Homework 3 of CS 165A (Winter 2019)

University of California, Santa Barbara

Assigned on February 21, 2019 (Thursday)
Due at 12:30 pm on March 7, 2019 (Tuesday)

Notes:

- Be sure to read "Policy on Academic Integrity" on the course syllabus.
- Any updates or correction will be posted on the course Announcements page and piazza, so check there occasionally.
- You must do your own work independently.
- Please typeset your answers and you must turn in a hard copy to the CS 165A homework box in the copy room of Harold Frank Hall before the due time or turn in at the beginning of due date's class.
- We also encourage you to submit a digital copy on the GauchoSpace for record purpose, we won't grade this.
- Keep your answers concise. In many cases, a few sentences are enough for each part of your answer.

Did you receive any help whatsover from anyone in solving this assignment?

Did you give any help whatsover to anyone in solving this assignment ?

Problem 1 (20')

Give a complete problem formulation for each of the following. Choose a formulation that is precise enough to be implemented. Specify the initial state, goal state, successor function, and cost function.

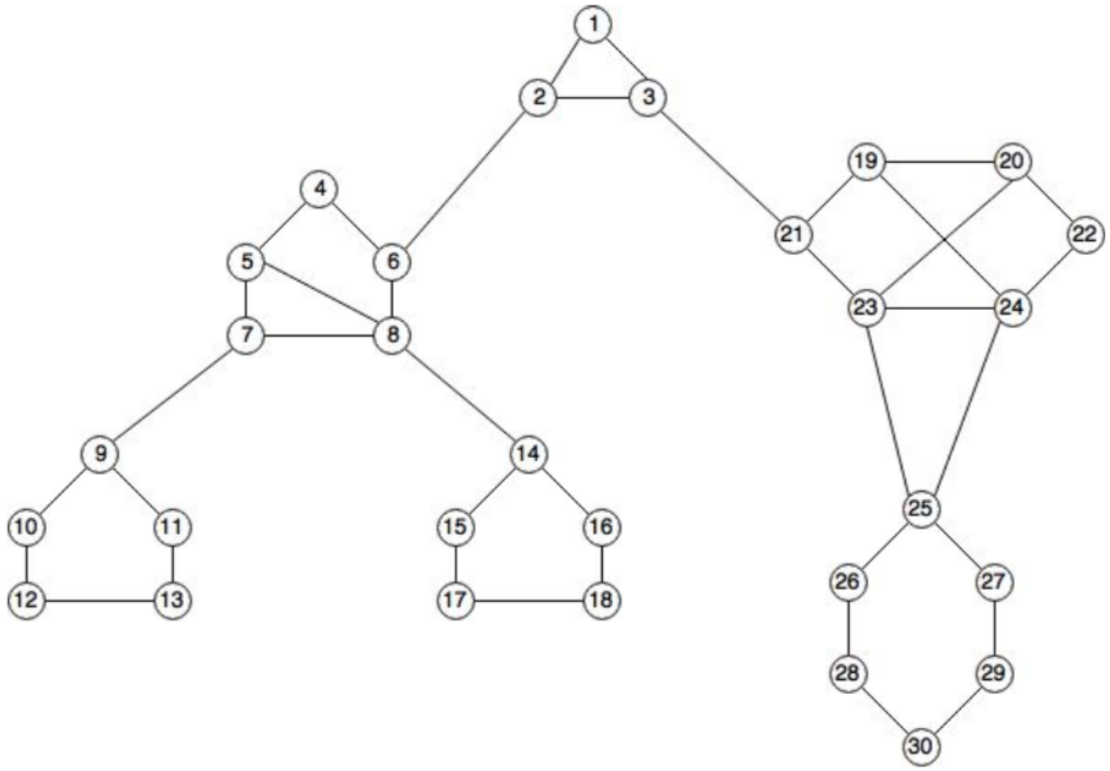
(a) Using only four colors, you have to color a planar map in such a way that no two adjacent regions have the same color.(5')

(b) A 3-foot-tall monkey is in a room where some bananas are suspended from the 8-foot ceiling. He would like to get the bananas. The room contains two stackable, movable, climbable 3-foot-high crates.(5')

(c) You have a program that outputs the message "illegal input record" when fed a certain file of input records. You know that processing of each record is independent of the other records. You want to discover what record is illegal.(5')

(d) You have three jugs, measuring 12 gallons, 8 gallons, and 3 gallons, and a water faucet. You can fill the jugs up or empty them out from one to another or onto the ground. You need to measure out exactly one gallon.(5')

Problem 2 (20')



Consider the state space shown above. Assume state 12 is the start state and state 30 is the goal state.

(a) Assuming a uniform cost of 1 on each edge, simulate the execution of BFS, DFS, IDS (assuming that the depth increases by 1 beginning from 3 to 5) and show the order of states visited. Assume that lower number children are visited first.

(b) Now, simulate the execution of bidirectional search (assuming uniform cost of 1 on each edge and BFS as the basic search from each end). At which state do the two searches meet?

(c) Now, we consider non-uniform weights on edges. Assume that edges between even-even and odd-odd numbered edges have a cost of 1 and those between even-odd numbered edges have a cost of 2. Repeat the goal search using uniform-cost search.

(d) Now, we add a heuristic h to the search. Denote states 1-3 as cluster A, 4-8 as cluster B, 9-13 as cluster C, 14-18 as cluster D, 19-24 as cluster E, and 25-30 as cluster F. Heuristic h estimates costs to the goal state 30 as follows:

   i. $h(30) = 0$

ii. h(all nodes in cluster F except 30) = 1

iii. h(all nodes in cluster E)=2

iv. h(all nodes in cluster A)=3

v. h(all nodes in cluster B) = 4

vi. h(all nodes in cluster C)=5

vii. h(all nodes in cluster D) = 5

a. Is this heuristic admissible? Prove or disprove.

b. Is it consistent? Prove or disprove.

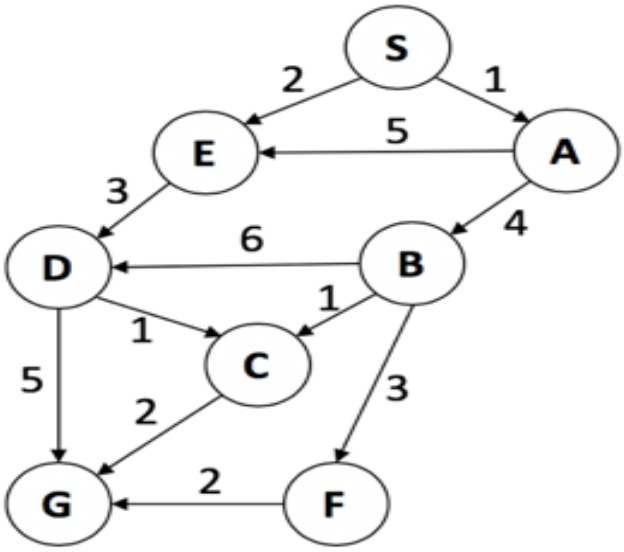
c. If the heuristic is consistent, repeat the search for the goal state using A* (GRAPH-SEARCH).

Problem 3 (15')

The state space description for a problem is shown below, with S being the start state and G being the goal state. Shown on the graph are path costs (g) between states. The table lists the estimated distance from a state to the goal. Assume these estimates are admissible. Perform an A* search for this problem, showing:

(a) the ordered list of the expanded node and all the current f-value calculated

(b) the best path from S to G. (That list starts with expanding S.).



Node	h
S	7
A	6
B	3
C	2
D	3
E	5
F	2

Now we move on to understand the theoretical guarantee of A* search.

(c) We claimed in the lecture that A* search algorithm is known to be "optimally efficient" for a given heuristic function h. The precise meaning of this statement is the following:

> **Optimal efficiency**: Let OPT be the optimal path cost of a problem. Among all optimal search algorithms that start from the from the same start node and uses the same heuristic function h, the A* Search algorithm expands the minimum number of paths p for which $f(p) := \text{Cost}(p) + h(p[-1]) < \text{OPT}$.
> *OPT denotes the optimal path cost.
> *$p[-1]$ is the python notation that denotes the last node of a path p.

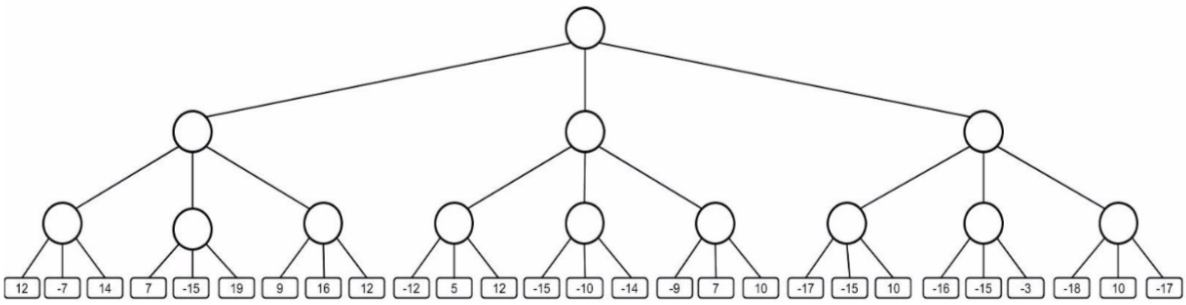
Prove the above statement.

(Hint: Proof by contradiction by following the steps below.

1) Assume there is another optimal search algorithm A' that expands one redundant path that A* search does not.
2) Construct an alternative search problem where the redundant path is part of the optimal path.
3) Show that A' behaves identically on the original search problem and this alternative search problem and hence show that A' cannot be optimal (on this alternative search problem).

Problem 4 (10')

Follow the Minimax algorithm and put a number in each rectangle and circle in the following game tree. Indicate which nodes (subtrees) are pruned and the type of pruning (alpha or beta). Indicate the values of alpha and beta at each pruning step.



Problem 5 (20') Bandits for Clinical Trials

The task of a clinical trial is to estimate the average treatment effect of applying a drug. Let $0 \leq Y \leq 1$ be the the response indicating whether the patient recovered after receiving a

treatment. Let T be the binary variable denoting the treatment assignment. $T = 1$ indicates that the actual drug is given and $T = 0$ indicates that the patient is assigned to a control group where placebos are given instead.

$$\text{ATE} = \mathbb{E}[Y|\text{do}(T = 1)] - \mathbb{E}[Y|\text{do}(T = 0)]$$

Where Y depends both on the treatment T and unknown health conditions of the patient. The do operator is there to indicates that we are fixing the value of T by intervention rather than conditioning on the value of T when the data is drawn from certain joint distribution of Y and T. $\mathbb{E}[Y|\text{do}(T = 1)]$ denotes the expected value of Y when all patients are treated with the actual drug, while $\mathbb{E}[Y|\text{do}(T = 0)]$ denotes the expected value of Y when all patients are given only placebos.

(a) The standard approach of clinical trial involves randomly assigning treatment and control. Let the probability of treatment be p and the probability of control be $1 - p$. Write down an unbiased estimator of the ATE using $(Y_1, T_1), ..., (Y_n, T_n)$ and p.

(b) In more advanced clinical trials, the treatments are determined as a function of the patients condition, measured using a feature vector $X \in \mathbb{R}^d$. Assume $X_1, ..., X_n \sim \mathcal{D}$ iid from an unknown patient distribution $\mathcal{D}$. Let μ be a randomized policy that chooses treatment $T_i = 1$ with probability $0 < \mu(X_i) < 1$. Write down a similar importance weighting estimator of the ATE using $(Y_1, X_1, T_1), ..., (Y_n, X_n, T_n)$ and μ.

(c) Suppose you are to use a 2-armed bandit algorithm rather than running a pre-determined clinical trial. One arm is treatment and the other arm is control. Explain in laymans term (so a doctor can understand) what does a regret bound mean in this setting.

(d) Suppose you are running a contextual bandit algorithm for this problem where we learn a reward estimator $\hat{Y}_t : \mathbb{R}^d \to \mathbb{R}$. $\hat{Y}_t$ is a function of $(Y_1, X_1, T_1), ..., (Y_{t-1}, X_{t-1}, T_{t-1})$ that we obtained by a supervised learning algorithm. Now let the contextual bandit algorithm that we use be ϵ-Greedy, which takes $T_t = \text{argmax}_T \hat{Y}_t$ with probability $1 - \epsilon_t$ and takes T_t uniformly at random with probability ϵ_t for a fixed sequence of $\epsilon_{1:n}$.

Write down an importance sampling based ATE estimator using $(Y_1, X_1, T_1), ..., (Y_n, X_n, T_n)$ and $\hat{Y}_1, ..., \hat{Y}_n$ and $\epsilon_1, , \epsilon_n$. Prove that it is unbiased.

Problem 6 (15') MDP for Rock-Paper-Scissors

We talked about adversarial search in two-player, perfect information, zero-sum game with deterministic transitions. Lets consider a game which fall into this category and we do not even have states at all Rock-Paper-Scissors. Two players are supposed to take actions together.

The payoff matrix for this game is given in Figure 1.

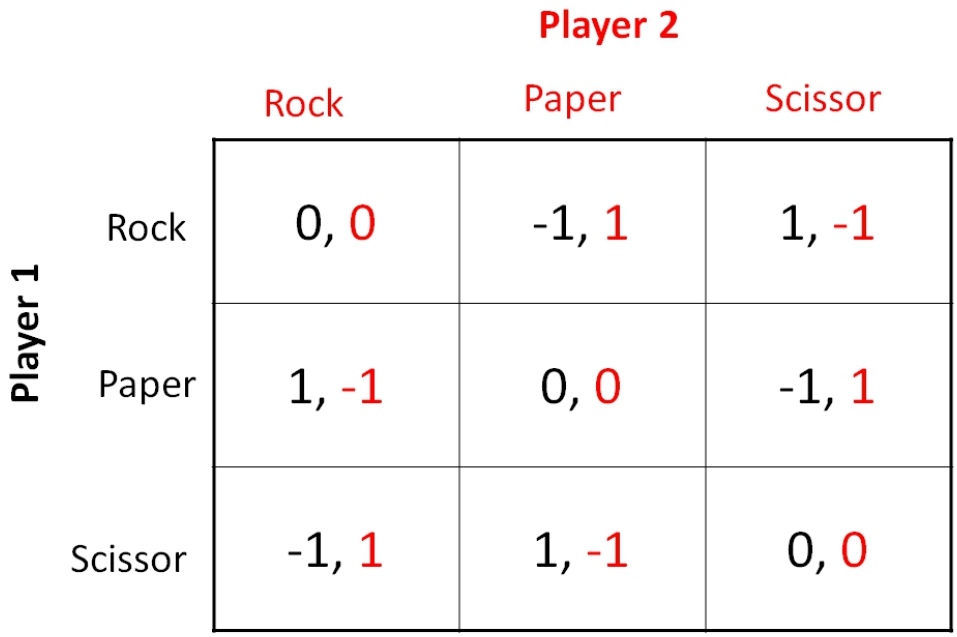


Figure 1: The payoff matrix of rock-paper-scissors.

It is well-known that the minimax strategy of this game is randomized, and it is to take each action uniformly at random with probability 1/3. However, this is not really an interesting strategy.

It is well-known that human beings are not able to generate random numbers. Let us consider an infinite sequence of Rock-Paper-Scissors and build a Markov Decision process to exploit this weakness of a human player.

Denote the sequence of actions of a human player by $a_1, ..., a_t, \in \mathcal{A}$, and the sequence of actions of the agent by $b_1, ..., b_t, \in \mathcal{A}$, where $\mathcal{A} = \{\text{Rock}, \text{Paper}, \text{Scissors}\}$.

The agent believes that human players action is a Markov Decision process where the state at time t is (a_{t-1}, b_{t-1}) for all $t = 2, 3, 4, .$

Note that this is a somewhat strange MDP, because the state is in fact given jointly by the action of of the two players in the past.

(a) Let the agent and human both take their first action uniformly at random. Then the agent runs a fixed (possibly randomized) policy $\mu : \mathcal{A}^2 \to \mathcal{A}$. $\mu(a|s)$ denotes the conditional probability table of taking action a at state s. Write down the human players MDP (Initial state distribution, state-transition matrix, reward distribution given state and action) and as a function of μ.

(b) By symmetry, if the human player is running a policy $\pi : \mathcal{A}^2 \to \mathcal{A}$, then the agent can view the world exactly the same as you derived in (a), except that we replace μ by π.

This means that we can drive an optimal policy to beat a human provided that we can estimate π. Assume π is known, write down the *Q-function* of this MDP as a function of π and the transitions, hence, work out the optimal policy.

(c) Let F be the function you derived in (b) that takes a human strategy π and output the optimal agent strategy $F(\pi)$. Similarly, by symmetry, when the agents strategy is μ, the optimal human player strategy will then be $F(\mu)$. If both parties update their policies alternatively, namely, $\mu_1 = F(\pi_1)$, $\pi_2 = F(\mu_1)$, $\mu_2 = F(\pi_2)$...

Find a fix point μ, π such that $\mu = F(\pi)$ and $\pi = F(\mu)$.