Homework 4 of CS 165A (Winter 2019)

University of California, Santa Barbara

Assigned on March 7, 2019 (Thursday)
Due at 12:30 pm on March 14, 2019 (Thursday)

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 (30') 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.

		Player 2		
		Rock	Paper	Scissor
Player 1	Rock	0, 0	-1, 1	1, -1
	Paper	1, -1	0, 0	-1, 1
	Scissor	-1, 1	1, -1	0, 0

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) (10') 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) as a function of μ.

(b) (10') 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) (10') 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 $\pi = F(\mu)$.

Problem 2 (40') Bellman equation of policy π

Consider an infinite horizon discounted MDP with discount factor γ. Let S and A be the number of states and actions there are and π be a policy. Let $V^\pi \in \mathbb{R}^S$ be the value function and $P^\pi \in \mathbb{R}^{S\times S}$ be the transition matrix under π, i.e., $V^\pi[s]$ is the value of policy π when we start from state s and $P^\pi[s', s]$ is the probability of transition from state s to state s' when the action is taken by policy π. Moreover, let $r^\pi \in \mathbb{R}^{S\times S}$ be the expected reward matrix under policy π, namely,

$$r^\pi[s', s] = \mathbb{E}_{a\sim\pi(s)}[r|s, s'].$$

We learned in the lecture that the corresponding Bellman equation for the value function is

$$V^\pi(s) = \sum_a \pi(s, a) \sum_{s'} P^a_{ss'}[r^a_{ss'} + \gamma V^\pi(s')].$$

(a) (10') Re-write the Bellman equation above in a matrix-form using V^π, P^π and r^π.

- Hint 1: Note that we are marginalizing a. Get rid of a in the above equation first before making it the matrix form.
- Hint 2: For writing the matrix form, you might need to use entrywise product, which we denote it using $\circ$. Also we use $\mathbf{1}_n$ to denote a vector of all 1 of dimension n.

(b) (10') Write down a closed-form solution for V^π.

(c) (10') Start from an arbitrary initialization V_0^π. Repeatedly apply the Bellman equation using the matrix equation you derive in Part (a) for k times. Write down an expression for V_k^π. Try to simplify your solution as much as possible.

(d) (10') Assume $\gamma < 1$. When does the iterative algorithm in (c) converge to the solution in (b) as $k \to \infty$? (Hint: P^π is a transition matrix (all rows are valid probabilities). All transition matrices have right eigenvalues between 0 and 1. What does it say about the operator norm of this matrix — its largest singular value?)

Problem 3 (15')

Assertion: According to some political pundits, a person who is radical (R) is electable (E) if he/she is conservative (C), but otherwise is not electable.

Which of the following are correct representations of this assertion? Briefly explain your reasoning.

- $(R \wedge E) \Leftrightarrow C$
- $R \Rightarrow (E \Leftrightarrow C)$
- $R \Rightarrow ((C \Rightarrow E) \vee \neg E)$

Problem 4 (15')

Consider the following sentence:

$$[(\text{Food} \Rightarrow \text{Party}) \vee (\text{Drinks} \Rightarrow \text{Party})] \Rightarrow [(\text{Food} \wedge \text{Drinks}) \Rightarrow \text{Party}].$$

(a) Convert the left-hand and right-hand sides of the main implication into CNF, showing each step.

(b) Using resolution, determine whether the sentence is valid, satisfiable (but not valid), or unsatisfiable.