MDP - Assignment - 3

NAGA NAANOTHAR
2021101128

(2)

Given,

Step cost = $R(I/A) = 0.04$

probability - direction of action = 0.7

- perpendicular to direction of action = 0.15

discount factor = $\gamma = 0.95$

delta = $\delta = 0.0001$

reward - goal state = +1

Penalty - red state = -1

Initialise: $U_0(I) = 0$

(let utility of wall = 0)

$$U_0 = \begin{bmatrix} 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

col(j) = 0, 1, 2 ; row(i) = 0, 1, 2, 3

Iteration 1:

$$U_{t+1}(I) = \max_A \left(R(I/A) + \gamma \cdot \sum_J p(J|I,A) * U_t(J) \right)$$

$$= R(I,A) + \max_A \left(\gamma \cdot \sum_J p(J|I,A) * U_t(J) \right)$$

(∵ $R(I,A)$ is constant)

$$\boxed{U_{t+1}(I) = R(I,A) + \gamma \cdot \max_A \left(\sum_J p(J|I,A) * U_t(J) \right)}$$

(∵ γ is constant)

Iteration-1:

$U_k[i][j]$ – represents a cell/state
k → row, col
in k^{th} iteration
$0 \le i < 4$
$0 \le j < 3$

$U_1[0][0]$

$= \max(-0.04) + \max \begin{pmatrix} 0.95^{*}(0.7^{*}0 + 0.15^{*}1 + 0.15^{*}0), & -up \\ 0.95^{*}(0.7^{*}1 + 0.15^{*}0 + 0.15^{*}0), & -left \\ 0.95^{*}(0.7^{*}0 + 0.15^{*}1 + 0.15^{*}0), & -down \\ 0.95^{*}(0.7^{*}0 + 0.15^{*}0 + 0.15^{*}0) & -right \end{pmatrix}$ (cyclic)

direction of action / perpendiculars

$= -0.04 + \max(0.1425, 0.665, 0.1425, 0)$

$= -0.04 + 0.665$

$= 0.625$

$U_1[0][1] = -0.04 + 0.95^{*}\max(0.7^{*}0)$

$U_1[0][1] = U_0[0][1]$ (absorption state – no movement)

$= +1$

$U_1[0][2] = U_0[0][2] = -1$ (")

$U_1[1][0] = -0.04 + 0.95^{*}\max \begin{pmatrix} 0.7^{*}0 + 0.15^{*}0 + 0.15^{*}0, \\ 0.7^{*}0 + 0.15^{*}0 + 0.15^{*}0, \\ 0.7^{*}0 + 0.15^{*}0 + 0.15^{*}0, \\ 0.7^{*}0 + 0.15^{*}0 + 0.15^{*}0 \end{pmatrix}$

$U_1[1][0] = -0.04$

$U_1[1][1] = -0.04 + 0.95^{*}\max \begin{pmatrix} 0.7^{*}1 + 0.15^{*}0 + 0.15^{*}0, \\ 0.7^{*}0 + 0.15^{*}1 + 0.15^{*}0, \\ 0.7^{*}0 + 0.15^{*}0 + 0.15^{*}0, \\ 0.7^{*}0 + 0.15^{*}1 + 0.15^{*}0 \end{pmatrix}$

$= -0.04 + 0.95^{*}(0.7)$

$= 0.625$

$$U_1[1][2] = -0.04 + 0.95 * \max \begin{pmatrix} 0.7*(-1) + 0.15*0 + 0.15*0, \\ 0.7*0 + 0.15*(-1) + 0.15*0, \\ 0.7*0 + 0.15*(0) + 0.15*0, \\ 0.7*0 + 0.15*(-1) + 0.15*0 \end{pmatrix}$$

$$= -0.04 + 0.95 * \max(-0.7, -0.15, 0, -0.15)$$

$$= -0.04 + 0$$

$$= -0.04$$

$$U_1[2][0] = -0.04 + 0.95 * \max \begin{pmatrix} 0.7*0 + 0.15*0 + 0.15*0, \\ 0.7*0 + 0.15*0 + 0.15*0, \\ 0.7*0 + 0.15*0 + 0.15*0, \\ 0.7*0 + 0.15*0 + 0.15*0 \end{pmatrix}$$

$$U_1[2][0] = -0.04$$

$$U_1[2][1] = (\text{wall } (\text{utility} = 0)) = 0$$

$$U_1[2][2] = -0.04 + 0.95 * \max \begin{pmatrix} 0.7*0 + 0.15*0 + 0.15*0, \\ 0.7*0 + 0.15*0 + 0.15*0, \\ 0.7*0 + 0.15*0 + 0.15*0, \\ 0.7*0 + 0.15*0 + 0.15*0 \end{pmatrix}$$

$$U_1[2][2] = -0.04$$

$$U_1[3][0] = -0.04 + 0.95 * \max \begin{pmatrix} 0.7*0 + 0.15*0 + 0.15*0, \\ 0.7*0 + 0.15*0 + 0.15*0, \\ 0.7*0 + 0.15*0 + 0.15*0, \\ 0.7*0 + 0.15*0 + 0.15*0 \end{pmatrix}$$

$$U_1[3][0] = -0.04$$

$$U_1[3][1] = -0.04 + 0.95 * \max \begin{pmatrix} 0.7*0 + 0.15*0 + 0.15*0, \\ 0.7*0 + 0.15*0 + 0.15*0, \\ 0.7*0 + 0.15*0 + 0.15*0, \\ 0.7*0 + 0.15*0 + 0.15*0 \end{pmatrix}$$

$$U_1[3][1] = -0.04$$

$$U_1[3][2] = -0.04 + 0.95 * \max \begin{pmatrix} 0.7*0 + 0.15*0 + 0.15*0, \\ 0.7*0 + 0.15*0 + 0.15*0, \\ 0.7*0 + 0.15*0 + 0.15*0, \\ 0.7*0 + 0.15*0 + 0.15*0 \end{pmatrix}$$

$$U_1[3][2] = -0.04$$

U_1 =

	0	1	2
0	0.625	1	-1
1	-0.04	0.625	-0.04
2	-0.04	0	-0.04
3	-0.04	-0.04	-0.04

Here as the agent did up but reached same cell $U_1[0][0]$

Iteration-2's

$$U_2[0][0] = -0.04 + 0.95 * \max\begin{pmatrix} 0.7*0.625 + 0.15*1 + 0.15*0, \\ 0.7*0.625 + 0.15*0.625 + 0.15*(-0.04), \\ 0.7*(-0.04) + 0.15*0.625 + 0.15*1, \\ 0.7*(0.625) + 0.15*0.625 + 0.15*(-0.04) \end{pmatrix}$$

$$= -0.04 + 0.95 * \max(0.5875, 0.78775, 0.21575, 0.52525)$$

$$= -0.04 + 0.95 * 0.7483625$$

$$= 0.7083625$$

$U_2[0][1] = U_1[0][1] = 1$

$U_2[0][2] = U_1[0][2] = -1$

$$U_2[1][0] = -0.04 + 0.95 * \max\begin{pmatrix} 0.7*0.625 + 0.15*0.625 + 0.15*(-0.04), \\ 0.7*0.625 + 0.15*0.625 + 0.15*(-0.04), \\ 0.7*(-0.04) + 0.15*0.625 + 0.15*(-0.04), \\ 0.7*(-0.04) + 0.15*0.625 + 0.15*(-0.04) \end{pmatrix}$$

$$= -0.04 + 0.95 * (0.52525)$$

$U_2[1][0] = 0.4589875$

$$U_2[1][1] = -0.04 + 0.95 * \max\begin{pmatrix} 0.7*1 + 0.15*(-0.04) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(1) + 0.15*(0) \\ 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(1) + 0.15*(0) \end{pmatrix}$$

$$= -0.04 + 0.95 * 0.688$$

$$= 0.6136$$

$$U_2[1][2] = -0.04 + 0.95 * \max\begin{pmatrix} 0.7*(-1) + 0.15*(0.625) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(-1) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(0.625) + 0.15*(-0.04) \\ 0.7*(0.625) + 0.15*(-1) + 0.15*(-0.04) \end{pmatrix}$$

$$= -0.04 + 0.95 * 0.2815$$

$$= 0.227425$$

$$U_2[2][0] = -0.04 + 0.95 * \max\begin{pmatrix} 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \end{pmatrix}$$

$$= -0.04 + 0.95*(-0.04)$$

$$= -0.078$$

$$((0.7 + 0.15 + 0.15)*(-0.04) = 1*-0.04 = -0.04)$$

$$U_2[2][1]\ (\text{wall}) = 0$$

$$U_2[2][2] = -0.04 + 0.95 * \max\begin{pmatrix} 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \end{pmatrix}$$

$$= -0.04 + 0.95*(-0.04)$$

$$= -0.078$$

$$U_2[3][0] = -0.04 + 0.95 * \max \begin{pmatrix} 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \end{pmatrix}$$

$$= -0.04 + 0.95 * (-0.04)$$

$$= -0.078$$

$$U_2[3][1] = -0.04 + 0.95 * \max \begin{pmatrix} 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \end{pmatrix}$$

$$= -0.04 + 0.95 * (-0.04)$$

$$= -0.078.$$

$$U_2[3][2] = -0.04 + 0.95 * \max \begin{pmatrix} 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \\ 0.7*(-0.04) + 0.15*(-0.04) + 0.15*(-0.04) \end{pmatrix}$$

$$= -0.04 + 0.95 * (-0.04)$$

$$= -0.078$$

$$\therefore U_2 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 0.7083625 & 1 & -1 \\ 0.4589875 & 0.6136 & 0.227425 \\ -0.078 & 0 & -0.078 \\ -0.078 & -0.078 & -0.078 \end{bmatrix} \end{array}$$

Thus, the values of U_1, U_2 match with the output of the code.