Entry number - 2015CS50282 **Name -** [redacted]

COL 333/671 Autumn 2015 Minor 2

Welcome to minor 2. The exam is for 1 hour 10 minutes. Please use only pens while answering questions. Do not use a pencil.

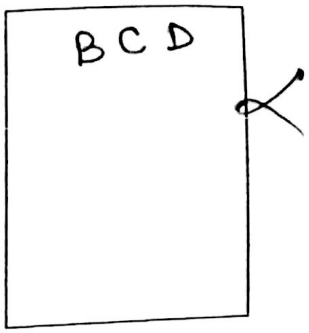
Questions numbered 1-13 are two points each. Answer in the box provided.

Question Number	Maximum Marks	Marks Obtained
1-16	40	19.5 + 3 = 22.5
17	10	2
18	15	00
19	35	23

47.5

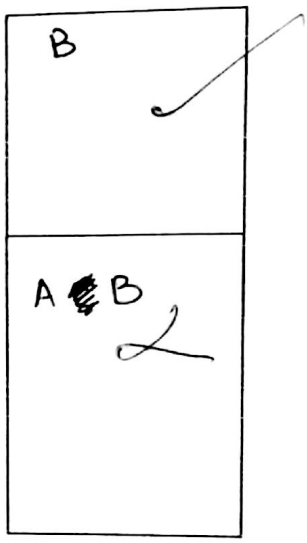
1. What all is true about discount factor in MDPs

(A) its use in MDPs is related to the idea of discounting objects to attract customers to purchase it
(B) its rationale is similar to that in economics – same amount of money is worth more today than tomorrow
(C) it makes infinite horizon MDPs well formed, since long-term rewards can no longer diverge
(D) its value is between 0 and 1, typically very close to 0.



2. Which of the following algorithms is likely to converge with a fewer number of iterations?

(A) Value Iteration
(B) Policy Iteration

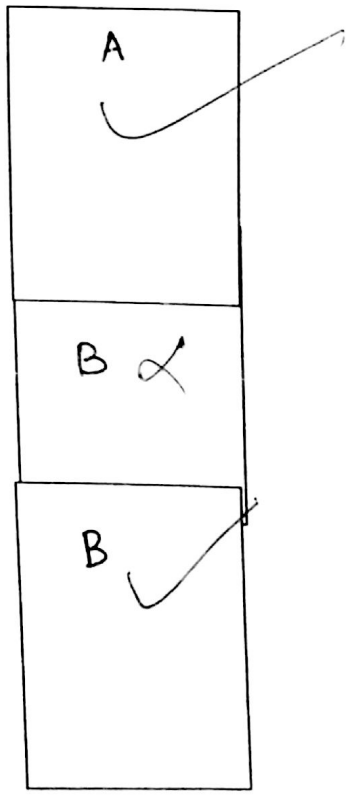


3. Which of these agents cannot be satisfactorily modeled as a POMDP

(A) route planning for a real robot operating in a physical environment
(B) a poker playing agent
(C) radiation planning agent for cancer patients
(D) none of the above

4. The optimal value function for a finite horizon POMDP is piecewise linear and convex even when the discount factor is less than 1.

(A) True
(B) False



5. LAO* algorithm outputs

(A) a path from the start state to the goal
(B) an acyclic AND/OR graph rooted at the start state
(C) a complete policy
(D) none of the above

6. Uniform sampling in a single state MDP has sublinear regret

(A) True
(B) False

7. Suppose we run LAO* algorithm on an undiscounted indefinite horizon MDP that has a proper (partial) policy rooted at the start state. Then, LAO* will

(A) never visit any state that is not reachable from the start state
(B) never visit any state that can't reach the goal (i.e., $V^*(s) = \infty$)
(C) visit all states reachable from the start state
(D) visit all states that can reach the goal (i.e., $V^*(s) < \infty$)

AB

8. A POMDP is just an MDP in a countably infinite state space

(A) True
(B) False

A

9. The Equal likelihood criterion and the criterion of realism 0.5 will result in the same best action.

(A) True
(B) False

A

10. If the utility of money m for an agent is e^m, then that agent is a

(A) risk-averse agent
(B) risk-neutral agent
(C) risk-prone agent

A

11. Suppose an agent ran to convergence Q-learning with a UCB exploration function in an MDP with known reward, but unknown transition model. Later we changed the reward function of the MDP and informed that to the agent. The agent can now obtain an optimal policy to the new MDP by

(A) running value iteration
(B) starting Q-learning initialized by the existing value function
(C) starting Q-learning initialized by a zero value function
(D) resuming the previous Q-learning computation on the new MDP

C

12. What is/are the reason(s) it is impractical to perform Q-learning for a game like Pacman?

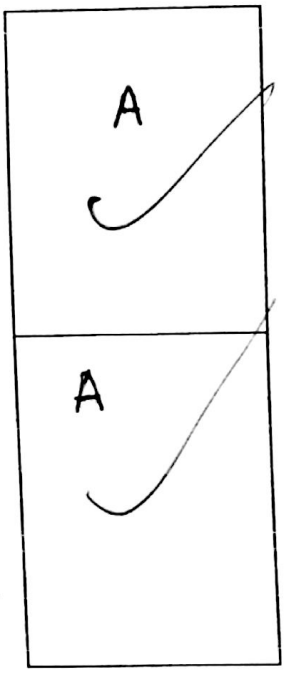
(A) the number of states is too high
(B) the branching factor is too high
(C) the value function is unbounded

A

(D) eating the power pellet reverses the role of the ghosts

13. The total number of possible policies in an MDP is

(A) $|A|^{|S|}$
(B) $|S|^{|A|}$
(C) $|A|^{|A|}$
(D) $|S|^{|S|}$



14. The iLAO* (improved LAO*) algorithm improves upon LAO* because

(A) iLAO* has a better running time than LAO*
(B) iLAO* typically uses less space compared to LAO*
(C) iLAO* converges to optimal solution where LAO* doesn't

15. [9 points] Match the problem settings with the algorithm that is best suited for that setting.

Known transition and cost model; known start state	F
Strong simulator; limited time available per decision	E
Known transition and cost model; evaluate value of a policy	H
Known transition and cost model; unknown start state	A
Weak simulator; estimate value of a known policy	B
Weak simulator; off policy learning	C

7.5

A	Prioritized sweeping
B	TD learning
C	Q learning
D	UCT
E	Policy improvement via rollout
F	iLAO*
G	A*
H	Solving system of linear equations

16. [3 points] You are given an MDP planning task (known transition and costs), where you want to compute the optimal policy. One question you are pondering is whether to compute and store V*(s) or Q*(s, a). Name one argument in favor of V*, and one different argument in favor of Q*.

(3) $V^*(s)$: Storing $V^*(s)$ takes less space

$Q^*(s,a)$: This will make getting the policy easy. As we know in action to take give a state $\pi(s)$ = argmax $Q^*(s,a)$

17. [10 points] TD-Learning

(a) [2 points] In trying to evaluate value of a policy π using TD-learning, suppose from state (x,y), the agent moves to state (x',y') and obtains an immediate reward r. What will be the update of $\hat{V}^\pi(x,y)$ assuming discount factor γ and learning rate α?

(1) Ans $V^\pi(x,y) \leftarrow V^\pi(x,y) + \alpha\left(r + \gamma V^\pi(x',y') - \alpha V^\pi(x,y)\right)$

(b) [3 points] Suppose we wish to update the value after taking two actions instead of one. That is, in state (x,y), the agent takes action $\pi(x,y)$ and reaches (x',y'), with reward r. Then it takes action $\pi(x',y')$, and reaches (x'',y'') with reward r'. Suppose we now wish to update $\hat{V}^\pi(x,y)$. What will be the new update equation?

(1)

$$V^\pi(x,y) \leftarrow V^\pi(x,y) + \alpha\left(r + \gamma\left(r' + \gamma V^\pi(x'',y'')\right) - \alpha V^\pi(x,y)\right)$$

(c) [5 points] Now, suppose you wish to use function approximation. We define the approximate value of a state as: , where x_g, y_g are both constants. Write down all the TD-learning update equations to learn the parameters when state (x,y) transitions to state (x',y'), and with an immediate reward r.

(0) as: $\hat{V}^\pi(x,y) = \theta_0 + \theta_1 x + \theta_2 y + \theta_3 (\log((x-x_g)^3 + (y-y_g)^3))$

$V^\pi(x,y) \leftarrow V^\pi(x,y) + \alpha(\text{Sample} - \hat{V}^\pi(x,y))$

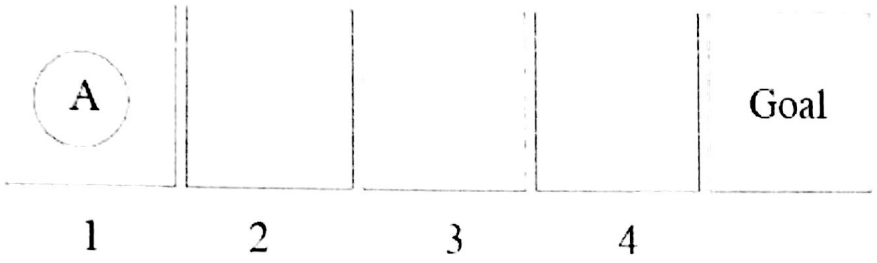
$\text{Sample} = r + (\theta_0 + \theta_1 x' + \theta_2 y' + \theta_3 (\log((x'-x_g)^3 + (y'-y_g)^3)))$

$V^\pi(x,y) \leftarrow V^\pi(x,y) + \alpha\left(r + (\theta_0 + \theta_1 x' + \theta_2 y' + \theta_3 (\log((x'-x_g)^3 + (y'-y_g)^3)))\right)$

18. [15 points] We are in an oversubscription planning setting. The rover on Mars is given many scientific goals $g_1, \ldots, g_k$, more than it can achieve in one charge. Each goal g_i on achievement provides a scientific reward of r_i. A reward on one scientific goal can only be obtained once. The rover has a total available battery B that it can spend before starting its "return to sunshine and recharge battery" routine. The rover is in a world state, which may include its location, its effector positions, etc., but does not include the available battery. It can execute actions, and each action transitions the world state according to and eats up units of battery. Rover's objective is to obtain rewards for as many goals as possible.

Cast this problem as an (un)-discounted infinite horizon reward maximization MDP M. Formally describe the state space, action space, transition function, and the reward function. If helpful, you can additionally define an applicability function denoting the set of actions that are allowed to execute in a given state. Do we need a discount factor less than 1 for M? Why or why not?

19. [35 points] 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 $\gamma = 1$.

$T(k, S, G) = k/6$,	$T(k, S, M) = 1 - (k/6)$,	for all $k = 1, 2, 3, 4$
$T(k, D, k+1) = y$,	$T(k, D, M) = 1 - y$,	for all $k = 1, 2, 3$
$R(k, S, G) = 2$, for all $k = 1, 2, 3, 4$.		All other transitions lead to zero rewards.

(a) [2 points] What is $V^{\pi}(1)$ for the policy that always shoots? Show your work.

$$V^{\pi}(s) = \sum_{s'} T(s,\pi(s),s') \left[R(s,\pi(s),s') + \gamma V^{\pi}(s')\right]$$

$$V^{\pi}(1) = T(1,S,G)\left[R(1,S,G)\right] + T(1,S,M)\left[R(1,S,M)\right]$$

$$= \frac{1}{6}[2] + \left(1 - \frac{1}{6}\right) \times 0 = 1/3$$

2

(b) [3 points] What is $Q^*(3, D)$ in terms of y? Show your work.

$$Q^*(s,a) = \sum_{s'} T(s,a,s') \left(R(s,a,s') + \gamma \max_{a'} Q^*(s',a')\right)$$

$$Q^*(3,D) = T(3,D,4)\left(0 + \max_{a'} Q^*(4,a')\right) + T(3,D,M)(0+0)$$

$$= y\left(\max\left(Q^*(4,D), Q^*(4,S)\right)\right)$$

4,D not allowed so

$$= y\left(Q^*(4,S)\right) = y\left(\frac{4}{6}(2)\right) = y\,4/3$$

3

(c) [8 points] Using y = 4/5, complete the first two iterations of Value Iteration. Assume that the states are updated in the order 1, 2, 3, and 4.

i	Vi(1)	Vi(2)	Vi(3)	Vi(4)
0	0	0	0	0
1	1/3	2/3	1	4/3
2	8/15	4/5	10/15	4/3

(d) [3 points] After how many iterations will Value Iteration compute the optimal values for all states when y = 4/5? Why?

$\left(\frac{4}{5}\right)^r V^{*}$

(e) [3 points] For what range of values of y is Q*(3, S) > Q*(3, D)? Show your work.

$Q^*(3,S) = T(3,S,G)(2) + T(3,S,M)(0)$

$= \frac{3}{6} \times 2 = 1$

$Q^*(3,D) = T(3,D,4)(0 + Q^*(4,S)) + T(3,D,M)(0)$

$= y\left(\frac{4}{6} \times 2\right)$

$= \frac{4}{3} y$

$1 > \frac{4}{3} y \rightarrow y < \frac{3}{4}$

For the rest of the questions we give the soccer agent A some additional information. We tell it that there could be an opponent defending robot O. A has no way of knowing whether O is present or absent, but does know the statistical properties of its environment. When O is absent y = 4/5. When O is present y=1/5.

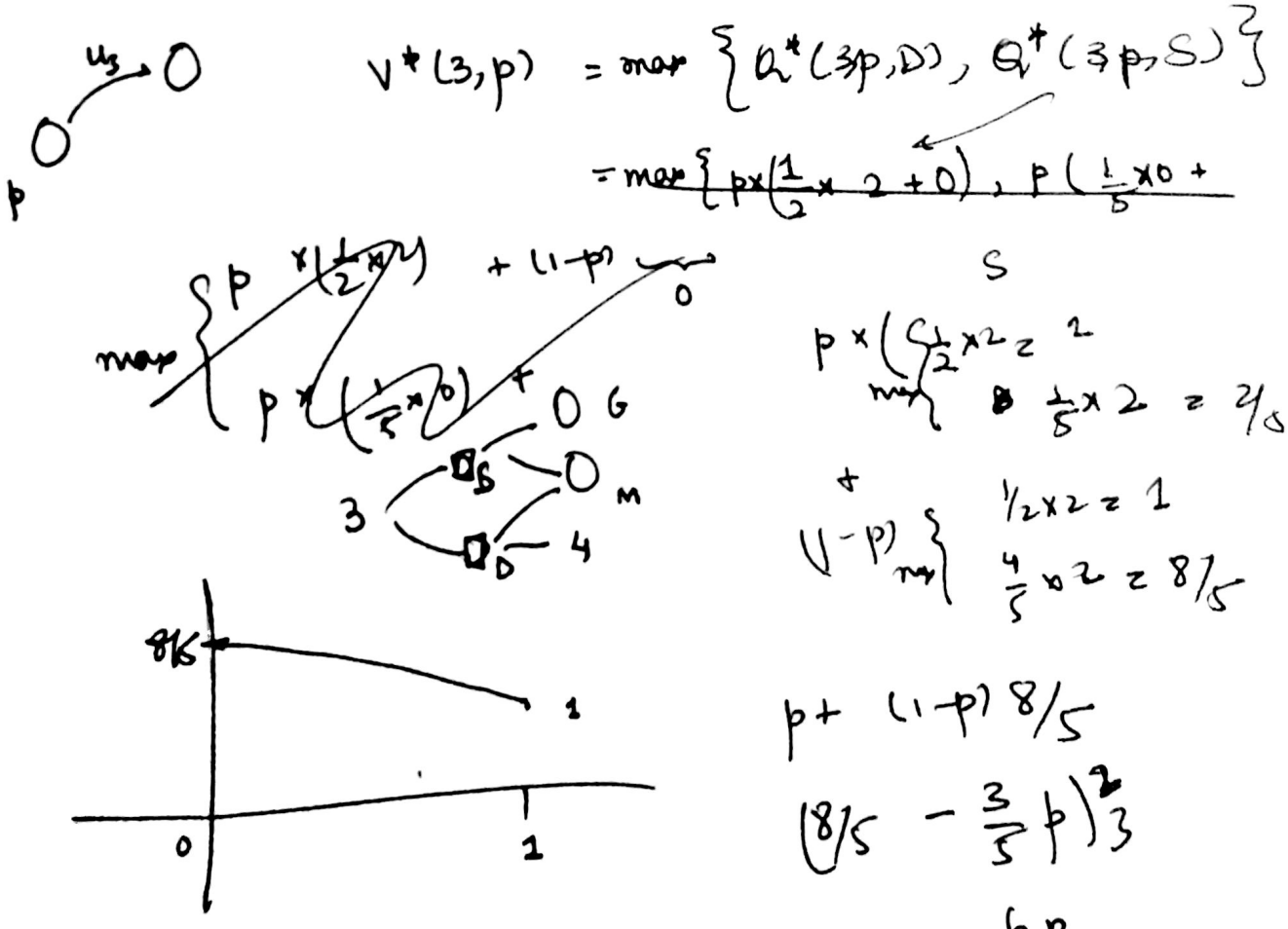
(f) [8 points] Since O's presence is unknown, A will model the decision making problem as a POMDP. Suppose the current belief is (s,p), i.e., A is in state s, and O is present with probability p. Compute the expected value V*(3, p) in terms of p. What is the optimal policy for belief (3,p)? Show your work.

$V^*(3,p) = \max\left[\frac{1}{5}(V^*(4)) + \frac{5}{6} \times 2\right]$

$= \max\left[\frac{1}{5} \times \frac{4}{6} \times 2 + 1\right]$

$= \frac{4}{15} + 1$

$\pi(3) = G$ is the optimal policy



(g) [8 points] Imagine that A can also use its laser range finder and take a sensing action L, which perfectly reports whether O is present or not. What is the value of information provided by the action L in belief (3, 0.5)? Show your work.

$V^*(3) = \max\left[0.5 \times \left[\frac{1}{5} \times V(4)\right] + \cancel{0.5\left[\frac{4}{5} \times V^*(4)\right]} + \frac{3}{6} \times 2\right]$

$= \max\left[\frac{1}{2}\left(\frac{1}{5} \times \frac{4}{6} \times 2 + \frac{4}{5} \times 2 \times \frac{4}{6}\right) + 1\right]$

$= \max\left[\left(\frac{4}{6}\right), 1\right]$

$= 1.$

So the value of information = prevd-

$V^*(3) = \cancel{0.5} \times (1)\frac{1}{2} + \frac{1}{2}\left(\frac{4}{5} \times \frac{2}{3} \times 2\right)$

$= \frac{1}{2}\left[1 + 16/15\right)$

$= \frac{31}{30}$ ✓

~~with~~ while without ~~reward~~ Information

$V^*(3) = 1$

$\therefore$ VOI $= \frac{1}{30}$.

8