AI Planning for Autonomy (COMP90054_2020_SM1)

Started: Jun 29 at 9:00

Quiz Instructions

Academic Integrity Declaration

By commencing and/or submitting this assessment I agree that I have read and understood the University's policy on academic integrity. ↗

I also agree that:

1. Unless paragraph 2 applies, the work I submit will be original and solely my own work (cheating);
2. I will not seek or receive any assistance from any other person (collusion) except where the work is for a designated collaborative task, in which case the individual contributions will be indicated; and,
3. I will not use any sources without proper acknowledgment or referencing (plagiarism).
4. Where the work I submit is a computer program or code, I will ensure that:
   a. any code I have copied is clearly noted by identifying the source of that code at the start of the program or in a header file or, that comments inline identify the start and end of the copied code; and
   b. any modifications to code sourced from elsewhere will be commented upon to show the nature of the modification.

This exam begins at 9:00 AM Australian Eastern Standard Time (AEST) on Monday 29/06/2020 in Canvas (lms.unimelb.edu.au). The exam must be completed by 11:00 AM AEST on Monday 29/06/2020. This exam has 120 minutes of writing time.

Questions

↻ Refresh
- ✓ Question 1
- ✓ Question 2
- ✓ Question 3
- ✓ Question 4
- ✓ Question 5
- ✓ Question 6
- ✓ Question 7
- ✓ Question 8

Time Running: Hide
Attempt due: Jun 29 at 11:00
0 Minutes, 10 Seconds

Multiple-choice questions [10 marks]

Question 1 — 1 pts

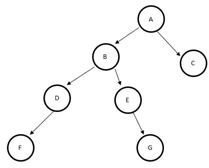
Consider a policy π that takes a state and returns the action a that should be chosen in state s.

What type of policy is this?

- ○ A random policy
- ○ A local policy
- ○ An initial policy
- ○ A stochastic policy
- ◉ A deterministic policy

Question 2 — 1 pts

Consider the search tree shown in the figure below. Assume the Goal state is G and that ties are broken alphabetically (e.g. B before C). Using Iterative Deepening Search on the tree above, how many times will node D be visited before a solution is found?



- ○ 1
- ◉ 2
- ○ 5
- ○ 3
- ○ 4

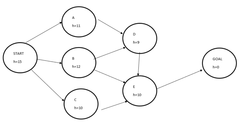
Question 3 — 1 pts

What is the correct formula for policy extraction from a value function?

- ◉ argmax_a(s,a) Sum_{s_in S} P_a(s'|s) [r(s, a, s') + γ V(s')]
- ○ π(s) = Q(s,a)
- ○ argmax_a(s,a) Sum_{s_in S} P_a(s'|s) [r(s, a, s') + γ Q(s',a)]
- ○ π(s,a) = argmax_{a' in A(s)} Q(s,a')
- ○ argmax_a(s,a) Sum_{s_in S} P_a(s' | s) [r(s, a, s') + γ max_{a' in A(s')} Q(s',a')]
- ○ π(s,a) = Q(s,a)
- ○ argmax_a(s,a) Sum_{s_in S} P_a(s'|s) [r(s, a, s') + γ max_{a' in A} γ Q(s',a')]

Question 4 — 1 pts

Consider the search tree shown in the diagram below. Using the Weighted A* Algorithm with $W=2$ and assuming a uniform cost of 3 to move between nodes, which will be the fourth node expanded.



- ○ B
- ◉ E
- ○ D
- ○ A
- ○ C

Question 5 — 1 pts

Select the statement below that is false

- ○ Q-learning and SARSA are both temporal difference learning
- ◉ Q-learning learns a Q-function and SARSA learns a policy
- ○ SARSA is temporal different learning
- ○ Q-learning is temporal different learning
- ○ n-step reinforcement learning can be applied to both Q-learning and SARSA

Question 6 — 1 pts

Below is the Bellman-Ford Table for $h^{MD}(I)$ for a particular problem where I is the initial state of the problem.

completed(A)	completed(B)	completeC)	completed(D)	completed(E)
Infinity	Infinity	Infinity	Infinity	Infinity
5	8	Infinity	7	1
5	3	2	4	1
5	3	2	4	1

If the goal is {completed(A), completed(B)}, what is the value of $h^{MD}(I)$?

- ◉ 8
- ○ Infinity
- ○ 5
- ○ 3
- ○ 15
- ○ 1

Question 7 — 1 pts

Below is the Bellman-Ford Table of $h^{max}(I)$ for a particular problem where I is the initial state of the problem.

completed(A)	completed(B)	completeC)	completed(D)	completed(E)
Infinity	Infinity	Infinity	Infinity	Infinity
5	8	Infinity	7	1
5	3	2	4	1
5	3	2	4	1

If the goal is {completed(A), completed(B)}, what is the value of $h^{max}(I)$?

- ○ Infinity
- ○ 3
- ○ 7
- ◉ 5
- ○ 8
- ○ 1

Question 8 — 1 pts

Select the **false** statement below about reward shaping.

- ○ Potential-based reward shaping is guaranteed to converge to an optimal policy for any representation in which it would converge without shaped rewards
- ○ Rewarding shaping does not modify the reward function
- ○ Reward shaping can be used with Q-learning and with SARSA
- ◉ Potential functions in reward shaping are heuristic functions
- ○ Reward shaping can be used to derive a better policy than without reward shaping

Question 9 — 1 pts

Consider the following game. Select the pure strategy Nash equilibrium for this game.

	Player 2	
Player 1	**Left**	**Right**
Up	4, 6	4, 5
Down	3, 7	5, 6

- ○ 3,7
- ◉ 4,6
- ○ 4,5
- ○ 5,6

Question 10 — 1 pts

Which of the following statements is false?

- ○ All consistent, goal aware heuristics are admissible
- ○ The hadd heuristic is inadmissible in general
- ○ Depth first search is complete for acyclic state spaces
- ◉ The IDA* algorithm is bounded suboptimal for admissible heuristics
- ○ The hmax heuristic is always admissible

Short-answer questions [20 marks]

Question 11 — 2 pts

Consider a reinforcement learning agent that is trying to learn how fast a vacuum cleaning robot can travel without over-heating.

There are two states: *cool* and *fast*.

There are two actions: *slow* and *fast*.

If the robot goes fast, it is more likely to transition to a warm state than it is goes slow.

Using a learning rate of 0.5 and a discount factor of 0.8, we arrive at the following Q table:

Q(cool, fast)	9
Q(cool, slow)	9
Q(warm, fast)	4
Q(warm, slow)	6

The agent executes the action *fast* in the state *cool*, receives a reward of 6.0, and is now in the

warm state. It will execute the action slow next.

Calculate the new value for $Q(cool, fast)$ using 1-step Q-learning to 2 decimal places.

9.9

Question 12

2 pts

Consider two players in a role-playing game, who have different skills available to them. On some turns in the game, the players need to compete for a resource. They can either choose to use their *Negotiate* strategy or use their *Power* strategy. The payoffs resulting from the *Power* strategy depend on the particular power of both of the players (those payoffs are represented by *a* and *b* below). The game is described below, and are described below.

	Player 2	
Player 1	**Power**	**Negotiate**
Power	a, b	10, 7
Negotiate	10, 13	2, 6

Consider situation in with both players have the same *Power* strategy as each other, so *a*=*b*. What is the smallest value of *a* that gives both players a strongly dominant strategy?

11

Question 13

2 pts

Autonomous weapons are weapons that can search for and fire at given targets without human intervention.

AI techniques for sequential decision making, such as planning, reinforcement learning, and game theory, can be used to improve the accuracy of autonomous weapons, potentially saving lives of civilians.

Given one short argument (2-3 sentences) that justifies the use of autonomous weapons from a utilitarianism or deontological perspective.

Specify which ethical perspective you take.

Note: you do not need to agree with the argument that you give.

12pt Paragraph

Argument: There is never complete certainty about whether a "target" is actually hostile so the autonomous weapon could potentially kill a civilian. However, from a utilitarianism perspective, using the autonomous weapon would save more civilian lives than it would remove so it is worth the risk.

Ethical perspective: The weapon should not be used due to the potential to harm civilians although it would save some. It's rather like the trolley problem.

p 72 words

Question 14

2 pts

Match the techniques below with their properties. Multiple techniques can match to one property.

Policy iteration: Model-based

Monte-Carlo Tree Search: Model-free, model-based, or s

Q-learning: Model-free

The following description applies to the two questions in this exam.

Consider a robot called *MedAssist*, which takes medical kits from their storage location to an operating theatre in a hospital. When it is not being used, it stays at its base station to charge.

It can be in one of five states:

1. *Base*: It is at its base station
2. *No Kit*: It is not at its base station and does not have a medical kit
3. *Kit 1*: It has collected medical kit 1
4. *Kit 2*: It has collected medical kit 2
5. *Delivered*: It has delivered a medical kit 1 or 2

There are three actions available:

1. *get_kit1*: *MedAssist* goes to collect medical kit 1. There is a 0.8 chance kit 1 will be there (transition to state *Kit 1*) and 0.2 chance kit 1 will be missing (stay in state *No Kit*). No immediate reward is received
2. *get_kit2*: *MedAssist* goes to collect medical kit 2. There is a 1.0 chance kit 2 will be there (transition to state *Kit 2*). No immediate reward is received.
3. *deliver*: Deliver the kit that is being carried

The MDP transition probabilities and rewards are:

s	a	s'	P(s, a, s')	r(s, a, s')
Base	get_kit1	No Kit	0.8	0
Base	get_kit1	Kit 1	0.2	0
Base	get_kit2	Kit 2	1.0	0
No Kit	get_kit2	Kit 2	1.0	0
Kit 1	deliver	Delivered	1.0	10
Kit 2	deliver	Delivered	1.0	5

Question 15

2 pts

Using value iteration, we end up with the following value function for *MedAssist* after four iterations using a discount factor $\gamma = 0.9$.

State	Base	No Kit	Kit 1	Kit 2	Delivered
Value	2.3	4.4	9	6	0

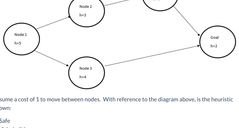
Apply one more iteration to calculate the value $V(Base)$ after five iterations.

Enter your final answer to two decimal places in the box below

6.74

Question 16

5 pts



Assume a cost of 1 to move between nodes. With reference to the diagram above, is the heuristic shown:

(i) Safe
(ii) Admissible
(iii) Consistent
(iv) Goal aware
(v) Equivalent to h^*

For each point, why or why not?

12pt Paragraph

(i) Yes, safe: all nodes shown have a path to the goal and none of them have h = infinity.

(ii) Not admissible: it overestimates. Take node 3 with $h^* = 1$ but h = 4.

(iii) Not consistent: h(n3) = 4 and h(goal) + c(n3, goal) = 2 + 1 = 3 so h(n3) > h(goal) + c(n3, goal)

(iv) Not goal aware: goal has h = 2 which is not h = 0.

(v) Not equivalent to h^*: take node 4 which has optimal cost 1 from the goal but h = 4.

p 101 words

Question 17

5 pts

With reference to the diagram above, answer the following questions:

(i) State the nodes that will be pruned by the IW(1) algorithm.
(ii) State the nodes that will be pruned by the IW(2) algorithm.
(iii) Will the goal state (Node 11) be found using either algorithm?

12pt Paragraph

(i) 6, 10

(ii) None

(iii) Yes, found by IW(1) and hence would be found by IW(2)

p 19 words

Long-answer questions [20 marks]

Question 18

5 pts

A robot (R) can move horizontally and vertically to adjacent cells within a walled room as depicted in the figure below. Note that the robot cannot move diagonally between cells and cannot move into a wall (W) or into the box. The robot can swap its position with that of a box (B) in order to manipulate the position of the box. For this to happen, the robot has to be at a position p_1 adjacent to p_2. The goal is to get the Box to the target (T).

W	W	W	W	W	W
W	B		W	W	W
W	R			W	W
W					W
W	W		T	W	W
W	W	W	W	W	W

The robot has two actions *move* and *swap*. Describe briefly in STRIPS how to model the domain described. Include a specification of the parameters the two actions will take, and the preconditions and postconditions of each action. Also include a description of the initial and goal states corresponding to the diagram above.

You are allowed to use variables as arguments for the actions (action schemes), specifying the values of the variables. Note: it is not compulsory to use PDDL syntax, as long as you can convey the main ideas.

12pt Paragraph

Atoms

F = {robot(X, Y), box(X, Y), wall(X, Y) | X, Y ∈ {0, 1, 2, 3, 4, 5}}

Operators

O = {move(X, Y, X', Y'), swap(X, Y, X', Y')}

PreimageO(X, Y, X', Y') = {robot(X, Y), ¬wall(X', Y'), ¬box(X', Y'), ((X' - X) = 1 ∧ Y' = Y) ∨ ((Y' - Y) = 1 ∧ X' = X) | X, Y, X', Y' ∈ {0, 1, 2, 3, 4, 5}}
AddS(move(X, Y, X', Y')) = {robot(X', Y')}
DelS(move(X, Y, X', Y')) = {robot(X, Y)}

PreimageO(X, Y, X', Y') = {robot(X, Y), ¬wall(X', Y'), box(X', Y'), ((X' - X) = 1 ∧ Y' = Y) ∨ ((Y' - Y) = 1 ∧ X' = X) | X, Y, X', Y' ∈ {0, 1, 2, 3, 4, 5}}
AddS(swap(X, Y, X', Y')) = {robot(X', Y'), box(X, Y)}
DelS(swap(X, Y, X', Y')) = {robot(X, Y), box(X', Y')}

Initial situation

I = {wall(0, 0), wall(0, 1), ... (all walls), box(4, 1), robot(3, 1)}

Goal

G = {box(1, 3)}

p 188 words

Question 19

5 pts

Consider a STRIPS problem with the following two actions:

$r(X, Y)$:
Preconditions: $near(X, Y)$, $on(X)$
Postconditions: $on(Y)$

$q(X, Y)$:
Preconditions $near(X, Y)$, $on(X)$, $on(Y)$
Postconditions $done(Y)$; $\neg on(Y)$ (i.e. the action deletes on (Y))

Suppose X and Y can each take the value $\{1, 2\}$ and that the cost of executing each action is 1.

The initial conditions are the following: $I := \{on(1), near(1, 2), near(2, 1)\}$

The goal is to achieve the following: $G := \{done(1), done(2)\}$

Find the value of $h^{max}(1)$. Find the value of $h^{f}(1)$ using the best supporter function induced by h^{max}. Show all working.

12pt Paragraph

hmax table:

iter	oe(1)	oe(2)	done(1)	done(2)	near(1, 1)	near(1, 2)	near(2, 1)	near(2, 2)
0	0	inf	inf	inf	inf	0	0	inf
1	0	1	inf	inf	inf	0	0	inf
2	0	1	2	2	inf	0	0	inf
3	0	1	2	2	inf	0	0	inf

Now computing hff:

Open := [done(1), done(2)]

bsidone(1)) = q(2, 1) => Open := [oe(2), done(2)], RPlan := [q(2,1)]

bsidone(2)) = r(1, 2) => Open := [done(2)], RPlan := [q(2, 1), r(1, 2)]

bsidone(2)) = q(1, 2) => Open: [], RPlan := [q(1, 2), q(2, 1), r(1, 2)]

Hence, hff(1) = sum action costs in RPlan.

hff(1) = 3

p + strong 139 words

Question 20

1 pts

MedAssist (see the description earlier) is updated, and it no longer knows the probability transition matrix, so it does not know the probability of Kit 1 being present. It also has an additional action called return, which returns to the base station, receiving no reward.

Assume the following Q-function for MedAssist, implemented as a Q-table. Note that some Q-values are omitted as they are not important for this question:

Q	Value
Q(Base, get_kit1)	0
Q(Base, get_kit2)	5
Q(NoKit, get_kit1)	9
Q(NoKit, get_kit2)	0
Q(Kit 1,deliver)	4
Q(Kit 2,deliver)	4
Q(Kit 1,return)	0
Q(Kit 2,return)	0

Consider the following trace from the state Base: get_kit1 → No Kit → get_kit2 → Kit 2 → deliver → Delivered

Assuming that α = 0.5 and γ = 0.8, perform a 3-step update for Q(Base, get_kit1) using Q-learning. Show your working.

Enter your final answer to two decimal places in the box below, and your working in the next question box.

1.02

Question 21

4 pts

Enter your working for the above question on reinforcement learning:

Note: If you get the incorrect answer and show no working, then you have not demonstrated any understanding, so will receive no marks. Thus, it is important that you show working in case of minor calculation errors, etc.

12pt Paragraph

Rewards:

Kit 1,deliver|Delivered|10|10
Kit 2,deliver|Delivered|10|5

States: Base -> No kit -> Kit 2 -> Delivered

G = 0 + 0.8(0) + 0.8^2(0) + 0.8^3 Q(Kit 2,deliver)
= 0.8^3 (4)
= 2.048

Q(Base, get_kit1) = Q(Base, get_kit1) + 0.5(G - Q(Base, get_kit1))
= 0 + 0.5(G - 0)
= 1.024

p + span 70 words

Question 22

5 pts

Consider an AI agent for trading financial securities. Each day, the agent needs to make the decision whether to buy or sell some of its holdings. The payoff depends on whether the market wants to buy or sell on those days.

The agent has the following information:

- If I decide to sell and the other traders buy, my utility will increase by $5 million dollars. The market's happiness will increase by $2 million.
- If I decide to buy and the other traders try to buy too, my utility will be $0. The other market's happiness will be $4 million though, because they keep their money to spend on other securities.
- If I decide to sell and the other traders try to sell too, my utility will decrease by $1 million because I wasted my time and sold nothing. The market's happiness will be $3M, because they still got to keep their money.
- If I decide to buy and the other traders sell, my utility will increase by $1 million dollars. The market's happiness will be $2 million.

What should the agent's strategy be today? Show your working.

12pt Paragraph

		Market	
		Buy	Sell
Agent	Buy	0, 4	1, 2
	Sell	5, 2	-1, 3

Agent should buy. If agent sells and market sells, payoff for agent is -1. Payoff for buy will be 0 or 1.

p 36 words

Saved at 11:00 Submit Quiz