CS 188 Introduction to
Summer 2020 Artificial Intelligence

Exam Prep 3

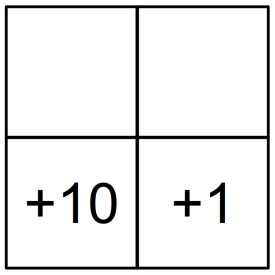
Q1. Feature-Based Q-Learning

1. When using features to represent the Q-function is it guaranteed that the feature-based Q-learning finds the same optimal $Q*$ as would be found when using a tabular representation for the Q-function?

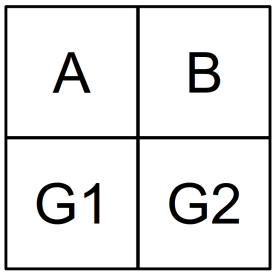
Q2. Q-learning

Consider the following gridworld (rewards shown on left, state names shown on right).

Rewards



State names



From state A, the possible actions are right(→) and down(↓). From state B, the possible actions are left(←) and down(↓). For a numbered state (G1, G2), the only action is to exit. Upon exiting from a numbered square we collect the reward specified by the number on the square and enter the end-of-game absorbing state X. We also know that the discount factor $\gamma = 1$, and in this MDP all actions are **deterministic** and always succeed.

Consider the following episodes:

Episode 1 ($E1$)

s	a	s'	r
A	↓	$G1$	0
$G1$	exit	X	10

Episode 2 ($E2$)

s	a	s'	r
B	↓	$G2$	0
$G2$	exit	X	1

Episode 3 ($E3$)

s	a	s'	r
A	→	B	0
B	↓	$G2$	0
$G2$	exit	X	1

Episode 4 ($E4$)

s	a	s'	r
B	←	A	0
A	↓	$G1$	0
$G1$	exit	X	10

(a) Consider using temporal-difference learning to learn $V(s)$. When running TD-learning, all values are initialized to zero. For which sequences of episodes, if repeated infinitely often, does $V(s)$ converge to $V^*(s)$ for all states s?

(Assume appropriate learning rates such that all values converge.)
Write the correct sequence under "Other" if no correct sequences of episodes are listed.

☐ $E1, E2, E3, E4$ ☐ $E1, E2, E1, E2$ ☐ $E1, E2, E3, E1$ ☐ $E4, E4, E4, E4$
☐ $E4, E3, E2, E1$ ☐ $E3, E4, E3, E4$ ☐ $E1, E2, E4, E1$

☐ Other _____

(b) Consider using Q-learning to learn $Q(s, a)$. When running Q-learning, all values are initialized to zero.
For which sequences of episodes, if repeated infinitely often, does $Q(s, a)$ converge to $Q^*(s, a)$ for all state-action pairs (s, a)

(Assume appropriate learning rates such that all Q-values converge.)
Write the correct sequence under "Other" if no correct sequences of episodes are listed.

☐ $E1, E2, E3, E4$ ☐ $E1, E2, E1, E2$ ☐ $E1, E2, E3, E1$ ☐ $E4, E4, E4, E4$
☐ $E4, E3, E2, E1$ ☐ $E3, E4, E3, E4$ ☐ $E1, E2, E4, E1$

☐ Other _____

Q3. Reinforcement Learning

Imagine an unknown game which has only two states $\{A, B\}$ and in each state the agent has two actions to choose from: {Up, Down}. Suppose a game agent chooses actions according to some policy π and generates the following sequence of actions and rewards in the unknown game:

t	s_t	a_t	s_{t+1}	r_t
0	A	Down	B	2
1	B	Down	B	-4
2	B	Up	B	0
3	B	Up	A	3
4	A	Up	A	-1

Unless specified otherwise, assume a discount factor $\gamma = 0.5$ and a learning rate $\alpha = 0.5$

(a) Recall the update function of Q-learning is:

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha(r_t + \gamma \max_{a'} Q(s_{t+1}, a'))$$

Assume that all Q-values initialized as 0. What are the following Q-values learned by running Q-learning with the above experience sequence?

$Q(\text{A, Down}) =$ _____, $Q(\text{B, Up}) =$ _____

(b) In model-based reinforcement learning, we first estimate the transition function $T(s, a, s')$ and the reward function $R(s, a, s')$. Fill in the following estimates of T and R, estimated from the experience above. Write "n/a" if not applicable or undefined.

$\hat{T}(\text{A, Up, A}) =$ _____, $\hat{T}(\text{A, Up, B}) =$ _____, $\hat{T}(\text{B, Up, A}) =$ _____, $\hat{T}(\text{B, Up, B}) =$ _____

$\hat{R}(\text{A, Up, A}) =$ _____, $\hat{R}(\text{A, Up, B}) =$ _____, $\hat{R}(\text{B, Up, A}) =$ _____, $\hat{R}(\text{B, Up, B}) =$ _____

(c) To decouple this question from the previous one, assume we had **a different experience** and ended up with the following estimates of the transition and reward functions:

s	a	s'	$\hat{T}(s, a, s')$	$\hat{R}(s, a, s')$
A	Up	A	1	10
A	Down	A	0.5	2
A	Down	B	0.5	2
B	Up	A	1	-5
B	Down	B	1	8

(i) Give the optimal policy $\hat{\pi}^*(s)$ and $\hat{V}^*(s)$ for the MDP with transition function $\hat{T}$ and reward function $\hat{R}$.
Hint: for any $x \in \mathbb{R}$, $|x| < 1$, we have $1 + x + x^2 + x^3 + x^4 + \cdots = 1/(1-x)$.

$\hat{\pi}^*(A) =$ _____, $\hat{\pi}^*(B) =$ _____, $\hat{V}^*(A) =$ _____, $\hat{V}^*(B) =$ _____.

(ii) If we repeatedly feed this new experience sequence through our Q-learning algorithm, what values will it converge to? Assume the learning rate α_t is properly chosen so that convergence is guaranteed.

- ◯ the values found above, $\hat{V}^*$
- ◯ the optimal values, V^*
- ◯ neither $\hat{V}^*$ nor V^*
- ◯ not enough information to determine