CSE 471

Problem Set 2

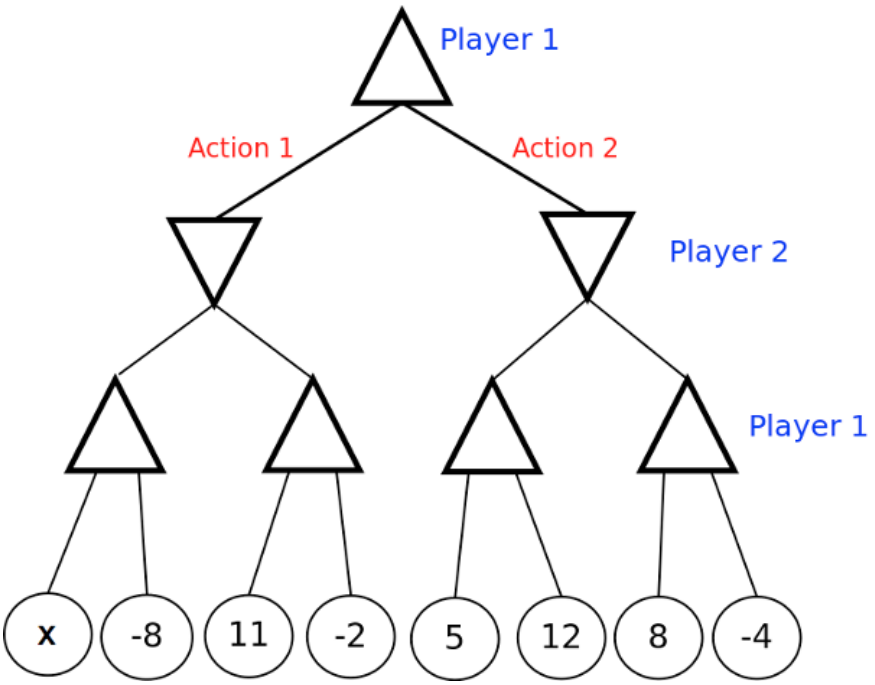
For the following questions, please keep your answers as brief as possible. No reason to fill full pages.

Part A (Game Trees):

1. This problem exercises the basic concepts of game playing, using tic-tac-toe as an example. We define X_n as the number of rows, columns or diagonals with exactly n X's, and no O's. Similarly, O_n is the number of rows, columns or diagonals with just n O's. The utility function assigns +1 to any position with X_3=1 and -1 to any position with O_3=1. All other terminal positions have utility 0. For non-terminal positions, we use a linear evaluation function defined as $Eval(s) = 3X_2(s)+X_1(s)- (3O_2(s)+O_1(s))$.

    a. Show the whole game tree starting from an empty board down to depth 2 (i.e. one X and one O on the board), taking symmetry into account.
    b. Mark on your tree the evaluations of all the positions at depth 2.
    c. Using the min-max algorithm, mark on your tree the backed-up values for the positions at depths 1 and 0 and use those values to choose the best starting move.
    d. Circles the nodes at depth 2 that would *not* be evaluated if alpha-beta pruning were applied, assuming the nodes are generated in *the optimal order for alpha-beta pruning*.

2. Consider the following game tree, where one of the leaves has an unknown payoff, x. Player 1 moves first, and attempts to maximize the value of the game.



a. Assume Player 2 is a minimizing agent (and Player 1 knows this). For what values of x is Player 1 guaranteed to choose Action 1 for their first move?
b. Assume Player 2 chooses actions at random with each action having equal probability (and Player 1 knows this). For what values of x is Player 1 guaranteed to choose Action 1?
c. Is it possible to have a game, where the minimax value is strictly larger than the expectimax value?

Part B (MDPs):

3. Consider an undiscounted MDP having three states (1,2,3), with rewards -1, -2 and 0 respectively. State 3 is a terminal (sink) state. In states 1 and 2, there are two possible actions: a and b. The transition model is as follows:
    a. in state 1, action a moves the agent to state 2 with probability 0.8 and makes the agent stay put with probability 0.2.
    b. in state 2, action a moves the agent to state 1 with probability 0.8 and makes the agent stay put with probability 0.2.

c. in either state 1 or state 2, action b moves the agent to state 3 with probability 0.1 and makes the agent stay put with probability 0.9.

Answer the following questions:

3.1. Write down the transition matrices Mij for each action.

3.2. How many possible policies are there for this problem? What do you think will be the optimal policy? (guess, without doing the problem). Explain the intuition behind your guess.

3.3. [Value Iteration] Suppose we initialize the value vector to be the immediate reward of the states. Do one (synchronous) iteration of the value iteration method. Show the resulting value vector.

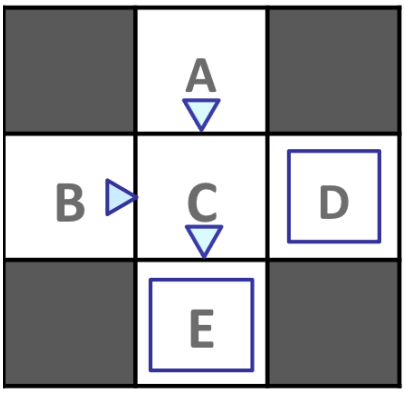
3.4. [Policy Iteration] Suppose we initialized the policy vector to be "do action b" in both states 1 and 2. Apply policy iteration, showing each step in full, to determine the optimal policy. [Hint: the policy evaluation equations are quite easy to solve in this case].

3.5. What is the value vector corresponding to the optimal policy in d?

3.6. What happens to policy iteration if the initial policy has action a in both states? Does discounting help? Does the optimal policy depend on the discount factor?

Part C (Model Based RL and Direct Evaluation):

Input Policy π



Assume: $\gamma = 1$

Observed Episodes (Training)

Episode 1

A, south, C, -1
C, south, E, -1
E, exit, x, +10

Episode 2

B, east, C, -1
C, south, D, -1
D, exit, x, -10

Episode 3

B, east, C, -1
C, south, E, -1
E, exit, x, +10

Episode 4

A, south, C, -1
C, south, E, -1
E, exit, x, +10

4. Answer the following
    a. What transition model can be learned from the above observed episodes? For all the state-action-state pairs, find the transition probabilities that can be learned from the above episodes.
    b. What value estimates can be obtained for the states by direct evaluation from the above episodes?

5. Consider an MDP with 3 states, A, B and C; and 2 actions Clockwise and Counterclockwise. We do not know the transition function or the reward function for the MDP, but instead, we are given samples of what an agent experiences when it interacts with the environment (although, we do know that we do not remain in the same state after taking an action). In this problem, we will first estimate the model (the transition function and the reward function), and then use the estimated model to find the optimal actions.

    To find the optimal actions, model-based RL proceeds by computing the optimal V or Q value function with respect to the estimated T and R. This could be done with

any of value iteration, policy iteration, or Q-value iteration. For this exercise, we will go with Q-value iteration.

Consider the following samples that the agent encountered.

s	a	s'	r
A	Clockwise	B	0.0
A	Clockwise	B	0.0
A	Clockwise	B	0.0
A	Clockwise	C	-10.0
A	Clockwise	C	-10.0
A	Counterclockwise	C	-8.0
A	Counterclockwise	C	-8.0
A	Counterclockwise	B	0.0
A	Counterclockwise	B	0.0
A	Counterclockwise	C	-8.0

s	a	s'	r
B	Clockwise	A	-3.0
B	Clockwise	A	-3.0
B	Clockwise	A	-3.0
B	Clockwise	A	-3.0
B	Clockwise	C	0.0
B	Counterclockwise	A	-10.0
B	Counterclockwise	A	-10.0
B	Counterclockwise	A	-10.0
B	Counterclockwise	A	-10.0
B	Counterclockwise	C	0.0

s	a	s'	r
C	Clockwise	A	0.0
C	Clockwise	B	6.0
C	Clockwise	B	6.0
C	Clockwise	A	0.0
C	Clockwise	A	0.0
C	Counterclockwise	B	-8.0
C	Counterclockwise	B	-8.0
C	Counterclockwise	B	-8.0
C	Counterclockwise	A	0.0
C	Counterclockwise	B	-8.0

We start by estimating the transition function, T(s,a,s') and reward function R(s,a,s') for this MDP.

5.1. Find the missing values (M,N,O and P) in the following table for T(s,a,s') and R(s,a,s').

Discount Factor, $\gamma = 0.5$

s	a	s'	T(s,a,s')	R(s,a,s')
A	Clockwise	B	M	N
A	Clockwise	C	O	P
A	Counterclockwise	B	0.400	0.000
A	Counterclockwise	C	0.600	-8.000
B	Clockwise	A	0.800	-3.000
B	Clockwise	C	0.200	0.000
B	Counterclockwise	A	0.800	-10.000
B	Counterclockwise	C	0.200	0.000
C	Clockwise	A	0.600	0.000
C	Clockwise	B	0.400	6.000
C	Counterclockwise	A	0.200	0.000
C	Counterclockwise	B	0.800	-8.000

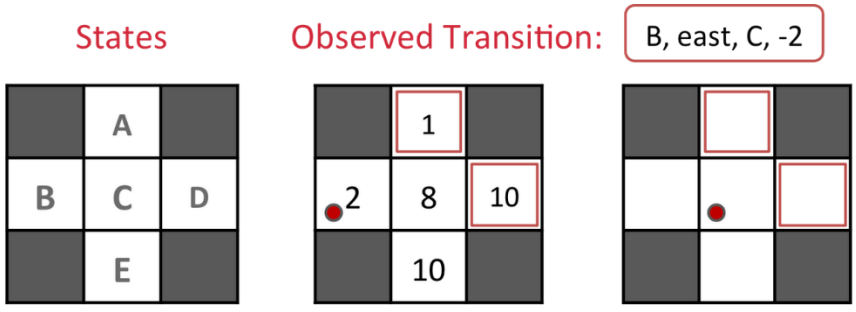
Now we will run Q-iteration using the estimated T and R functions. The values of $Q_k(s,a)$ are given in the table below.

	A	B	C
Clockwise	-4.24	-3.76	0.72
Counterclockwise	-4.56	-9.36	-7.76

5.2. Find the values for $Q_{k+1}(s,a)$.

Part D (Temporal Difference Learning and Model Free RL):

6. Consider the gridworld shown below. The left panel shows the name of each state A through E. The middle panel shows the current estimate of the value function V^{π} for each state. A transition is observed, that takes the agent from state B through taking action east into state C, and the agent receives a reward of -2. Assuming $\gamma = 1$, $\alpha = \frac{1}{2}$. What are the value estimates for each of the states after the TD learning update?



7. Consider an MDP with 3 states, A, B and C; and 2 actions Clockwise and Counterclockwise. We do not know the transition function or the reward function for the MDP, but instead, we are given with samples of what an agent actually experiences when it interacts with the environment (although, we do know that we do not remain in the same state after taking an action). In this problem, instead of first estimating the transition and reward functions, we will directly estimate the Q

function using Q-learning. Assume, the discount factor, $\gamma = \frac{1}{2}$ and the step size for Q-learning, $\alpha = \frac{1}{2}$.

Our current Q function $Q(s, a)$ is as follows:

	A	B	C
Clockwise	1.501	-0.451	2.73
Counterclockwise	3.153	-6.055	2.133

The agent encounters the following samples.

s	a	s'	r
A	Counterclockwise	C	8.0
C	Counterclockwise	A	0.0

Process the samples given above. Find the Q-values for each state-action pair after both samples have been accounted for.