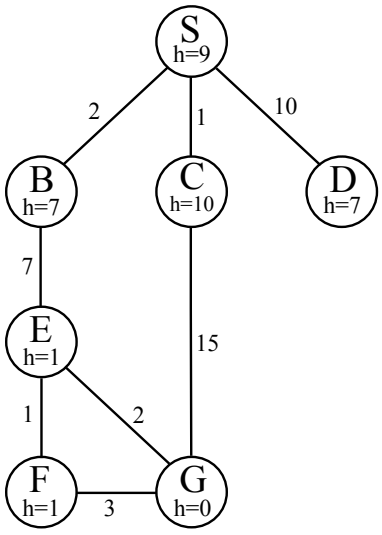
Table 1: For instructor's use

Question	Points Scored	Possible Points
1		12
2		12
3		8
4		12

This exam is closed book. You are allowed 2 sheets of notes (4 pages front and back). You may use any format for your notes that you like. Please explain all of your answers fully to receive full credit.

Here is some extra space. **Show all of your work on the questions!** If you need more paper just ask. Good luck!!

Question 1. Consider the graph below, where S is the start state and G is the goal state.



For each of the following search strategies, give the *solution path* that would be returned. Break any ties alphabetically (i.e., nodes for states earlier in the alphabet are expanded first during ties). Costs are given for the edges and heuristic values, h, are given for each state.

(a) (1 pts) Depth-first graph search

S-B-E-F-G if successors are pushed right to left, S-C-G if pushed left to right

(b) (1 pts) Breadth-first graph search

S-C-G, shortest path with edge costs of 1, frontier at termination will be D E

(c) (1 pts) Uniform cost graph search

S-B-E-G, shortest path counting edge costs

(d) (1 pts) Greedy best-first graph search

S-B-E-G

(e) (2 pts) A* graph search

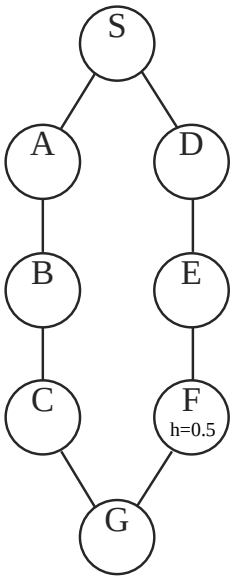
S-B-E-G

(f) (2 pts) Give the *Explored List* produced by A* graph search in part (e)

S B E C F G, note following expansion of E the frontier is C(11), F(11), G(11), D(17).

For the following question parts, all of the edges in the graphs have cost 1.

(g) (4 pts) Suppose that you are designing a heuristic h for the graph on the right. You are told that $h(F) = 0.5$, but given no other information. What ranges of values are possible for $h(D)$ if the following conditions must hold? Your answer should be a range, e.g. $2 \leq h(D) < 10$. You may assume that h is nonnegative.

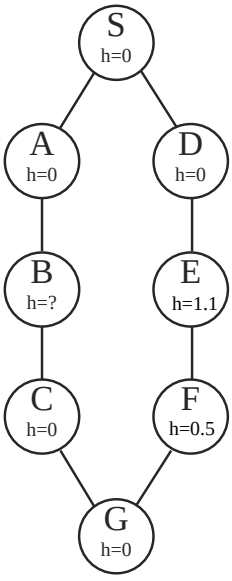


1. h must be admissible
   $0 \leq h(D) \leq 3$
   The path to goal from D is 3.

2. h must be consistent
   $0 \leq h(D) \leq 2.5$
   In order for $h(E)$ to be consistent, it must hold that $h(E) - h(F) \leq 1$, since the path from E to F is of cost 1. Similarly, it must hold that $h(D) - h(F) = h(D) - 0.5 \leq 2$, or $h(D) \leq 2.5$.

EXTRA CREDIT - 3 Points

Answering this part is *not required* and is worth 3 points of *extra credit* if completed correctly.



Now suppose that $h(\text{F}) = 0.5$, $h(\text{E}) = 1.1$, and all other heuristic values except $h(\text{B})$ are fixed to zero (as shown on the right).

Find a value for $h(\text{B})$ that yields an admissible heuristic AND results in the following behavior for A* search:

B is expanded before E, and E is expanded before F

$0.0 \leq h(\text{B}) \leq 1.1$

Question 2. Mark each of the following statements as *TRUE or FALSE*. **(2 pts each)**
If FALSE, **rewrite the sentence** changing just a few words to make it true.

- In an *episodic* environment, an AI agent performs a series of tasks where each task is an independent activity that doesn't depend on the task that came before it. TRUE

- In A* search, as soon as a goal state is ~~generated by get-successors~~ popped from the frontier, it will pass the goal test and the search will terminate. FALSE

- The primary difference between iterative deepening DFS and BFS is that iterative deepening ~~is guaranteed to find the optimal state with fewer node expansions (i.e. shorter explored list) than BFS.~~ uses less memory than BFS while retaining completeness. FALSE

- Let P be a joint probability table for a set of discrete random variables. If you pick one variable and *marginalize* it out of P, the resulting table will be smaller (contain fewer numbers). TRUE

- The reason that Q-Learning is an *off-policy* method is because it always updates the Q-function based on the best action that is available, regardless of the current policy being used to collect the data. TRUE

- For a given MDP, suppose that you have a value function V^k which is *not optimal*. You explore two ways to get V^{k+1}. First, you perform one round of value iteration (updating the value of each state) to get V_v^{k+1}. Second, starting again with V^k, you perform policy improvement followed by policy evaluation to get V_p^{k+1} (i.e. one round of policy iteration). It follows that $V_v^{k+1} \neq V_p^{k+1}$. TRUE

Question 3. Consider the following probability tables that define the joint probability density $P(A, B, C)$, which is given by $P(A, B, C) = P(A)P(B)P(C|A, B)$.

	$C = F$	$C = T$
$A = F, B = F$	1	0
$A = F, B = T$	0.5	0.5
$A = T, B = F$	1	0
$A = T, B = T$	0	1

$A = F$	0.75
$A = T$	0.25

$B = F$	0.25
$B = T$	0.75

Compute the values of the following probabilities:

(a) (2 pts) $P(A = T, B = T) = P(A = T)P(B = T) = 3/16 = 0.1875$, by independence of A and B.

Hint: A and B are independent.

(b) (3 pts) $P(C = T) = \sum_{A,B} P(A, B, C) = \sum_{A,B} P(A)P(B)P(C = T|A, B) =$
$P(A = F)P(B = F)P(C = T|A = F, B = F)+$
$P(A = F)P(B = T)P(C = T|A = F, B = T)+$
$P(A = T)P(B = F)P(C = T|A = T, B = F)+$
$P(A = T)P(B = T)P(C = T|A = T, B = T) =$
$\frac{3}{4} \times \frac{1}{4} \times 0 + \frac{3}{4} \times \frac{3}{4} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{4} \times 0 + \frac{1}{4} \times \frac{3}{4} \times 1 = 15/32 = 0.46875$

(c) (3 pts) $P(A = T, B = T|C = T) = \frac{P(A=T,B=T,C=T)}{P(C=T)} = \frac{P(A=T)P(B=T)P(C=T|A=T,B=T)}{P(C=T)} =$
$(\frac{1}{4}\frac{3}{4}1)/\frac{15}{32} = 2/5 = 0.4$

Question 4. **Blackjack!** In this problem you will play a simplified version of the game of blackjack that we discussed in class. Your goal is to beat the dealer's total without going over 21, and we assume here that the dealer always has a fixed total of 15. Let the state S_t represent the *total points* from adding up your cards, unless the total is 22 or higher in which case the state is *bust*. The initial state S_0 is the *sum of the first two cards* you are dealt.

At each turn, your two action choices are *hit* or *stay*. If you choose to *hit*, you receive no immediate reward and are dealt an additional card. If you *stay*, you receive a reward:

$$R(S_t, stay) = \begin{cases} 0 & \text{if } S_t = 15 \\ +10 & \text{if } 16 \leq S_t \leq 21 \\ -10 & \text{if } S_t < 15 \text{ or } S_t = \textit{bust}, \end{cases}$$

and then the game terminates. If your state is *bust*, then you must choose the action *stay*.

The standard 52-card deck contains 4 of each of the 13 cards: 2 through 10, J, Q, K, and A. To simplify the problem, you should assume that each of these cards is always *equally likely* to be drawn (i.e. cards are drawn *with replacement*[1]). Each number card is worth its face value (e.g. the 5 card is worth 5 points), the cards J, Q, and K are worth 10 points, and A is *always* worth 11 points.

(a) (1 pts) What is the state space for this MDP (i.e. the possible values for the state S_t)?

$\{4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, bust\}$, note smallest hand is $(2, 2)$.

(b) (3 pts) Compute the following three probabilities:

$P(S_0 = 21) = \frac{4}{13} \times \frac{1}{13} + \frac{1}{13} \times \frac{4}{13} = 8/169 = 0.047$. Note there are two hands that make 21: $(10, A)$ or $(A, 10)$, and there are four 10 cards out of 13 $(10, J, Q, K)$. You can do 4/52 instead of 1/13.

$P(S_1 = 16 | S_0 = 12, A_0 = \textit{hit}) = 1/13 = 0.077$. You must get a 4.

$P(S_1 = \textit{bust} | S_0 = 14, A_0 = \textit{hit}) = 7/13 = 0.54$. You need an 8 or higher to bust, so 7 cards in total: $\{8, 9, 10, J, Q, K, A\}$

[1]This means that after a card is dealt, it's value is noted and then it's shuffled back into the deck before the next card is dealt. Since the number of cards in the deck never changes, the probability of drawing any specific card, such as a 2 or a J, is always the same.

(c) (4 pts) You decide to solve the MDP using *value iteration*. After performing k iterations, the current value function $V^k(S)$ is given in the table below.

Calculate $V^{k+1}(12)$. Assume a *discount factor* of $\gamma = 0.75$.
The update equation for value iteration is:

$$V^{k+1}(S_t) = \max_A \{R(S_t, A) + \gamma \sum_{S_{t+1}} P(S_{t+1}|S_t, A)V^k(S_{t+1})\}.$$

S	$V^k(S)$
13	2
14	10
15	10
16	10
17	10
18	10
19	10
20	10
21	10
bust	-10

$$V^{k+1}(12) =$$

Note that states that are not relevant to the question have been omitted from the table.

$$V^{k+1}(12) = 0.75 \times \max_{S,H} \begin{cases} \text{stay} : -10 \quad \text{(Reward for terminating} < 15) \\ \text{hit} : 8/13 \times 10 + 5/13 \times -10 = 30/13 = 2.31 \end{cases}$$

Note: For the *hit* action: 8 cards (2 through 9) lead to state with value 10, 5 cards (10 through A) lead to bust state with value -10.

(d) (4 pts) You decide to try *Q-Learning*, since you think the probability model for the cards might be wrong. Given the partial table of initial Q-values below, *fill in* the partial table of *updated* Q-values below using the episode data provided. Assume a learning rate of 0.5 and a discount factor of 1. Leave blank any values which Q-learning does not update. The update equation for Q-Learning is:

$$Q^{k+1}(S_t, A_t) = (1-\alpha)Q^k(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_{A'} Q^k(S_{t+1}, A'))$$

Initial Q-values

S	A	$Q(S,A)$
19	hit	-2
19	stay	5
20	hit	-4
20	stay	7
21	hit	-6
21	stay	8
bust	stay	-8

Episode

S	A	R	S	A	R	S	A	R
19	hit	0	21	hit	0	bust	stay	-10

Updated Q-values

S	A	$Q(S,A)$
19	hit	3
19	stay	
20	hit	
20	stay	
21	hit	-7
21	stay	
bust	stay	-9

Note: You *do not* need to indicate which iteration of Q-Learning produced each updated value. Just record the updated Q-function after all the episode data has been processed.

$Q(19, \text{hit}) = 0.5(-2) + 0.5(0 + \max\{-6, 8\} = 3$ Note: max over two rows with state 21
$Q(21, \text{hit}) = 0.5(-6) + 0.5(0 + \max\{-8\}) = -7$
$Q(\text{bust}, \text{stay}) = 0.5(-8) + 0.5(-10) = -9$ Note: game terminates, no successor