COMP.4200/5430, Introduction to Artificial Intelligence
Fall 2019

Exam 2 – November 14, 2019

Name: PHONG VO # 01790283

This is a closed book exam. Notes and computational aids are not allowed.

Problem	Topic	Possible Points	Your Score
3	MDPs I	10	10
4	MDPs II	10	10
Total		20	20

Good!

(10 points.) **MDPs and Reinforcement Learning**

Consider an autonomous robot which can either move FAST or SLOW in any time step. Moving FAST generally gives a reward of +2, while moving SLOW gives a reward of only +1. However, the robot must also take into account its internal temperature, which can be either HOT or OK. Driving SLOW tends to lower the temperature, while driving FAST tends to raise it. If the robot is HOT, there is a danger if it overheating, at which point it must stop, cool down, and make repairs. The MDP transitions and rewards are specified as follows:

s	a	s'	$T(s,a,s')$	$R(s,a,s')$
OK	SLOW	OK	1.0	+1
OK	FAST	OK	0.5	+2
OK	FAST	HOT	0.5	+2
HOT	SLOW	OK	1.0	+1
HOT	FAST	HOT	0.5	+2
HOT	FAST	OK	0.5	-10

Note that while repairs are costly, the robot is OK afterwards (the last row in the table).

(1) (5 pts): Run two rounds of value iteration in the table below, using a discount of 0.8. You may skip the greyed-out square.

s	V_0	V_1	V_2
OK	0	2 ✓	3.2 ✓
HOT	0	1 ✓	

5

(1) (5 pts): Run Q-learning with a discount of 0.8 and a learning rate of 0.5, using the transition samples below. Do not copy over q-values which have not changed in a given step.

Assume the agent experiences the samples:

OK, FAST, HOT, reward +2, calculate Q_1
HOT, FAST OK, reward -10, calculate Q_2
OK, SLOW, OK, reward +1, calculate Q_3

s	a	Q_0	Q_1	Q_2	Q_3
OK	SLOW	0	—	—	.9 ✓
OK	FAST	0	1.0 ✓		—
HOT	SLOW	0			
HOT	FAST	0		-4.6 ✓	

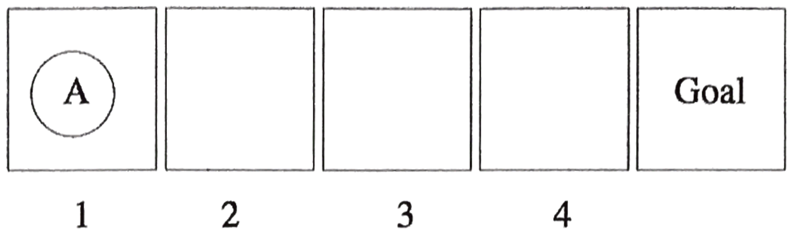
5

$$V_{k+1}(s) = \max_{\alpha'} \sum_{s'} T(s,\alpha,s')\left[R(s,a,s') + \gamma * V(s')\right]$$

$$Q_1 = Q_0 + \alpha\left[R(s,a,s\right.$$

2. (10 points) MDPs: Robot Soccer

A soccer robot **A** is on a fast break toward the goal, starting in position 1. From positions 1 through 3, it can either shoot (S) or dribble the ball forward (D); from 4 it can only shoot. If it shoots, it either scores a goal (state G) or misses (state M). If it dribbles, it either advances a square or loses the ball, ending up in M.



In this MDP, the states are 1, 2, 3, 4, G and M, where G and M are terminal states. The transition model depends on the parameter y, which is the probability of dribbling success. Assume a discount of $\gamma = 1$.

$T(k,S,G) = \frac{k}{6}$	$T(k,S,M) = 1 - \frac{k}{6}$	for $k \in \{1,2,3,4\}$
$T(k,D,k+1) = y$	$T(k,D,M) = 1 - y$	for $k \in \{1,2,3\}$
$R(k,S,G) = 1$	for $k \in \{1,2,3,4\}$,	and rewards are 0 for all other transitions

(a) (2 pt) What is $V^\pi(1)$ for the policy π that always shoots?

$$V^\pi(1) = T(1,S,G)\cdot R(1,S,G) + T(1,S,M)\cdot R(1,S,M) = 1/6$$
$$= 1/6 \cdot 1 + \frac{5}{6} \cdot 0$$

(b) (2 pt) What is $Q^*(3,D)$ in terms of y?

$$Q^*(3,D) = T(3,D,4)\cdot[R(3,D,4) + V^*(4)] + T(3,D,M)\cdot R(3,D,M)$$
$$= y(0 + 2/3) + (1-y)\cdot 0 = \frac{2}{3}y$$

(c) (2 pt) Using $y = \frac{3}{4}$, complete the first two iterations of value iteration.

i	$V_i^*(1)$	$V_i^*(2)$	$V_i^*(3)$	$V_i^*(4)$
0	0	0	0	0
1	1/6	1/3	1/2	2/3
2	1/4	3/8	1/2	2/3

(d) (2 pt) After how many iterations will value iteration compute the optimal values for all states?

It will do after 3 iterations.

(e) (2 pt) For what range of values of y is $Q^*(3,S) \geq Q^*(3,D)$?

$$Q^*(3,S) \geq Q^*(3,D)$$
$$\Leftrightarrow T(3,S,G)\cdot 1 \geq T(3,D,4)\cdot T(4,S,G)\cdot 1$$
$$\Leftrightarrow \frac{1}{2} \geq y \cdot \frac{2}{3} \Leftrightarrow y \leq \frac{3}{4}$$

3

$$\Rightarrow 0 \leq y \leq 3/4$$