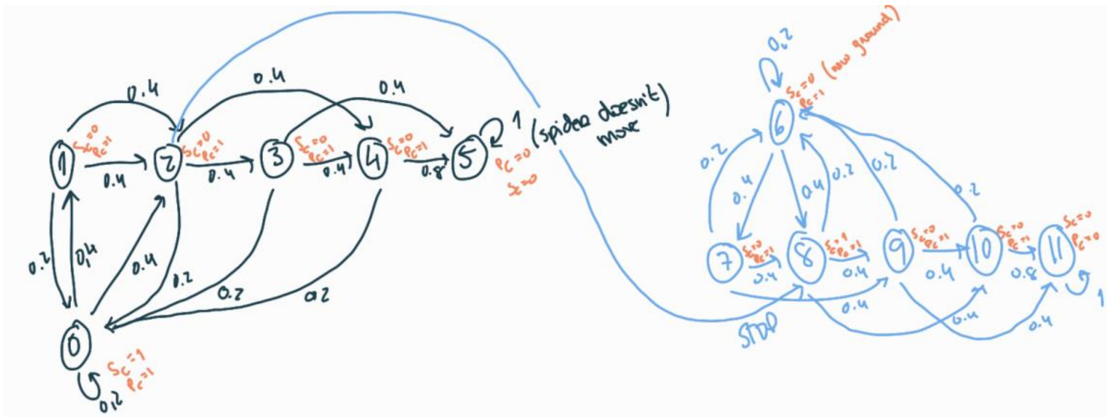
Homework 2

Group 30

a)



$$P_{play} = \begin{array}{cc} \begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \end{matrix} & \\ \begin{bmatrix} 0.2 & 0.4 & 0.4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2 & 0 & 0.4 & 0.4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2 & 0 & 0 & 0.4 & 0.4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2 & 0 & 0 & 0 & 0.4 & 0.4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2 & 0 & 0 & 0 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.4 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0 & 0.4 & 0.4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0 & 0 & 0.4 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0 & 0 & 0 & 0.4 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0.8 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} & \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \end{matrix} \end{array}$$

$$C = \begin{array}{cc} \begin{matrix} Play & Stop \end{matrix} & \\ \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 0 \end{bmatrix} & \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \end{matrix} \end{array}$$

- $S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$
- $A = \{P, S\}$
- Transition Matrices: P_{play}, P_{stop}
- Cost Matrix: C
- Discount Factor: $\gamma = 0.9$

$$P_{stop} = \begin{array}{cc} \begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \end{matrix} & \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} & \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \end{matrix} \end{array}$$

b)

The two deterministic policies at state 2 are choosing "play" or "stop"

→ $\pi(2) = \text{play}$

$P_\pi(2) = [0.2 \quad 0 \quad 0 \quad 0.4 \quad 0.4 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$

$c_\pi(2) = 1 \qquad \gamma = 0.9$

→ The cost-to-go for this policy would be:

$$J^\pi(2) = c_\pi(2) + \gamma \sum_{y \in \mathcal{X}} P_\pi(y \mid \underset{\substack{\downarrow \\ \text{state we currently are}}}{2}) \times J^\pi(y)$$

$$J^\pi(2) = 1 + 0.9\,(0.2\,J^\pi(0) + 0.4\,J^\pi(3) + 0.4\,J^\pi(4))$$

In order to find out the values that we are missing:

$J^\pi(5) = 0$, because according to the homework any action after reaching level 5 will have a cost of 0

$$J^\pi(4) = 1 + 0.9\,(0.2\,J^\pi(0) + 0.8\,\overbrace{J^\pi(5)}^{0}) =$$
$$= 1 + 0.9 \times 0.2\,J^\pi(0) = 1 + 0.18\,J^\pi(0)$$

$$J^\pi(3) = 1 + 0.9\,(0.2\,J^\pi(0) + 0.4\,J^\pi(4) + 0.4\,J^\pi(5)) =$$
$$= 1 + 0.9\,(0.2\,J^\pi(0) + 0.4\,(1 + 0.18\,J^\pi(0))) =$$
$$= 1 + 0.9\,(0.4 + 0.272\,J^\pi(0)) = 1.36 + 0.2448\,J^\pi(0)$$

$$J^\pi(2) = 1 + 0.9\,(0.2\,J^\pi(0) + 0.4\,(1.36 + 0.2448\,J^\pi(0)) + 0.4\,(1 + 0.18\,J^\pi(0))) =$$
$$= 1 + 0.9\,(0.944 + 0.36992\,J^\pi(0)) = 1.8496 + 0.3329\,J^\pi(0)$$

$$J^\pi(1) = 1 + 0.9\,(0.2\,J^\pi(0) + 0.4\,J^\pi(2) + 0.4\,J^\pi(3)) =$$
$$= 1 + 0.9\,(0.2\,J^\pi(0) + 0.4\,(1.8496 + 0.3329\,J^\pi(0)) + 0.4\,(1.36 + 0.2448\,J^\pi(0))) =$$
$$= 1 + 0.9\,(1.2838 + 0.4311\,J^\pi(0)) = 2.1554 + 0.3880\,J^\pi(0)$$

$$J^\pi(0) = 1 + 0.9\,(0.2\,J^\pi(0) + 0.4\,J^\pi(1) + 0.4\,J^\pi(2)) =$$
$$= 1 + 0.9\,(0.2\,J^\pi(0) + 0.4\,(2.1554 + 0.3880\,J^\pi(0)) + 0.4\,(1.8496 + 0.3329\,J^\pi(0))) =$$
$$= 1 + 0.9\,(1.602 + 0.4884\,J^\pi(0)) = 2.4418 + 0.4396\,J^\pi(0)$$

$$J^\pi(0) = 2.4418 + 0.4396\,J^\pi(0) \Leftrightarrow J^\pi(0) = 4.3572$$

$$J^\pi(2) = 1.8496 + 0.3329\,J^\pi(0) = 3.30 \rightarrow \text{cost for policy play}$$

Now for the action "stop"

$J^\pi(11) = 0$, because according to the homework any action after reaching level 11 will have a cost of 0

The cost for "stop" in state 2 is the discounted value of $J^\pi(8)$

Since we assume that in all other states the policy is optimal, we will do the "play" action, from state 8 forward

$$J^\pi(10) = 1 + 0.9(0.2 J^\pi(8) + 0.8 \overbrace{J^\pi(11)}^{0}) =$$

$$= 1 + 0.9 \times 0.2 J^\pi(8) = 1 + 0.18 J^\pi(8)$$

$$J^\pi(9) = 1 + 0.9(0.2 J^\pi(8) + 0.4 J^\pi(10) + 0.4 J^\pi(11)) =$$

$$= 1 + 0.9(0.2 J^\pi(8) + 0.4(1 + 0.18 J^\pi(8))) =$$

$$= 1 + 0.9(0.4 + 0.272 J^\pi(8)) = 1.36 + 0.2448 J^\pi(8)$$

$$J^\pi(8) = 1 + 0.9(0.2 J^\pi(8) + 0.4(1.36 + 0.2448 J^\pi(8)) + 0.4(1 + 0.18 J^\pi(8))) =$$

$$= 1 + 0.9(0.944 + 0.36992 J^\pi(8)) = 1.8496 + 0.3329 J^\pi(8)$$

$$J^\pi(8) = 1.8496 + 0.3329 J^\pi(8) \Leftrightarrow J^\pi(8) = 2.7726$$

Finally, the cost for stopping in state 2:

$$J^\pi(2) = c^\pi(2) + 0.9 J^\pi(8) = 0 + 0.9 \times 2.7726 \simeq \underline{\underline{2.4953}} \rightarrow \text{cost for policy stop}$$

$\rightarrow \pi = \text{stop}$

c)

The optimal policy for state 2 is to choose "stop" since this action has the lowest expected cost-to-go (2.4953 < 3.30)