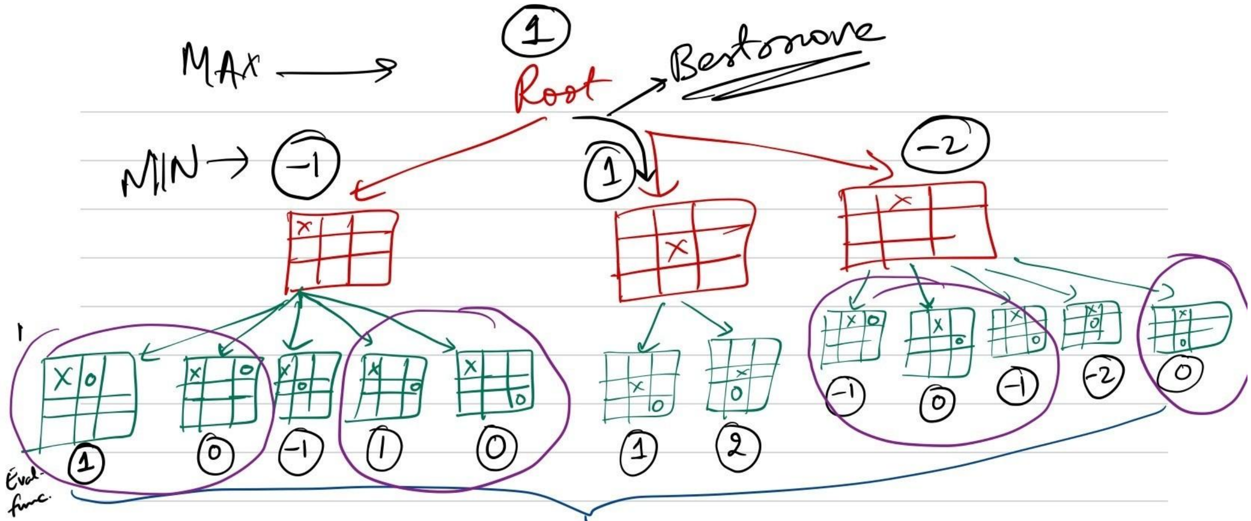
Q1



1. Drawing the above tree
2. Evaluation at depth 2
3. MIN - MAX values
4. Pruning the circled nodes: Optimal ordering is the center-X branch to be the first branch & then the other branches with the first child being less than the minimum of the center-X branch

Q2. a.

ANSWER

x>8

EXPLANATION

Depending on the value of x, there can be three different values for taking Action 1:

If $x < -8$, then Action 1 results in -8;

If $-8 \leq x \leq 11$ then Action 1 results in x;

If $x > 11$ then Action 1 results in 11.

Action 2 always results in a utility of 8

Hence Action 1 is optimal for Player 1 if $x > 8$

b.

ANSWER

x>9

EXPLANATION

Action 2 gives Player 1 a utility of 10, so the average of x and 11 must be greater than 10. $(x + 11)/2 > 10 \rightarrow x > 9$

c.

ANSWER

No

EXPLANATION

The minimax value can never be strictly greater than the expectimax value for the same tree because in minimax Player 2 always chooses the worst possible move for Player 1, while in expectimax, those same nodes average that value with other higher values. Thus, the utility at a node under expectimax is always at least as high as the utility of the same node under minimax.

Q3.

3.1 <u>Transition Matrix:</u>

For action a:

	I	II	III
I	0.2	0.8	0
II	0.8	0.2	0
III	0	0	0

For action b

	I	II	III
I	0.9	0	0.1
II	0	0.9	0.1
III	0	0	0

3.2 $A^{|S|} = 2^2 = 4$

Optimal Policy : $\left.\begin{array}{l}\text{if in 1 do b} \\ \text{if in 2 do a}\end{array}\right\} \rightarrow$ Intuition: Maximize Reward

3.3 $V_0^{I} = -1$ | $V_1^{I} = -1 + \max\left(\left((0.8)(-2) + (0.2)(-1)\right), \left((0.9)(-1) + (0.1)(0)\right)\right) = -1.9$

$V_0^{II} = -2$ | $V_2^{II} = -2 + \max\left(\left((0.8)(-1) + (0.2)(-2)\right), \left((0.9)(-2) + (0.1)(0)\right)\right) = -3.2$

3.4 → Initial Policy Eval [b, b]

$V^{I} = -1 + 0.9V^{I} + 0.1V^{III} \Rightarrow V^{I} = -10$

$V^{II} = -2 + 0.9V^{II} + 0.1V^{III} \Rightarrow V^{II} = -20$

$V^{III} = 0$

→ Evaluation

Can be done with argmax $(\Sigma T \times V)$ or $\boxed{R + \Sigma T \cdot V}$

- $R + \Sigma T.(V)$

$V^{I} = -19$ (for action a) $<$ -10 (for action b) [Dont change policy]

$V^{II} = -14$ (for action a) $>$ -20 (for action b) [Change policy to b]

- argmax $(\Sigma T \times V)$

$\pi_{V^{I}} = \text{argmax}\{\overset{b}{(0.9)^*(-10) + (0.1)(0)}, \overset{a}{(0.8)^*(-20) + (0.2)(-10)}\}$

$= \text{argmax}\{\overset{b}{-9}, \overset{a}{-18}\} = b$

$\pi_{V^{II}} = \text{argmax}\{\overset{b}{0.9^*(-20) + (0.1)(0)}, \overset{a}{(0.8)^*(-10) + (0.2)(-20)}\}$

$= \text{argmax}\{\overset{b}{-18}, \overset{a}{-12}\} = a$

→ Policy Eval [b, a]

$V^{I} = -1 + 0.9V^{I} + 0.1V^{III} = -10$

$V^{II} = -2 + 0.8V^{I} + 0.2V^{II} \quad V^{II} = -10/0.8 = -12.5$

$V^{III} = 0$

→ Evaluation

- $R + \Sigma T(V)$ [No Change]

$V^{I} = -1 + (0.8)(-12.5) + (0.2)(-10) = -13$ (for action a) $<$ -10 (for action b)

$V^{II} = -2 + (0.9)(-12.5) + (0.1)(0) = -13.25$ (for action b) $<$ -12.5 [for action a]

- argmax $(\Sigma T \times V)$ [No Change]

$\pi_{V^{I}} = \text{argmax}\{\overset{a}{(0.8)(-12.5) + (0.2)(-10)}, \overset{b}{(0.9)(-10) + (0.1)(0)}\}$

$= \text{argmax}\{\overset{a}{-12}, \overset{b}{-9}\} = b$

$\pi_{V^{II}} = \text{argmax}\{\overset{a}{(0.8)(-10) + (0.2)(-12.5)}, (0.9)(-12.5) + (0.1)(0)\} = \{$

$= \text{argmax}\{\overset{a}{-10.5}, \overset{b}{-11.25}\} = a$

No Change in Policy Terminate

3.5. <-10,-12.5,0>

3.6. Policy evaluation equations become unsolvable

$V(1) = -1 + 0.8*V(2) + 0.2*V(1)$

$V(2) = -2 + 0.8*V(1) + 0.2*V(2)$

Discounting helps in making these equations solvable and choice of the discount factor determines the policy. If a small discount factor is chosen then the future plays a negligible role as the agent gets greedy and will probably choose action b in state 2.

Q4. a.

- T(A,south,C) = 1
  The action south is taken twice from state A, and both times results in state C.
  2/2 = 1
- T(B,east,C) = 1
  The action east is taken twice from state B, and both times results in state C.
  2/2=1
- T(C,south,E) = 0.75
  The action south is taken four times from state C, and results in state E three times.¾=0.75
- T(C,south,D) = 0.25
  The action south is taken four times from state C, and results in state D one time.
  ¼ =0.25

b.

V(A) = 8, V(B) = -2, V(C) = 4, V(D) = -10, V(E) = 10

EXPLANATION

The estimated value of $\hat{V}^{\pi}(s)$ is equal to the average value achieved starting from that state.

$\hat{V}^{\pi}(A)$: Episodes 1 and 4 start from state A and both result in a utility of 8. $\frac{8+8}{2} = 8$

$\hat{V}^{\pi}(B)$: Episodes 2 and 3 start from state B. Episode 2 results in -12, while episode 3 results in 8. $\frac{8-12}{2} = -2$

$\hat{V}^{\pi}(C)$: State C is visited in every episode. The remaining rewards from C in episodes 1, 3, and 4 total 9, while the remaining rewards in episode 2 total -11.
$\frac{9+9+9-11}{4} = 4$

$\hat{V}^{\pi}(D)$: State D is only visited in episode 2 and has a remaining utility of -10.

$\hat{V}^{\pi}(E)$: State E is visited in episodes 1, 3, and 4 and has a remaining utility of 10 in each state. $\frac{10+10+10}{3} = 10$

Q5.

5.1

M=0.6

The clockwise action was taken 5 times from A, and went to B 3 times.

N=0

The transition (A,clockwise,B) happened 3 times and had reward 0 every time.

O=0.4

The clockwise action was taken 5 times from A, and went to C 2 times.

P=-10

Both of the occurrences of (A,clockwise,C) had reward -10.

5.2

Q(A, clockwise) = -4.984

EXPLANATION

$$V_k(B) = max(Q_k(B, clockwise), Q_k(B, counterclockwise)) = -3.76$$

$$V_k(C) = max(Q_k(C, clockwise), Q_k(C, counterclockwise)) = 0.72$$

$$Q(A, clockwise) = T(A, clockwise, B) \times (R(A, clockwise, B) + \gamma V_k(B)) +$$

$$T(A, clockwise, C) \times (R(A, clockwise, C) + \gamma V_k(C))$$

$$= .6 \times (0 + .5 \times -3.76) + .4 \times (-10 + .5 \times .72) = -4.984$$

Q(A, counterclockwise) = -5.336

EXPLANATION

$$Q(A, counterclockwise) = T(A, counterclockwise, B) \times (R(A, counterclockwise, B) + \gamma V_k(B)) +$$

$$T(A, counterclockwise, C) \times (R(A, counterclockwise, C) + \gamma V_k(C))$$

$$= .4 \times (0 + .5 \times -3.76) + .6 \times (-8 + .5 \times .72) = -5.336$$

Q(B, clockwise) = -4.024

EXPLANATION

$$V_k(A) = max(Q_k(A, clockwise), Q_k(A, counterclockwise)) = -4.24$$

$$Q(B, clockwise) = T(B, clockwise, A) \times (R(B, clockwise, A) + \gamma V_k(A)) +$$

$$T(B, clockwise, C) \times (R(B, clockwise, C) + \gamma V_k(C))$$

$$= .8 \times (-3 + .5 \times -4.24) + .2 \times (0 + .5 \times .72) = -4.024$$

Q(B, counterclockwise) = -9.624

EXPLANATION

$$Q(B, counterclockwise) = T(B, counterclockwise, A) \times (R(B, counterclockwise, A) + \gamma V_k(A)) +$$

$$T(B, counterclockwise, C) \times (R(B, counterclockwise, C) + \gamma V_k(C))$$

$$= .8 \times (-10 + .5 \times -4.24) + .2 \times (0 + .5 \times .72) = -9.624$$

Q(C, clockwise) = 0.376

EXPLANATION

$$Q(C, clockwise) = T(C, clockwise, A) \times (R(C, clockwise, A) + \gamma V_k(A)) +$$

$$T(C, clockwise, B) \times (R(C, clockwise, B) + \gamma V_k(B))$$

$$= .6 \times (0 + .5 \times -4.24) + .4 \times (6 + .5 \times -3.76) = .376$$

Q(C, counterclockwise) =-8.328

EXPLANATION

$$Q(C, counterclockwise) = T(C, counterclockwise, A) \times (R(C, counterclockwise, A) + \gamma V_k(A)) +$$

$$T(C, counterclockwise, B) \times (R(C, counterclockwise, B) + \gamma V_k(B))$$

$$= .2 \times (0 + .5 \times -4.24) + .8 \times (-8 + .5 \times -3.76) = -8.328$$

Q6.

EXPLANATION

The only value that gets updated is $\hat{V}^{\pi}(B)$, because the only transition observed starts in state B.

$\hat{V}^{\pi}(A) = 1$

$\hat{V}^{\pi}(B) = .5 * 2 + .5 * (-2 + 8) = 4$

$\hat{V}^{\pi}(C) = 8$

$\hat{V}^{\pi}(D) = 10$

$\hat{V}^{\pi}(E) = 10$

Q7.

EXPLANATION

For each s, a, s', r transition sample, you update the Q value function as follows:
$Q(s,a) = (1-\alpha) * Q(s,a) + \alpha * (R(s,a,s') + \gamma * max_{a'} Q(s',a'))$

First we update $Q(A, counterclockwise) = .5 \times 3.153 + .5 \times (8 + .5 \times 2.73) = 6.259$

Then we update $Q(C, counterclockwise) = .5 \times 2.133 + .5 \times (0 + .5 \times 6.259) \approx 2.631$

Note that we use the updated value of $Q(A, counterclockwise)$ in the second update.

Because there are only two samples, the other four values stay the same.