Entry number: 2021CS50512 **Name:** KUSHAGRA GUPTA

COL 333/671 Autumn 2023 Major

Welcome to the major exam. This exam is for 2 hours. Please use only pens while answering questions. Do not use a pencil. Please do not write outside the margins – that part may not get graded.

If we find any form of collaboration with any other person during the course of exam, it will result in a straight zero on the exam – no exceptions.

Before starting the exam, close your eyes and take three deep breaths. Your performance in the exam is not an accurate reflection of your understanding of the material. Nevertheless, if you are relaxed, you will likely perform better.

Question Number	Maximum Marks	Marks Obtained
1	30	
2	12	
3	17	
4	17	
5	15	
6	36	
7	9	
8	10	
9	4	
Total	**150**	

1. [30 points] Answer the following objective questions. Multiple choices may be correct. You get credit only if you pick all correct choices.

1.1) Which of the following are true regarding the definition of AI?

B C D

(A) Thinking rationally is not an ideal definition because rationality is not an achievable concept.
(B) Acting like humans is not an ideal definition because it presupposes that humans are the final benchmark in intelligence, when they are not.
(C) Thinking like humans is not an ideal definition because humans often do not think methodically; instead take intuitive emotional decisions, which is not appropriate for AI.
(D) Acting rationally is not an ideal definition because, as per strong AI hypothesis, actions are not always indicative of an intelligent thought process.
(E) None of these

1.2) Which of the following are true regarding uninformed search algorithms

B C D.

(A) The ratio of nodes expanded by iterative deepening depth first search to depth limited search tends to one, as branching factor increases
(B) The ratio of nodes expanded by breadth first search to bidirectional search tends to two, as depth of goal increases
(C) Iterative lengthening search is not as effective as iterative deepening search when the minimum cost in graph is small
(D) Beam search has much better asymptotic space complexity than both depth first and breadth first search.
(E) None of these

1.3) If *h1* and *h2* are admissible heuristics (all values positive), then which of the following are also admissible

A C.

(A) *max(h1, h2)*
(B) *1+min(h1, h2)*
(C) $\alpha h1 + (1-\alpha)h2$ where $0 \leq \alpha \leq 1$
(D) $\sqrt{h1.h2}$
(E) None of these

1.4) Assume that we have a function $y = (x - 1)^4$. Starting at $x = 2$, which of the following values of the step size λ will allow gradient descent to converge to the global minimum?

A B C

(A) 0.05
(B) 0.2
(C) 0.5
(D) 0.75
(E) None of these

1.5) What techniques are common between Deep Blue and AlphaGo?

C D.

(A) Deep Neural Networks
(B) Alpha-beta Pruning
(C) Monte-Carlo Tree Search
(D) Database for all moves in beginning and end of the game
(E) None of these

1.6) Which of these are true about constraint satisfaction algorithms

ABCD

(A) K-consistency can be seen as a generalization of arc consistency
(B) If there is an *alldiff* constraint between k variables in a CSP, then its underlying constraint graph will have at least $0.5k^2$ edges
(C) Solving tree-structured CSPs requires no backtracking at all
(D) If a CSP, with nk variables with each domain size d, is known to be decomposable into k independent subproblems of n variables each, then worst case complexity of solving this CSP is $O(kd^n)$
(E) None of these

1.7) Which of the following are examples where fuzzy logic is applicable?

CD

(A) A person is fairly certain that India will win the next cricket match and therefore bets 100 Rs. in India's favour
(B) A person cancels his trip to a museum tomorrow because he thinks it will rain
(C) While playing darts, the thrown dart lands at the edge of two different score regions and the average score of the two is taken
(D) A doctor orders two different diagnostic tests for the same disease to be certain that a patient definitely has it.
(E) None of these

1.8) Which of the following are true regarding the various KR approaches?

ABD

(A) In probability theory, the ontological commitment is degrees of belief
(B) In propositional logic, the epistemological commitment is {true, false, unknown}
(C) In first order logic, the ontological commitment is {true, false, unknown}
(D) In first order logic, the ontological commitment is {facts, objects, relations}
(E) None of these

1.9) Which of the following are true about Bayesian networks?

E

(A) When computing P(Z|E) a variable in ancestor(Z ∪ E) is irrelevant
(B) Given all parents, a variable is independent of all its descendents
(C) If an undirected path between two nodes is cutoff, then they are independent
(D) Parent and child nodes of a given node are collectively called its Markov blanket
(E) None of these

Undirected path after connecting spouses

1.10) Which of the following are true about sampling algorithms in Bayesian networks?

AD.

(A) Bayes net allows easy generation of samples from its prior distribution
(B) Likelihood weighting generates samples from the posterior distribution
(C) Rejection sampling (after rejection) creates samples, as if generated from posterior distribution
(D) Gibbs sampling (after mixing) generates samples from posterior distribution
(E) None of these

1.11) Which of the following are true about randomized algorithms studied in the course?

ABC

(A) Simulated annealing uses randomization to select the next state among the neighbors
(B) Backtracking search with randomization helps in avoiding local optima
(C) Q-learning uses randomization in action selection to visit unexplored states or actions
(D) Maximum likelihood learning uses randomization for smoothing the parameters learned in Bayesian network
(E) None of these

1.12) Which of these are true about Markov decision processes?

ACD

(A) MDPs (as studied in class) assume full observability of the environment
(B) Goal states may have transitions to other states in the MDP
(C) Discount factor is not needed for mathematical well-formedness in finite-horizon MDPs
(D) We assume that the reward and cost models are independent of the previous state transition history, given the current state.
(E) None of these

1.13) Which of these are true regarding Boltzmann exploration?

ABCD

(A) All actions have almost equal probability of being executed initially
(B) It leans more towards exploitation as temperature is increased
(C) It satisfies the GLIE property
(D) The probability of an action being chosen at a particular state varies exponentially with its Q-value at that point in time
(E) None of these

1.14) Which of these are true regarding training in feed-forward neural networks?

ABCD

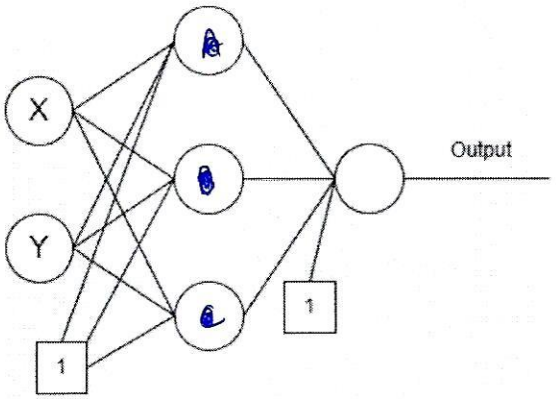
(A) Backpropagation estimates gradient of loss w.r.t. each parameter
(B) Gradient descent may get stuck in local optima or saddle points
(C) Backpropagation computes gradients layer-wise starting from the layer closest to the output
(D) If the neural network does not have any non-linearity then all parameters will have constant-valued gradients
(E) None of these

1.15) Which of these are not one of the tenets of GDPR?

AD

(A) Robustness to data attacks
(B) Maintaining data confidentiality
(C) Limits on the purpose/usage of data
(D) High accuracy of data
(E) None of these

2. [12 points] In the futuristic metropolis of Robo-City, the city's infrastructure is controlled by neural networks. As the city's lead Neural Network Architect (pun intended), you've been tasked with designing systems to manage the flow of robot traffic and ensure efficient energy distribution. You're presented with three distinct blueprints, each representing a different control system:



- Blueprint A (The Intersection Controller): if ((not X) and Y) then output >0 else <0
- Blueprint B (The Energy Equalizer): if (X=Y) then output >0 else <0
- Blueprint C (The Central Hub Regulator): if (X=0.5 and Y=0.5) then output >0 else <0

Note that blueprints A and B only admit Boolean values of X and Y (0 or 1), i.e., there are four possible inputs. However, for blueprint C, either X and Y are Boolean or (X=0.5, Y=0.5), i.e., it admits five possible input configurations.

Your job is to determine which of these blueprints can be learned by your neural network design.

(a) How many parameters does your neural network have? 9

(b) Can your neural network above (without any non-linearity) represent these three blueprints. If your answer for a blueprint is No, calculate the least error (number of misclassified inputs) achieved by your network for that.

Yes, the above neural network can represent the first blueprint if not 0, 1 is linearly separable

X or Y.
if not X and Y } Yes can be represented by 1 neural layer.

$X + (Y-1) - -$

Similarly for the second case yes it is possible

The third case is not possible we need the input (0.5, 1) & (1, 0.5) are misclassified

(c) Robo-City council passed a directive that to save costs, all bias terms will be removed from the neural network. Now, can your neural network above represent the three blueprints. Answer this question for all the blueprints where the answer to part (b) was yes.

Yes No, third can be represented

(d) A breakthrough in neural network technology introduces the ReLU non-linearity. The three nodes in the hidden layer apply ReLU before sending output to the final node. Now, can the blueprints be represented by the neural network? Answer for each blueprint.

Biases are not 0

(e) Continuing from part (d), if the answer is yes for Blueprint C, output any one set of network weights that produce the desired output. If the answer is no, output any one set of network weights that produce least error. For uniformity, let us call hidden nodes H1, H2, and H3. And the output node O.

3. [17 points] Consider an agent that plays a game at the casino. He is given a fair dice with K faces numbered 1 to K and is given only one action a – roll the dice. The agent takes a, i.e., rolls this dice, exactly T times. At every time step, if he gets the face i on the dice, then the casino pays him $i + n_i$ rupees, where n_i represents the no. of times the face i has been obtained so far. At the start, all n_is are initialized to be zero.

(a) [9 points] Formulate this problem as a reward-based MDP *without* a goal state. Define a finite state space, transition function, reward model, and start state. Let discount factor be 1. Be comprehensive in your description, paying special attention to all corner cases.

✓ State space : No. of times each die face has been rolled so far.
i.e it is a vector (K dimensional) representing no. of times face i has been rolled for ith entry.

✓ Initial State (start state s_0) → | 0 | 0 | 0 | 0 | 0 |
1 2 - - - K.

Transition function $T(s, a, s')$ → s' obtained by at this roll (action) the face I obtain would move me to a new state s' by adding 1 to the face obtained for entry.

Action → roll the die & observe the die roll/face

✓ Reward → $R(s, a, s')$ where a is the action.
if a returns face f, $R(s, a, s') = s[f] + f$ $(f \in [K])$

$T(s, a, s') = ?$
$s'[f] = 1 + s[f]$. & $s'[f'] = s[f'] \; \forall f' \neq f$.

✓ $T(s, a, s') = \frac{1}{|K|}$, probability for the state action pairs/triples above.

$\gamma = 1$ for all $a \in [K]$, $s'[a] = s[a] + 1$, $s'[f'] = s[f'] \; \forall f' \neq a$

Note) finite state as every K-dimensional vector with integer entries

(b) [8 points] Let V function have its typical meaning, i.e., it defines the expected long term reward obtained starting in state s. Using your notation above, write the system of equations that govern this MDP. Furthermore, solve the MDP for the start state (s_0), i.e., compute $V(s_0)$ in a closed form.

No goal state,

$$V(s) = \sum_{a} \sum_{s'} T(s,a,s')[R(s') + \gamma V(s')] \quad \text{for all } s$$

given by the action space.

$$V(s) = \frac{1}{K}[\gamma V(s') + r] \quad \text{for } s' \text{ is given next.}$$

$1/K$ ( ) for s_0:

$V(s'_1)$ is

for the given $\gamma = 1$

to solve these equations

We will solve the system of equations

$$V(s_0) = \frac{1}{K}\left[1 + \gamma V(s_1) + 2 + \gamma V(s_2) - - (k+1) + \gamma V(s_k)\right)$$

where s_i is given as 0 0 0 1 0 0 (i-th)

$$= \frac{k+1}{2} + \gamma\left(\frac{V(s_1) - - V(s_k)}{k}\right) \qquad \frac{k(k+1)}{2}$$

This will be our answer in closed form

4. [17 points] You are exploring a mysterious island with a reputation for its magical artifacts. As you delve deeper into the island's dense jungle, you come across four ancient places of worship: a temple, a church, a mosque and a gurudwara. According to local legends, each place possesses a unique mystical power. You talk to the monks at each of these. They give you some insights.

Each monk makes 2 statements except Iman.

I) The priest of the temple shares, "If you approach the mosque, you'll gain insight into the mysteries of Fire. Moreover, if the church guides you, you'll harness the essence of Fire."

II) The imam of the mosque confides, "Neither the church nor the gurudwara holds the secrets of Water's power."

III) The pastor at the church declares, "Follow the path to the hindu temple, and you'll be instilled with the secrets of the Earth. Opt for the gurudwara, and you'll be left empty-handed."

IV) The granthi of the gurudwara reveals, "If you seek the strength of Air, the church will confuse you. Choose the temple, and you shall never learn about mysteries of Air."

You know that all these monks are lying – none of their statements are true. They just want you to get confused and not fulfill your mission. Use propositional logic to infer which place of worship holds the power of Water. First define relevant symbols and convert each of these statements into expressions in propositional logic. Later use one or more inference approaches discussed in class to determine the desired answer. Show each step of your work.

~~Priest => M => F~~

M → Mosque, T → Temple, G → Gurudwara, C → Church.

each variable takes exactly 1 values for which it is true (unique).
↳ F, W, A, E & each element true for only 1 worship place

Priest says
$(M = F \ \& \ C = F)$ which are false.

Iman says
$(\neg(C = W) \ \& \ \neg(G = W))$ which is false

Pastor → $((M = E) \ \& \neg(G = E))$, are

Granthi → $(\neg(C = A) \ \& \neg(T = A))$, are

. we have $\neg(M = F)$, $\neg(C = F)$, $\neg((C = W) \vee (G = W))$,
$\neg(M = E)$, $\neg(G = E)$, $\neg(C = A)$, $\neg(T = A)$ in our KB
alongwith one-hot encoding constraints for every (element, worship) pair i.e $M = F \vee M = A \vee M = E \vee M = W$
for each piece of worship

and $\neg(M = F) \vee \neg(M = W)$
$\neg(M = A) \vee \neg(M = E)$ } 6 such for mosque

--- 24 such and 4 earlier.

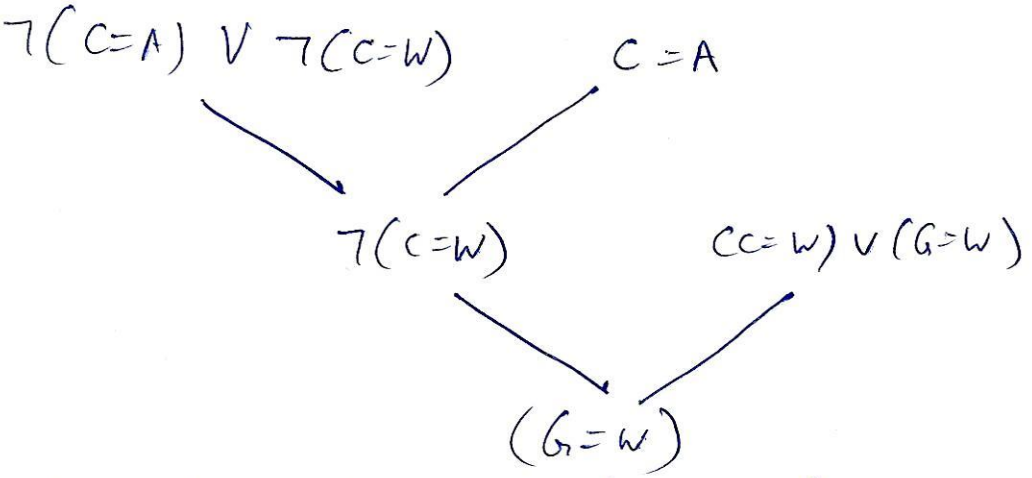
This will ensure that one & only one element for each worship place & vice-versa not vice-versa)).

Secrets & mysteries are same.

Now, we know that
$((C = W) \vee (G = W))$ will be true ~~& both can't be true~~.

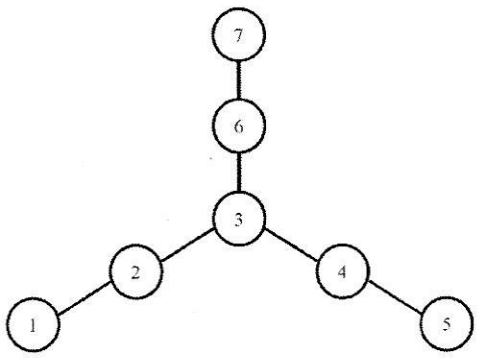
also we know that
$(C = A)$ is true from 4th statement

By resolution of $\neg(C=A) \vee \neg(C=W)$, $(C=A)$,
$(C=W) \vee (G=W)$ we obtain



∴ the Grandmaster has the power of healer

5. [15 points] You are interstellar police and you have intercepted a group of rebels to a cluster of seven small planets. You must apprehend them quickly. Of course, the rebels try to dodge you by moving from planet to planet. They have a slow spaceboat and in each time step, they can only travel to an adjacent planet. However, you have a state of the art space-cruiser and, in each time step, you can quickly warp (jump) to any (one) planet (including the planet you started from) before the rebels finish their trip, and catch them if they are traveling there. Your job is to find a sequence of jumps such that no matter where the rebels start from, and what path they follow, they will be caught, assuming that they always move and never stay at one planet for two contiguous time points. Note that you never know the location of rebels unless you are at the same location as them.



Your goal is to formulate this problem as deterministic single-agent state space search. Define state space, action space, transition function, goal test, and start state. Draw one step search tree (i.e., start state, all actions and all one-step successors of these actions). Assume that at the start, you are in none of the seven planets and the rebels are in any one of them.

~~Goo~~ State Space : Where am I right now ? and where are the rebels ?

Pair $\langle P, R \rangle$ where $R \in [1, -- 7]$

$P \in [\phi, 1 -- 7]$ ϕ only at start

Start state $P = \phi$, $R \in [1, -- 7]$ or $R = R_0$.

Action Space → P jumps to any planet (including its own)

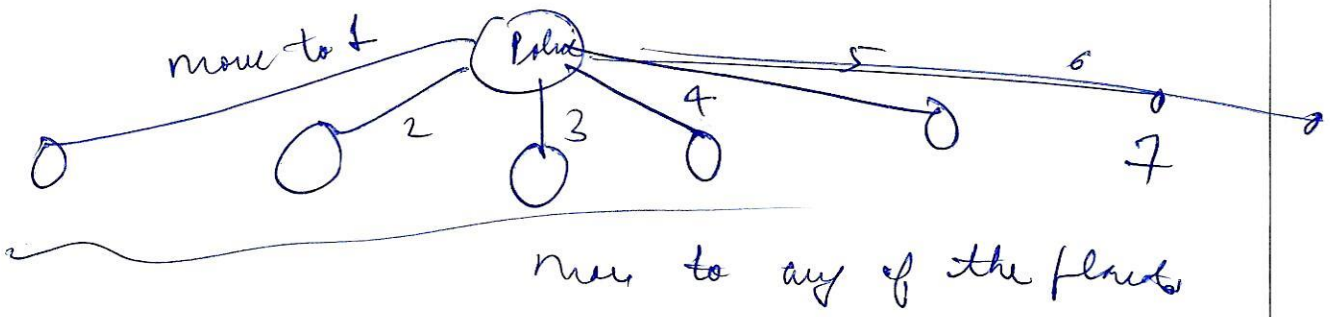
R jumps from ~~about~~ its planet to any one adjacent one.

eg : ~~A ∈~~ (a possible action A → ((1,3) to (7,6)) etc

$T(s, a, s')$.

Transition f" is given by moving from one state to the other.

Goal when $P = R$,



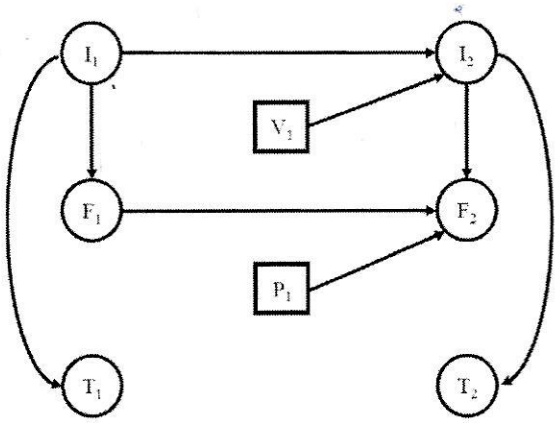
(b) [3 points] Can you think of an admissible heuristic for this problem? A better heuristic fetches more points.

An admissible heuristic could be searching for the n

(c) [**Bonus** 3 points] Find an optimal solution to the problem assuming that, at start, rebels are at an odd planet. *Hint: notice that rebels jump from odd to even and even to odd in each time step.*

(d) [**Bonus** 3 points] Find an optimal solution assuming the rebels can start anywhere.

6. [36 points] Consider the Bayesian network below.



If V_1=False:
$I_2=I_1$ with prob 0.9, I_2 = other values with prob 0.05 each

If V_1=True:

	I_2=H	I_2=M	I_2=L
I_1=H	0.4	0.5	0.1
I_1=M	0.1	0.2	0.7
I_1=L	0	0.2	0.8

If P_1=True, then $P(F_2=Y)$

	I_2=H	I_2=M	I_2=L
F_1=Y	0.7	0.4	0.1
F_1=N	0.1	0.05	0.01

If P_1=False, then $P(F_2=Y)$

	I_2=H	I_2=M	I_2=L
F_1=Y	0.9	0.7	0.5
F_1=N	0.5	0.2	0.01

Here I represents that the patient has infection and it takes three values: high (H), medium (M) and low (L). F represents whether the patient has fever and has two values: Yes (Y) and No (N). The square nodes represent two possible actions: take an antiviral medicine (V) and take a paracetamol (P). The subscripts 1 and 2, represent time instants (let 1 be the first time instant). Even though the Bayes net is shown only for two time instants, the same structure (and CPT tables) is repeated for other time instants too – e.g., there will be an edge from I_2 to I_3 and V_2 to I_3, and so on.

Let CPTs for t=1 be $P(I_1=L) = 0.8$, $P(I_1=M) = 0.1$, and $P(F_1=Y \mid I_1=H) = 0.8$, $P(F_1=Y \mid I_1=M)=0.4$, $P(F_1=Y \mid I_1=L) = 0.35$. Moreover, there is a blood test T, which tells us whether the infection is low (L) or not (NL) with prob. 0.8. I.e., $P(T=L \mid I=L) = 0.8$, $P(T=L \mid I=H \text{ or } M)=0.2$ and similarly, $P(T=NL \mid I = H \text{ or } M) = 0.8$, $P(T=NL \mid I=L) = 0.2$. It does not distinguish between whether the infection is H or M.

(a) [1 point] is I_1 conditionally independent of F_3 given I_2? No (as F_2 lies on one path I_1 to F_3 so not d-separated)

(b) [1 point] is F_1 conditionally independent of F_3 given I_2 and F_2? Yes (d-separated)

For parts (c)-(f) assume that the patient has fever at t=1, but his infection levels are unknown. **Also, throughout this question, approximate final answers to two decimal digits, and as required, use the approximate values for next parts.**

(c) [3 points] Compute the distribution $P(I_1|F_1=Y)$. Show your work.

$$P(I_1 \mid F_1=Y) \propto P(I_1 \cap F_1=Y)$$
$$P(F_1=Y \mid I_1) \cdot P(I_1)$$

$P(I_1=H \mid F_1=Y) = (0.8 \times 0.1) \propto$ as $P(I_1=H) = 1-0.9 = 0.1$

$P(I_1=M \mid F_1=Y) = \propto (0.1 \times 0.4)$

$P(I_1=L \mid F_1=Y) = \propto (0.8 \times 0.35)$

Normalizing to 1, $\alpha(0.08 + 0.04 + 0.28) = 1$

$\alpha = 2.5$

∴ CPT →

$P(I_1 = H \mid F_1 = Y)$	0.2
$P(I_1 = M \mid F_1 = Y)$	0.1
$P(I_1 = L \mid F_1 = Y)$	0.7

(d) [6 points] Compute posterior of I_2 if F_1=Y and the patient took paracetamol at time step 1. Also, compute the posterior if instead patient took antiviral at time step 1. Show your work.

Posterior of I_2 if paracetamol.

$P(I_2 = i \mid F_1 = Y, P_1 = \text{True}, V_1 = F)$.

$= \alpha P(I_2 = i, F_1 = Y, P_1 = \text{True}$

$= \alpha$

first row of table 2 gives us.

for $i = H$, it is $= \alpha \sum_{F_2} P(I_2 = H, F_1 = Y, P_1 = T, F_2)$

$= \alpha \sum_{F_2}$.

$P(I_2 = i \mid F_1 = Y, P_1 = T) = \sum_{F_2} P(I_2 = i, F_2 \mid F_1 = Y, P_1 = T)$

for $F_2 = Y$, first row of table 2

$I_2 = H$ gives us $= \alpha P(F_2 = Y \mid I_2 = H, F_1 = Y, P_1 = T)$.

for $I_2 = H$, $P(I_2 = H \mid F_1 = Y, P_1 = T) = P(I_2 = H, F_2 = Y \mid - - -) + P(I_2 = H, F_2 = N \mid - - -)$.

for $I_2 = H$, $F_1 = Y$, $P_1 = T$, $V_1 = F$., sum over all F_2, I_1.

$\sum P(I_2 = H, F_1 = Y, P_1 = T, V_1 = F, F_2, I_1) = ?$.

as $F_1 = Y$, $I_1 = H$: 0.2×0.9 ; $I_1 = L$: 0.1×0.05 ; $I_1 = M$: 0.1×0.05

$P(I_2 = H, I_1 = H \mid F_1 = Y, P_1 = T) \cdot P(I_1 = H \mid F_1 = Y)$ add them to get $P(I_2 = H \mid F_1 = Y)$.

for $F_2 = Y$

$0.19 = P(I_2 = M \mid F_1 = Y)$, Similarly $= 0.1 \times 0.9 + 0.1 \times$

$P(I_2 = L \mid F_1 = Y)$

$P(I_2 = M \mid F_1 = Y)$

~~$0.2 \times 0.9 \times 0.7$~~

For parts (e) and (f) let rewards be measured in health points to the body. Due to side effects, taking an antiviral costs 30, and taking a paracetamol costs 10. If you reach low infection you get a reward of 500, medium infection results in a reward of 50, and high infection has zero reward.

(e) [4 points] Given F_1 is Y, what is the right medicine to take at time step 1, assuming only one more time step to go (i,e,, you get the reward based on the state at time 2). Show your work.

(f) [12 points] Continuing from part (e), how much should you be willing to pay (in health points) for the blood test at time step 1, if there is only one step to go? To help reduce your calculation, we inform you that if test shows NL, the optimal medicine is antiviral, and if test result is L, then paracetamol is better. Show your work.

(g) [8 points] Now assume that the Bayes net only has I and F nodes (and all relevant edges), and there is no intervention at any time step, i.e., no test is conducted and no medicine is given. The fever values f_t at each time step t are observed, and our goal is to estimate $P(I_{t+1} \mid f_{1:t})$.

Let $\delta_{t+1}(i)$ represent $P(I_{t+1}=i \mid f_{1:t})$, where $i \in \{H, M, L\}$. Derive a recursive equation for $\delta_{t+1}(i)$. At each step of derivation, indicate the probability rule/reason why it is valid.

7. [9 points] We have an MDP with state space S, action space A, reward $R(s,a,s')$ and discount factor γ. Our eventual goal is to learn a policy that can be used by a robot in the real world. However, we only have access to simulation software, not the robot directly. We know that the simulation software is built using the transition model $T_{sim}(s,a,s')$, which is unfortunately different than the transition model that governs our real robot $T_{real}(s,a,s')$. Without changing the simulation software, we want to use the samples drawn from the simulator to learn Q-values for our real robot.

(a) Write down the update equation for Q-learning for a sample (s,a,s',r), if $T_{sim} = T_{real}$

$T_{sim}(s,a,s') = T_{real}(\quad)$

$(s,a,s',r) = \alpha(\quad) + (1-\alpha)(\quad)$

(b) Assuming the samples are drawn from the simulator, suggest a new update rule that will learn the correct Q-value functions for the real world robot? Assume both T_{sim} and T_{real} are known to the agent.

(c) Suppose now you wish to use function approximation. Let the state be comprised of two real valued variables x and y. You define the value of a state as:

$U(x,y) = \theta_0 + \theta_1 x + \theta_2 y + \theta_3 \sqrt{(x-x_g)^2 + (y-y_g)^2}$, where x_g, y_g are both constants. Write down all the TD-learning update equations (to estimate value of a policy π) to learn the parameters, assuming $T_{sim} = T_{real}$.

8. [10 points] Suppose our search space is a tree but has negative costs.

(a) [2 points] Explain why any optimal uninformed algorithm will have to explore the whole of search space.

Because we have negative costs &
we may go into -ve cycles &

(b) [2 points] Will A* search with admissible heuristic return an optimal solution? Why/why not?

Due to negative costs, ⟶ does not underestimate

$h(n) \leq h(n^*)$ $\quad h(n) < c(n, a, n') + h(n')$

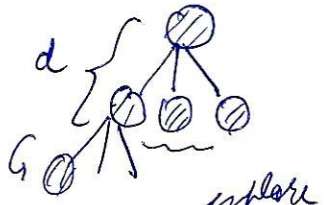
(c) [6 points] Let the max depth of the tree be finite (m) and each edge cost > -r. Suppose we ran a search algorithm and found the first goal G at depth d, with path cost c. Give an optimality bound for (c-c*) in terms of d, r and m, where c* is the optimal cost, if the search algorithm is:

(i) depth first search

we know that c^* can be at minimum $-rm$.

$c^* > -rm$.

$\therefore c - c^* < d + rm$.



(ii) breadth first search

$|(c-c^*)| < (m-d)[r] + d$.

(iii) uniform cost search

$|c-c^*| < (m-d)[r] + d$

9. [4 points] Explain the use of supervised learning and reinforcement learning for the training of ChatGPT. For supervised learning, explain the input, the desired output and how a supervised dataset is created. For reinforcement learning, describe the state (input), the action, and the environment which gives reward.

For supervised learning we provide the input dataset to the model, which is obtained by looking at all (phrase, next word) pairs from the multiple sources available throughout the world in Chat GPT.

So the collection of sentences yield us this sentence completion dataset.

Eg: (AI is a good course) generates AI is, AI is a, AI is a good, AI is a good course)

The desired output is the weights/parameters for the possible words that may be predicted next.

The input is any partial sentence.

For the reinforcement learning approach, it asks itself a lot of questions and obtains answer to these to generate new sentences. Here, the state (input) is the current information / (phrase, next word) pairs in it, the action is asking itself a question. And it gets the reward if it increases the size of its dataset over which it is trained, by asking itself a question.

[EXTRA SHEET]