The Hebrew University of Jerusalem Introduction to Artificial Intelligence Problem Set 4 Reinforcement Learning, Game Theory

Due Date 35.06.20$\widetilde{70}$20

1 Question 1 - Q Learning

[34 points]

An enterprising student builds a q-learning agent to buy and sell real estate. The agent begins owning Nothing. When the agent owns Nothing, it can try to buy land ($BuyL$) or buy a mansion ($BuyM$). If successful, the agent then owns Land or a Mansion. When the agent owns property, he can only try to $Sell$. Once the property is sold, the scenario ends in the Terminal state. Assume a discount factor of $\gamma = 1$.

1. At first, the student makes decisions for two episodes and the agent just learns $Q(s, a)$. Fill in the table with Q-value estimates after each observation. Use a learning rate of $\alpha = 0.5$. All estimates begin at 0. Terminal state T has no actions and a value of 0.

Ovservation				**New Q Esitmates**			
State s	**Action** a	**Successor** s'	**Reward** r	$Q(N, BuyL)$	$Q(N, BuyM)$	$Q(L, S)$	$Q(M, S)$
Ininital values:				0	0	0	0
N	$BuyL$	L	-10				
L	$Sell$	T	20				
N	$BuyL$	N	-2				
N	$BuyM$	M	-20				
N	$Sell$	M	-2				
M	$Sell$	T	60				

2. What is the optimal policy from this q-learning agent, and what is the q-learning agent's estimate of the expected sum of discounted future rewards starting in N under this policy?

3. If we instead use these observations to estimate a transition model, what would it be? Fill out the transition probabilities in the following table.

$T(N, BuyL, N) =$		$T(L, Sell, L) =$	
$T(N, BuyL, L) =$		$T(L, Sell, T) =$	
$T(N, BuyM, N) =$		$T(M, Sell, M) =$	
$T(N, BuyML, M) =$		$T(M, Sell, T) =$	

4. If the same sequence of 6 observations in 1 repeated indefinitely and the q-learner very slowly decreased its learning rate, what would $Q^{\star}(N, BuyM)$ converge to? Justify your answer.

2 Question 2 - Reinforcement learning

[33 points] Given the deterministic world in figure 1 where:

- Possible moves shown by arrows
- Reward is indicated by numbers near the arrows (or zero if there is no number). For example, the reward of moving from the bottom right-state to the top-right state is 20.

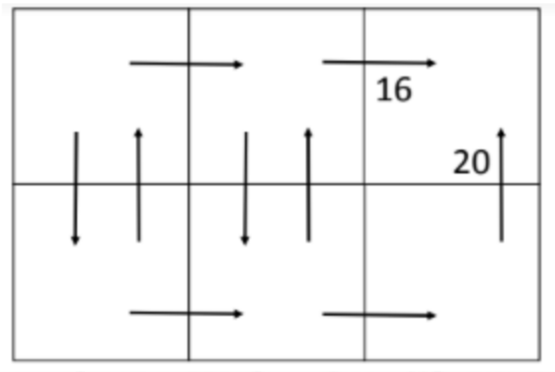

Figure 1: Deterministic world for question 2

1. Find the optimal $Q^{\star}(s, a)$ values for the discount factors 0.5 and 0.9, and fill them in in figures 2 and 3 respectively.
2. Based on your previous answer, find the optimal $V^{\star}(s)$ values for the discount factors 0.5 and 0.9 and fill them in on figures 4 and 5 respectively. Then mark the optimal policy on the figures in an obvious way (i.e. by circling or adding arrows).
3. In general, if we are doing reinforcement learning, why would we prefer an agent to get instantaneous rewards (in each state) which are good indicators of total rewards as opposed to getting all rewards at the end of the game? Give an answer which holds even if the discount factor is 1.

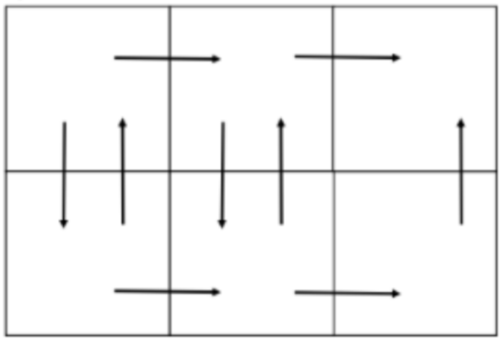

Figure 2: Fill in the $Q^\star$ values for $\gamma = 0.5$.

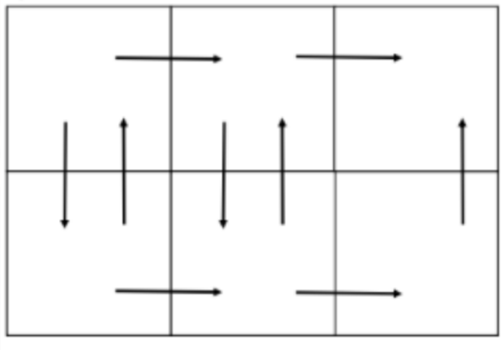

Figure 3: Fill in the $Q^\star$ values for $\gamma = 0.9$.

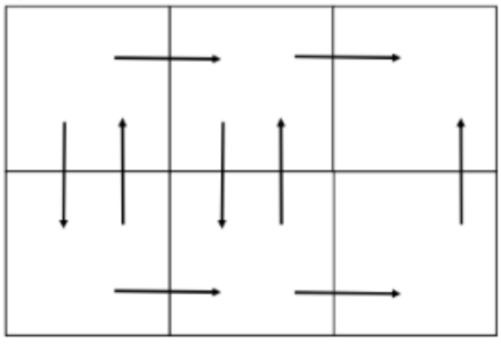

Figure 4: Fill in the $V^\star$ values for $\gamma = 0.5$ and mark the optimal policy.

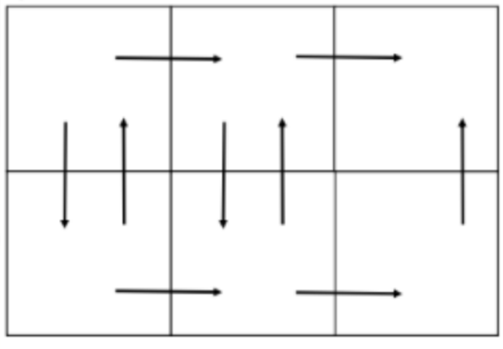

Figure 5: Fill in the $V^\star$ values for $\gamma = 0.9$ and mark the optimal policy.

3 Game Theory

3.1 Question 3

[15 points]

Show that a dominant strategy equilibrium is a Nash equilibrium, but not vice versa.

3.2 Question 4

[18 Points]

Consider the following normal form game:

Player 1 (row) can choose between actions A, B, and C, while Player 2 (column) can choose between a, b and c. The utility of an outcome is listed in the bottom left corner of the square for player 1, and in the top right for player 2.

	a	b	c
A	3 4	0 7	2 2
B	1 1	6 5	2 0
C	2 0	7 9	8 3

1. List all of the pure strategy Nash Equilibrium of this game.
2. List all of the dominated strategies for either player in this game.
3. Draw the game which results when all of the dominated strategies are removed.