NATIONAL UNIVERSITY OF SINGAPORE

SCHOOL OF COMPUTING

SEMESTER 1 (2021/2022)
Mid-Term Test Solution Sketches

CS4246/CS5446: AI Planning and Decision Making

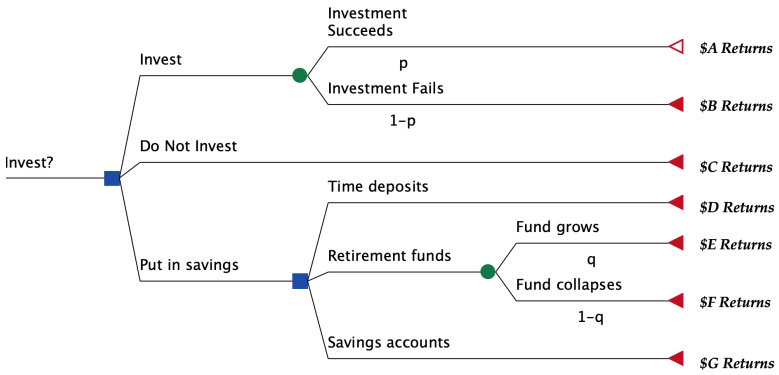
October 2021 Time Allowed: 120 minutes

Question 1 (20 marks)

For each of the following cases, answer **True or False.**

1. Both value iteration and policy iteration can be used to solve a Markov decision process with transition functions that change over time, with the convergence and optimality guarantees as discussed in class. **(2 marks)**

2. The information captured in the following decision tree cannot be represented as an influence diagram. **(2 marks)**



3. Unlike active adaptive dynamic programming (ADP) in reinforcement learning, model-free Q-learning does not focus on exploring the states that would lead to larger differences in the utilities learned in recent iterations. **(2 marks)**

4. If a sequence of primitive actions does not appear in any of the refinement methods of a hierarchical task network (HTN) planner, it will never appear as a solution for any planning problem solved by the planner. **(2 marks)**

5. In a world that can be specified by X state variables, where each state variable can take up to Y values, there can be up to X^Y belief states. **(2 marks)**

6. It is impossible to model a stochastic process whose current state (as denoted by the outcome of the state variable) S_t at time t is dependent on its immediate two previous states: S_{t-1} at time t – 1, S_{t-2} at time t – 2, as a first-order Markov process, i.e., a stochastic process whose current state S'_t at time t is dependent only on its previous state: S'_{t-1} at time t – 1. **(2 marks)**

7. Consider the following two alternatives for Agent J: **(2 marks)**

   Case A: In addition to what he currently owns, he has been given $500. He is now asked to choose one of these options:
   50% chance to win $500 (and 50% chance to win $0) OR get $250 for sure

   Case B: In addition to what he currently owns, he has been given $1000. He is now asked to choose one of these options:
   50% chance to lose $500 (and 50% chance to lose $0) OR lose $250 for sure.

   Utility theory will predict that Agent J will be indifferent between the two choices in Case A and indifferent between the two choices in Case B.

8. Consider the following two alternatives for Agent J: **(2 marks)**

   Case A: In addition to what he currently owns, he has been given $500. He is now asked to choose one of these options:
   50% chance to win $500 (and 50% chance to win $0) OR get $250 for sure

   Case B: In addition to what he currently owns, he has been given $1000. He is now asked to choose one of these options:
   50% chance to lose $500 (and 50% chance to lose $0) OR lose $250 for sure.

   Prospect theory will predict that Agent J will take the uncertain alternative : 50% chance to win $500 in Case A and 50% chance to lose $500 in Case B.

9. Algorithmic solutions to automatically choose the optimal trade-off points between fairness and accuracy in AI systems are available; these solutions can be applied in general for different group and individual fairness criteria and settings. **(2 marks)**

10. In the fight against the Covid-19 pandemic, clinical trials and intervention (test and treatment) planning may be modeled as Markov Decision Processes with large, varying state spaces and transition functions that change with time; these problems can be effectively solved by Monte Carlo Tree Search (MCTS). **(2 marks)**

[Solution]:

1. **False. Value iteration can work with heterogeneous transition models in finite horizon problems; extension to the basic formula with time indices is needed. Policy iteration assumes stationary (non-changing) policies, non-stationary policies do not have convergence guarantees.**

2. **False. Influence diagrams and decision trees are isomorphic. The same information can be captured in the influence diagram although not explicitly through the structure.**

3. **True. Model-free Q-learning cannot take advantage of the memory-based "prioritized" states concept in modeling learning and planning (policy generation or improvement) in reinforcement learning. In Q-learning, updating are based only on the experiences (trials); exploration can be random, calculated as a function of the experienced trials, or captured in a function approximation of the Q-functions.**

4. **True. This is because all the "legal" or desirable sequences would have been specified in the refinement methods or high level actions.**

5. **False. There will be Y^X belief states.**

6. **False. A second-order Markov process can be formulated as a first-order one with a 2-tuple state definition: $S_{t'} = (S_t, S_{t-1})$.**

7. **True. Utility theory will predict Agent J will make the same choice in each case, as they are the same outcomes, but not what they are.**
8. **False. Prospect theory will predict in the first case that Agent J will take the $250, because people are risk averse for gains. It will predict Agent J will gamble in the second case because people are risk seeking for losses.**

9. **(D) False. Human judgment is required to define and determine the trade-off formulations and thresholds.**

10. **(D) False. MCTS is effectively only if the problem environment is simple enough for fast multistep simulation.**

Question 2 (24+1 = 25 marks) Markov Decision Process

Happy is a new Rover powered by AI planning and decision making deployed on Mars. There are three buttons: 1, 2, and 3 on Happy that can be activated, one at a time, to automatically send Happy to each of the three bases on Mars: Apple (A), Banana (B), and Cherry (C) to gather rock samples. Happy receives 1 unit of power from base C, and none from any other bases. As the technology is new, the effects of activating the buttons can sometimes be uncertain. In particular, the following patterns are observed:

- At base A:

Button 1: 0.2 probability of no-effect, and 0.8 probability of getting to base B
Button 2: 0.2 probability of no-effect, and 0.8 probability of getting to base C

- At base B:

Button 2: 0.2 probability of no-effect, and 0.8 probability of getting to base C
Button 3: 0.2 probability of no-effect, and 0.8 probability of getting to base A

- At base C:

Button 1: 0.5 probability of getting back to base C, and 0.5 probability of getting to base A
Button 3: Always getting to base B

Happy needs to decide on the right button to activate at each base, so that it can maximize the power units it receives to work as long as possible on Mars. Assume that it will always get the power unit as long as it gets to base C, regardless of its previous base.

Happy's decision problem can be model as a Markov Decision Process (MDP) with:

- State-space $\boldsymbol{S} = \{A, B, C\}$
- Action-space $\boldsymbol{A} = \{1, 2, 3\}$
- Reward model: $R(A) = 0, R(B) = 0, R(C) = 1$

A. Does the following correctly depicts Happy's transition model T(s, a, s') for states s, s' and actions a? **[6 marks]**

- $T(A, 1, A) = 0.2, T(A, 1, B) = 0.8, T(A, 2, A) = 0.2, T(A, 2, C) = 0.8$
- $T(B, 2, B) = 0.2, T(B, 2, C) = 0.8, T(B, 3, B) = 0.2, T(B, 3, A) = 0.8$
- $T(C, 1, C) = 0.5, T(C, 1, A) = 0.5$
- $T(C, 3, B) = 1$

Ans: **A**

a. Yes
b. No

[Solution]: the transition functions are correct.

B. Assuming a discount factor gamma = 1, apply the value iteration algorithm for MDP and determine the utility values for the states for t = 1 and t = 2. (You do NOT need to show the detailed calculations.) What is the optimal policy for Happy at t = 2? **(6 marks)**

Ans: Assume "policy-at" is equivalent to the "pi" function notation:

- policy-at(A) = **[2]** , policy-at(B) = **[2]** , policy-at(C) = **[1]**

[Solution]: (up to 2 decimal points).

	Time or Iteration (t)					
	0	**1**	**2**	**3**	**4**	**5**
U(A)	**0**	**0**	**0.64**	**1.328**	**1.92**	**2.85**
U(B)	**0**	**0**	**0.64**	**1.328**	**1.92**	**2.85**
U(C)	**0**	**1**	**1.5**	**2.07**	**2.699**	**3.31**

Example, from time t = 1 to t = 5:

U(A) = R(A) + γmax (0.2U(A) + 0.8U(B) [1]
0.2U(A) + 0.8U(C)) [2]

U(B) = R(B) + γmax (0.2U(B) + 0.8U(C) [2]
0.2U(B) + 0.8U(A)) [3]

U(C) = R(C) + γmax (0.5U(C) + 0.5U(A) [1]
1U(B)) [3]

Optimal policy:

π*(A) = [2], π*(B) = [2], π*(C) = [1]

C. Happy uses its long ram to collect rock samples. There is a 0.8 probability that the rock will accidentally slip from its long arm every time it tries to pick up a rock. If this happens, Happy will try to pick the rock up again, until it succeeds. How many attempts does it take on average for Happy to successfully pick up a rock? **(6 marks)**

Ans: It will take **5** attempts on average.

[Solution]:

It will take on average 5 attempts. (Model the process as an MDP with two states: rock-on-floor and picked-up with cost (F) = 1 and cost (P) = 0)

D. Consider the case when a new rover, Jolly is deployed to join Happy on the same mission. Assume that both Happy and Jolly are made in exactly the same way, with the same button effects. Can the MDP as defined in (A) above be modified or adjusted to accommodate the new decision problem with two agents? **(6 marks)**

Ans: **A**

a. Yes. The state, action, transition, and reward definitions of the MDP can be amended to include information about the second rover.

b. No. The new problem can only be formulated in a multi-agent planning process.

[Solution]:

There are several ways of transforming an MDP into a multiagent MDP.

Simplest solution is to extend state description to accommodate second agent (position, time)

Some examples include placing the second agent's effects into the transition function. That is, assume the other agent is merely part of the environment. This assumes that the agents do not change their behavior since the transition function in an MDP must be fixed. Another way is to extend the definition of MDP to include multiple agents all of which can take an action at each time step. Also need to determine how the reward can be distributed among the agents.

Question 3 (20 marks) Classical Planning

Happy the Rover carries two containers: a 5-litre BIN and a 2-litre BAG to collect precious rock samples on Mars. After collecting and filling the 5-litre BIN with rock samples at the Jezero Crater (a site on Mars), it needs to transfer **precisely** 1-litre of the rock samples into the empty 2-litre BAG, before moving to the next site to collect more rock samples.

Happy can transfer rock samples from the BIN to the BAG, or vice versa, and it can discard the rock samples, but no additional rock samples are available other than from the originally filled 5-litre BIN.

You may assume that the rock samples can be transferred back and forth accurately to the BIN or BAG, i.e., disregard any remnants or losses in the transfers.

Assume that the states in the problem can be expressed as **S(b, g)**, where **b** is the amount of rock samples in the 5-litre (big) BIN, and **g** is the amount of rock samples in the 2-litre (small) BAG.

- **Briefly** answer all the following subparts.
- Clearly state the **assumptions** you make.
- Clearly **label** your answers to the subparts in the space provided below.
- You may use the **symbols**: &, ||, ~ to denote logical *"and", "or"*, and *"not-"* in your answers.

[Any reasonable assumptions are acceptable. Assumptions for the solution sketches as follows:

Assuming states in the form of S(b, g). Any other assumptions have to clearly state the transformation or mapping rules.

Assuming that there is no measuring cup, and that pouring amount can only be measured by the (full) content of either the origin (or "from") container (BIN or BAG) or the target (or "to") container (BIN or BAG). Any other assumptions have to clearly state the transformation or mapping rules.]

A. What is the state space of the problem? **(4 marks)**

- Define in terms of the range of values that **b** and **g** can take on, as well as the initial state, and the goal state.

[Solutions]: Credits for: Clear assumptions, correct representation of states, initial state, goal state, range for b, and range for g

States can be expressed as S(b, g), where $0 \leq b \leq 5, 0 \leq g \leq 2$ are integers subject to the constraint: $0 \leq b + g \leq 5$, as we only consider the initial 5-L of rocks in BIN.
Assume only discrete volume of rock samples in the containers, there are 15 states in the general state space: total possible combinations: |{0,1,2,3,4,5}| x |{0,1,2}| - |{(5,1),(5,2),(4,2)| 18 – 3 = 15, smaller if some problem specific constraints are imposed. Note that in general definition of the state space in the planning domain is different from the list of reachable states for a specific problem.
Start state is (5, 0), and goal state is (x, 1), where $0 \leq x \leq 5$

B. What are the actions needed to solve this problem? **(8 marks)**

- Briefly describe each of them and specify the actions in PDDL-like format. (You may make any assumptions about the action preconditions or effects, including using functions and other appropriate calculations.)

[Solutions]: Credits for: Feasible representation of actions based on clear assumptions; acceptable definitions of preconditions and effects.

One possible set of definitions:
EmptyBig(), EmptySmall() TransferFromBig(b, x), TransferFromSmall(g, x), where transfer is restricted from one container to another. Notice that the preconds and effects of the actions are defined in terms of the state (changes).

Assume that we can specify functions/calculations in the action effects, for all $0 \leq x \leq b, 0 \leq y \leq g$, and b, g, x, y are integers

EmptyBIN()
Precond: S(b, g)
Effect: S(0, g)

EmptyBAG()
Precond: S(b, g)
Effect: S(b, 0)

TransferFromBIN(b, x)
Precond: S(b, g)
Effect: S(b – x , min(g + x, 2))

TransferFromBAG(g, y)
Precond: S(b, g)
Effect: S(min(b + y, 5), g – y))

Alternate definitions more specific to the problem are acceptable.

C. Describe a plan that would solve the problem. **(8 marks)**

- List the actions and trace through the states when you execute the plan. (Note: You do NOT need to show the full search tree. It suffices to just show the final action sequence and the state transitions.)

[Solutions]: Credits for: Acceptable solution based on clear assumptions, correct applications of actions, and correct state-transition descriptions

Assuming that there is no measuring cup (as stated earlier), a possible solution is:

TransferFromBIN(5, 2), EmptyBAG(), TransferFromBIN(3, 2), EmptyBAG(), TransferfromBIN(1, 1)

State transitions: (5, 0) → (3, 2) → (3, 0) → (1, 2) → (1, 0) → (0, 1)

Question 4 (15 marks) Classical Planning

The Chief Engineer at Mars Mission Command is very impressed with the successful AI planning solutions deployed by Happy the Mars Rover. To further improve the performance of future rovers, she tries to combine some of the actions in the classical planners designed for Mars Rovers. You are to advise on the following:

Assume that the two actions to be combined are a_1 and a_2, the composite action $[a_1; a_2]$ means: "do a_1 then do a_2."

Assume also that the definitions follow the simpler PDDL-like representations, i.e., without any additional functional/computational assumptions made on the preconditions and effects.

- **Briefly** answer all the following subparts.
- Clearly state the **assumptions** you make.
- Clearly **label** your answers to the subparts in the space provided below.
- You may use the **symbols**: &, ||, ~ to denote logical *"and", "or"*, and *"not-"* in your answers.

A. What are the effects for this composite action, in terms of the original effects of a_1 and a_2? **(5 marks)**

[Solution]: Partial credits if only one condition is met or only partial descriptions are provided for each part. Descriptions on state changes are acceptable in some cases.

Assume that in an action schema, the PRECONDITION list contains the preconditions (literals) of an action, and the EFFECT list contains the effects (literals) of an acton.

The effect list for $[a_1; a_2]$ is:
The effects of a_2 AND the effects of a_1 THAT ARE NOT undone by the effects of a_2

(Note: The conditions that must have been true before a_1 are not effects of the composite action - this does not affect correctness, but having these as effects is redundant if they were true before).

B. When is the composite action impossible? That is, when is it impossible for a_2 to be done immediately after a_1? **(5 marks)**

The action $[a_1; a_2]$, is impossible if:

Case 1/ When the effects of a1 negate a precondition of a2 OR
Case 2/ At least one precondition of a2 is a negated precondition of a1 and this precondition of a2 is also not in the effects of a1.

C. Assuming the composite action is possible, what are the preconditions for this composite action? **(5 marks)**

If it is not impossible, the precondition-list for $[a_1; a_2]$ is
The preconditions of a_1 AND the preconditions of a_2 THAT ARE NOT achieved by the effects of a_1

Question 5 (20 marks) Probability Theory and Normative Decision Theory

On its way to collect rock samples, Happy the autonomous Rover on Mars has encountered many little Martians! Happy has noticed that 1 in every 1000 Martians is friendly. Of those Martians who are friendly, 30% have dark green noses, and 70% have light green noses. Happy assumes that all the other Martians have no green noses.

Happy has been given a test kit to detect from far away whether each Martian it encounters has green nose, and hence friendly, so that it can run away from those who may be unfriendly … The "Green Nose Test", however, has the following characteristics:

- For the Martians with dark green noses, fraction of positive (presence) test results is 90%
- For the Martians with light green noses, fraction of positive test results is 60%
- For the Martians with no green noses, fraction of positive test results is 10% (oh no!)

A. Of all the "Green Nose Tests" that have been performed, what is the probability that the test result is positive? **(5 marks)**

Ans: The probability of getting positive test result is: **~0.1**

[Solution] Let A1 be the event that Martians have "dark green noses"P(A1) = 0.0003

Let A2 be the event that Martians have "light green noses"P(A2) = 0.0007

Let A3 be the event that Martians have "no green noses"P(A3) = 0.9990

Let B be the event "a test result is positive"

P(B) = P(B|A1)P(A1) + P(B|A2)P(A2) + P(B|A3) P(A3)

= (0.9)(0.0003) + (0.6)(0.0007) + (0.1)(0.999)

= 0.10059

B. Is the "Green Nose Test" a good diagnostic test? **[5 marks]**
Ans: The diagnostic test is: **B**

a. Good
b. Bad

This is because the test is **with a very high false positive rate!**

[Solution] What is the probability that a Martian has green nose, given a +ve test result?

Let C be the event "A Martian has green nose" C = A1∪A2; P(C)= 0.001

P(C|B) = P(B|C)P(C) / P(B) **(*)**

P(B|C) = P(B|A1∪A2)

= P(B ∩(A1∪A2))/ P(A1∪A2)

= [P(B|A1)P(A1) + P(B|A2)P(A2)] / P(C) = 0.69 **Substitute back to (*):**

P(C|B) = ~0.00686

This is a terrible diagnostic test, with a very high false positive rate!

To help prepare for future human missions, Happy has to build enough safety zones – fenced areas in the rock sampling sites to keep out the unfriendly Martians. But it is unsure how many safety zones it has to build for a start. Happy has the option of building 0, 10, 20, or 30 zones in its current mission. The number of safety zones actually needed may be 0, 10, 20, or 30 zones, with **equal likelihood**.

Given the limited resources, assume that Happy's performance (which will be determined in future) is represented by the following utility function:

$$U(m, n) = -m - 2n$$

where m = additional number of safety zones needed because it didn't build enough
n = unnecessary number of safety zones unused because it built too many

C. What is the expected utility of having Happy build 30 safety zones? **(5 marks)**

Ans: The expected utility is: **– 30**

D. How many safety zones should Happy build? **(5 marks)**

Ans: Happy should build **10** safety zones

C and D:
[Solution]:

Happy's choice of actions is between building 0, 10, 20, or 30 safety zones. Thus Happy's decision problem is given by the table:

Consequences		**State: (No. of inadequate zones, No. of redundant zones)**			
(m_{ij}, n_{ij})	**No. of safety zones needed**	**0**	**10**	**20**	**30**
Action: No. of safety zones built	**0**	**(0, 0)**	**(10, 0)**	**(20, 0)**	**(30, 0)**
	10	**(0, 10)**	**(0, 0)**	**(10, 0)**	**(20, 0)**
	20	**(0, 20)**	**(0, 10)**	**(0, 0)**	**(10, 0)**
	30	**(0, 30)**	**(0, 20)**	**(0, 10)**	**(0, 0)**

The utilities are:

Action (Safety zones Built)	**Utility(m,n)= -m-2n**				**EU (with assumptions)**
	0	**10**	**20**	**30**	
0	**0**	**-10**	**-20**	**-30**	**-15**
10	**-20**	**0**	**-10**	**-20**	**-12.5**
20	**-40**	**-20**	**0**	**-10**	**-17.5**
30	**-60**	**-40**	**-20**	**0**	**-30**

Assume the probabilities of the states of needing 0, 10, 20, 30 safety zones are: a, b, c, d respectively, with a = b = c = d = 0.25

Expected utility of building:

0 safety zones: **0.25(0 – 10 – 20 – 30) = -15**
10 safety zones: **0.25(-20 + 0 – 10 – 20) = -12.5**
20 safety zones: **0.25(-40 – 20 + 0 – 10) = -17.5**
30 safety zones: **0.25(-60 – 40 – 20 + 0) = -30**

Expected Utility of building 30 safety zones is – 30. Happy should build 10 safety zones.