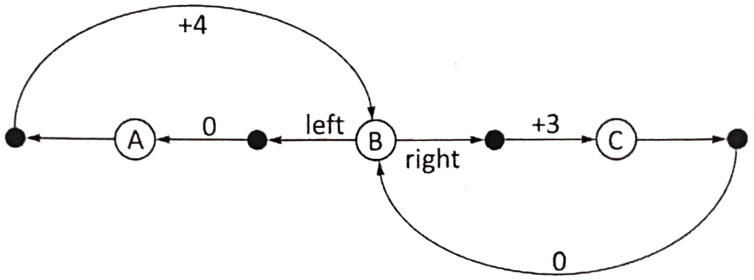


Fig.1

Note: Circles are the states and dots are the actions.
Hint: $1 + x^2 + x^4 + \cdots = \frac{1}{1-x^2}$

1) Right - $G_t = (3) + (0)(0) + (0)^2(3) \ldots = 3$

Left - $G_t = (0) + (0)(4) + (0)^2(0) \ldots = 0$

$\boxed{\text{For } \gamma = 0,\ \pi_{right} \text{ is the optimal policy}}$

2) Right - $G_t = (3) + (0.5)(0) + (0.5)^2(3) \ldots$

$= (3) + (0.5)^2(3) + (0.5)^4(3) \ldots$

$= 3\left(\frac{1}{1-(0.5)^2}\right) = \frac{3}{0.75} = 4$

Left - $G_t = (0) + (0.5)(4) + (0.5)^2(0) \ldots$

$= (0.5)(4) + (0.5)^3(4) + (0.5)^5(4) \ldots$

$= 4\left(\sum_{k=0}^{\infty} (0.5)^{2k+1}\right) = 4(0.666) = 2.66$

For $\gamma = 0.5$, π_{right} is the optimal policy

3) Right − $G_t = 3\left(\frac{1}{1-(0.75)^2}\right) = \frac{3}{0.4375} = 6.857$

Left − $G_t = 4\left(\sum_{k=0}^{\infty} (0.75)^{2k+1}\right) = 4(1.71) = 6.857$

For $\gamma = 0.75$, both π_{right} and π_{left} are optimal policies

4) Right − $G_t = 3\left(\frac{1}{1-(0.95)^2}\right) = \frac{3}{0.0975} = 30.769$

Left − $G_t = 4\left(\sum_{k=0}^{\infty} (0.95)^{2k+1}\right) = 4(9.744) = 38.976$

For $\gamma = 0.95$, π_{left} is the optimal policy

$$\sum_{k=0}^{\infty} (\gamma)^{2k+1} = \gamma\left(\frac{1}{1-\gamma^2}\right)$$

Problem 2: (20 points)

Consider a k-armed bandit problem with k=4 actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ε-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = 1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = 2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the ε case may have occurred, causing an action to be selected at random.

(1) Make a table including all the action-value estimates and the set of greedy actions for each time step in the initial sequence.

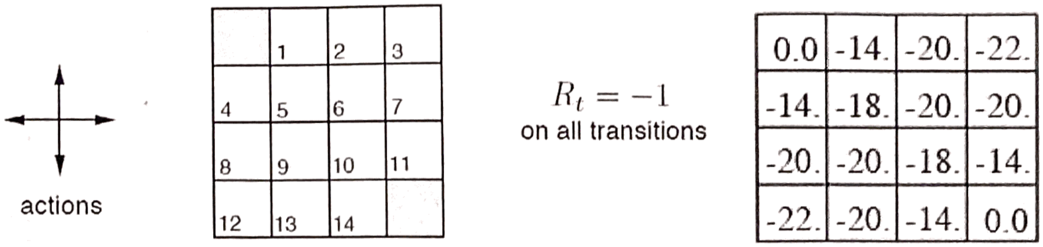
t	$Q_t(1)$	$Q_t(2)$	$Q_t(3)$	$Q_t(4)$	$\{A_t^*\}$	A_t	ε	R_t
1	0	0	0	0	$\{1,2,3,4\}$	1	maybe	1
2	1	0	0	0	$\{1\}$	2	yes	1
3	1	1	0	0	$\{1,2\}$	2	maybe	2
4	1	1.5	0	0	$\{2\}$	2	maybe	2
5	1	1.667	0	0	$\{2\}$	3	yes	0

(2) On which time steps did the ε case definitely occur? On which time steps could this possibly have occurred?

On time steps 2 and 5 the ε case definitely occured because the action chosen was not in the set of optimal actions. On time steps 1, 3, and 4 it could have possibly occured because the action chosen was in the set of optimal actions, but it could have been chosen due to the ε-case.

Problem 3: (30 points)

Consider the gridworld shown in Fig. 2(a).



	1	2	3
4	5	6	7
8	9	10	11
12	13	14	

(a)

0.0	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	0.0

(b)

Fig. 2

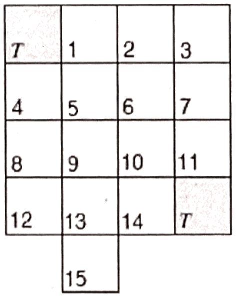
The nonterminal states are $S = \{1, 2, \dots, 14\}$. There are four actions possible in each state, $A = \{left, up, right, down\}$, which deterministically cause the corresponding state transitions, except that actions that would take the agent off the grid in fact leave the state unchanged. This is an undiscounted, episodic task. The reward is -1 on all transitions until the terminal state is reached. The terminal state is shaded in the figure (although it is shown in two places, it is formally one state). Now, we have achieved the state-value function as in Fig. 2(b) based on iterative policy evaluation.

Now a new state 15 is added to the gridworld as Fig. 3 and its actions, *left*, *up*, *right*, and *down*, take the agent to states 12, 13, 14, and 15, respectively. Suppose now the transition function of agent movement happens with probability (0.6, 0.1, 0.2, 0.1), i.e., *left*(60%), *up*(10%), *right*(20%), and *down*(10%).

Follows this policy, what is $v_\pi(15)$?

$V_\pi(15) = -21.\overline{11}$

$$\boxed{V_\pi(15) = -21.\overline{11}}$$



T	1	2	3
4	5	6	7
8	9	10	11
12	13	14	T
	15		

Fig. 3

$V_1(15) = (0.6)(-1+(-22)) + (0.1)(-1+(-20)) + (0.2)(-1+(-14))$
$+ (0.1)(-1+(0)) = -18.9 + -0.1 = -19$

$V_2(15) = -18.9 + (0.1)(-1+(-19)) = -20.9$

$V_3(15) = -18.9 + (0.1)(-1+(-20.9)) = -21.09$

$V_4(15) = -18.9 + (0.1)(-1+(-21.09)) = -21.109$

Problem 4: (25 points)

(1) In the class, we studied the logistic function (sigmoid unit), which is used for positive output of a neural network node. Consider the bilateral output (both positive and negative), you may have to use the function

$$\boxed{\frac{d}{dx}(h(x)) = 1-(h(x))^2}$$

$$h(x) = \tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Please derive the partial derivatives of *h(x)* with respective to *x*

Hint: as you did in the lecture, your results should be in terms of h(x), nothing else.

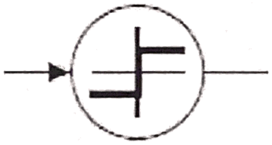
$$\frac{d}{dx}\left(\frac{e^x-e^{-x}}{e^x+e^{-x}}\right) = \frac{(e^x+e^{-x})(e^x+e^{-x})-(e^x-e^{-x})(e^x-e^{-x})}{(e^x+e^{-x})^2}$$

$$= \frac{4}{(e^x+e^{-x})^2} = \left(\frac{2}{(e^x+e^{-x})}\right)^2 = \left(\frac{1}{\cosh(x)}\right)^2 = \frac{1}{\cosh^2(x)} = \operatorname{Sech}^2(x)$$

$$\operatorname{Sech}^2(x) = 1-\tanh^2(x) = 1-(h(x))^2$$

(2) Briefly discuss why we need to replace the threshold function (presented in below) with the logistic (sigmoid) function.

Hint: here is a picture of the threshold function we learned in class



The threshold function is non-differentiable while the sigmoid function is. This allows neural networks to be trained through backpropagation via gradient based optimization. Additionally, the output of a sigmoid function represents probabilities, which are useful in situations such as binary classification problems.