2019101056

Aryan Jain

Page :

Date : / /

MDL Assgn 2 Part 1

Ans 1 At either of the state (A/B/C/R) we can have 4 possible moves which are Up (U), Down (D), Left (L), Right (R).

To (s') ⇒	A				B				C				R			
From (s) ⇓	L	R	U	D	L	R	U	D	L	R	U	D	L	R	U	D
A	0	0.2	0.2	0	0	0.8	0	0	0	0	0.8	0	0	0	0	0
B	0.8	0	0	0	0.2	0	0.2	0	0	0	0	0	0	0	0.8	0
C	0	0	0	0.8	0	0	0	0	0	0.75	0	0.2	0	0.25	0	0

∴ Here each value represents $P(s'|s,a)$ i.e. probability of reaching a state s' from ~~initial~~ given initial state s and given action a.

Ans 2 Going from $A \to R$ has different possibilities like

$A \to B \to R$

$A \to C \to R$

$A \to B \to C \to R$ } these will add on some extra cost than above two.

$A \to C \to B \to R$

⋮

$\therefore$ Best path will be choosen from $A \to B \to R$ and $A \to C \to R$

Given $P(B \to R) = 0.8$ Step Cost $(B \to R) = -4$

$P(C \to R) = 0.25$ Step Cost $(C \to R) = -3$

$\therefore$ Expected utility $(B \to R) = 0.8 \times (-4) = -3.2$

Expected utility $(C \to R) = 0.25 \times (-3) = -0.75$

~~$\therefore$ Optimal solⁿs follows MEU principle~~

By MEU principle (Maximum Expected Utility) we can say $A \to C \to R$ is the optimal path

Ans.3 roll-number = 2019101056

$R = Arr[2019101056 \ \% \ 15] = Arr[1] = 9$

$\Rightarrow$ $\boxed{\text{Reward Value } (R) = 9.}$

To perform the value iteration algorithm we need to apply Bellman equation viz.

$$U_{t+1}(I) = \max_{A}\left[R(I,A) + \gamma \sum_{J} P(J \mid I, A)\, U_t(J) \right]$$

Also for $t=0$ $\quad U_0(A)=0 \quad U_0(B)=0$

$\quad U_0(C)=0 \quad U_0(D)=9$

0	9
0	0

$\boxed{\text{Let's call this as Destination } D}$

$t=0$ a.) For state 'A' : R, U are only possible actions.

1st iteration $U_1(A) = \max \begin{cases} R(A,R) + \gamma[P(B \mid A,R) \cdot U_0(B) + P(A \mid A,R) \cdot U_0(A)] \\ R(A,U) + \gamma[P(C, A, U) \cdot U_0(C) + P(A \mid A, U) \cdot U_0(A)] \end{cases}$

$U_1(A) = \max \begin{cases} -1 + 0.2[0.8 \times 0 + 0.2 \times 0] \\ -1 + 0.2[0.8 \times 0 + 0.2 \times 0] \end{cases}$

$\therefore \boxed{U_1(A) = -1}$

b.) For state 'B': L, U are 2 possible actions.

$U_1(B) = \max \begin{cases} R(B,L) + \gamma[P(A \mid B,L) \cdot U_0(A) + P(B \mid B,L) \cdot U_0(B)] \\ R(B,U) + \gamma[P(D \mid B,U) \cdot U_0(D) + P(B \mid B,U) \cdot U_0(B)] \end{cases}$

Destination

~~To calculate~~ $R(B,U)$: The reward associated with Action = 'U' and initial state = 'B' is also dependent on the destination state.

$\therefore R(B,U) = \sum_{J} R(B,U,J) \cdot P(J \mid B,U)$

$\therefore R(B,U) = -4 \times 0.8 + (-1) \times 0.2 = -3.4$

$$\therefore U_1(B) = \max \begin{cases} -1 + 0.2[0.8 \times 0 + 0.2 \times 0] = -1 \\ -3.4 + 0.2[0.8 \times 9 + 0.2 \times 0] = -1.96 \end{cases}$$

$\therefore \boxed{U_1(B) = -1}$

c.) For state 'C' : R, D are 2 possible action.

$$U_1(C) = \max \begin{cases} R(C,R) + \gamma[P(D|C,R).U_0(D) + P(C,C,R).U_0(C)] \\ R(C,D) + \gamma[P(A|C,R).U_0(A) + P(C,C,D).U_0(C)] \end{cases}$$

$R(C,R) = -3 \times 0.25 + (-1) \times 0.75 = -1.5$

$$U_1(C) = \max \begin{cases} -1.5 + 0.2[0.25 \times 9 + 0.75 \times 0] = -1.05 \\ -1 + 0.2[0.8 \times 0 + 0.2 \times 0] = -1 \end{cases}$$

$\therefore \boxed{U_1(C) = -1}$

$\therefore$ New utilities values are

-1	10
-1	-1

$\max_X [U_{t+1}(X) - U_t(X)]$ ← Max difference = 1 > 0.01 ($\delta = 0.01$)

⇒ ~~Re~~ Not converged

$t=1$

2nd iteration

By taking the formula from last iteration and substituting $U_0(X)$ with $U_1(X)$, we have :-

a.) For state 'A'

$$U_2(A) = \max \begin{cases} -1 + 0.2[0.8 \times (-1) + 0.2 \times (-1)] = \cancel{-1.02}\ -1.2 \\ -1 + 0.2[0.8 \times (-1) + 0.2 \times (-1)] = \cancel{-1.02}\ -1.2 \end{cases}$$

⇒ ~~$\boxed{U_2(A) = -1.02}$~~ $\boxed{U_2(A) = -1.2}$

HAWAMAHAL
Page :
Date : / /

b) For state 'B'

$$U_2(B) = \max \begin{cases} -1 + 0.2[0.8\times(-1) + 0.2\times(-1)] = \cancel{-1.02}\ -1.2 \\ -3.4 + 0.2[0.8\times 9 + 0.2\times(-1)] = -2.00 \end{cases}$$

$\therefore$ ~~$\boxed{U_2(B) = -1.02}$~~ $\boxed{U_2(B) = -1.2}$

c) For state 'c'

$$U_2(c) = \max \begin{cases} -1.5 + 0.2[0.25\times 9 + 0.75\times(-1)] = \cancel{0}\ -1.2 \\ -1 + 0.2[0.8\times(-1) + 0.2\times(-1)] = \cancel{-1.02}\ -1.2 \end{cases}$$

$\therefore$ ~~$\boxed{U_2(c) = 0}$~~ ~~$\boxed{U_2(c) = -1.02}$~~ $\boxed{U_2(c) = -1.2}$

$\therefore$ New utilities values are

~~-1.02~~	~~9~~
~~-1.02~~	~~-1.02~~

-1.2	9
-1.2	-1.2

Max. difference $= \cancel{0.02}\ 0.2 > 0.01$ $(\delta = 0.01)$

$\Rightarrow$ Not converged

~~∴ New utilities values are~~

~~-1.203~~	~~9~~
~~-1.204~~	~~-1.204~~

∴ Max. difference =

t=2
3rd iteration

a) For state 'A'

$$U_3(A) = \max \begin{cases} -1+0.2\,[0.8\times(-1.2)+0.2\times(-1.2)] \\ -1+0.2\,[0.8\times(-1.2)+0.2\times(-1.2)] \end{cases}$$

$$\therefore \boxed{U_2(A) = -1.24}$$

b.) For state 'B'

$$U_3(B) = \max \begin{cases} -1+0.2\,[0.8\times(-1.2)+0.2\times(-1.2)] \\ -3.4+0.2\,[0.8\times 9+0.2\times(-1.2)] \end{cases}$$

$$\therefore \boxed{U_2(B) = -1.24}$$

c) For state 'C'

$$U_3(C) \begin{cases} -1.5+0.2\,[0.25\times 9+0.75\times(-1.2)] \\ -1+0.2\,[0.8\times(-1.2)+0.2\times(-1.2)] \end{cases}$$

$$\boxed{U_3(C) = -1.23}$$

∴ New utilities values are

-1.23	9
-1.24	-1.24

∴ Max. difference = 0.04 > 0.01 $(\delta = 0.01)$
⇒ Not converged

t=3
4th iteration

a.) For state 'A'

$$U_4(A) = \max \begin{cases} -1+0.2\,[0.8\times(-1.24)+0.2\times(-1.24)] \\ -1+0.2\,[0.8\times(-1.23)+0.2\times(-1.24)] \end{cases}$$

$$\therefore \boxed{U_4(A) = -1.2464}$$

Page:
Date: / /

~~b) For state 'B'~~

$$\cancel{U_4(B) = \max \begin{cases} -1 + 0.2\,[0.25\times 9 + 0.75\times(-1.24)] \\ -3.4 + 0.2\,[0.8\times 9 + 0.2\times(-1.24)] \end{cases}}$$

$\therefore$ $\cancel{\boxed{U_4(B) = -0.736}}$

b.) For state 'B'

$$U_4(B) = \max \begin{cases} -1 + 0.2\,[0.8\times(-1.24) + 0.2\times(-1.24)] \\ -3.4 + 0.2\,[0.8\times 9 + 0.2\times(-1.24)] \end{cases}$$

$$\boxed{U_4(B) = -1.248}$$

c.) For state 'c'

$$U_4(C) = \max \begin{cases} -1.5 + 0.2\,[0.25\times 9 + 0.75\times(-1.23)] \\ -1 + 0.2\,[0.8\times(-1.24) + 0.2\times(-1.23)] \end{cases}$$

$\therefore$ $\boxed{U_4(C) = -1.2345}$

$\therefore$ New utility values are

-1.2345	9
-1.2464	-1.248

$\therefore$ Max. difference $= 0.0064 < 0.01$

$\Rightarrow$ Converged (At 4^{th} iteration)

Ans 4 Now, we run the algorithm one last time to determine the policy.

~~Utilities were :-~~

$$\therefore U_5(A) = \max \begin{cases} -1 + 0.2\,[0.8\times(-1.248) + 0.2\times(-1.2464)] & ① \leftarrow \text{Right} \\ -1 + 0.2\,[0.8\times(-1.2345) + 0.2\times(-1.2464)] & ② \leftarrow \text{Up} \end{cases}$$

Action

~~Clearly $U_5(A)$ UP > Right~~

Clearly ② > ① (By comparison) $\Rightarrow$ $\boxed{\text{From A you should take an UP action}}$

① = -1.249536 ② = -1.247376

$U_5(B) \rightarrow$ Calculating it will be useless.

$$U_5(c) \begin{cases} -1.5 + 0.2[0.25 \times 9 + 0.75 \times (-1.2345)] \\ -1 + 0.2[0.8 \times (-1.2464) + 0.2 \times (-1.2345)] \end{cases}$$

$$U_5(c) \begin{cases} -1.235175 \leftarrow \text{Right} \\ -1.248804 \leftarrow \text{Down} \end{cases}$$

$U_5(c) = -1.235175$

$\Rightarrow$ From c u should move Right

$\Rightarrow$ $A \rightarrow c \rightarrow R$ is optimal path
(Which we also guessed in Ques. 2)

Ans. 5 As the reward values increases the optimal path $A \rightarrow C \rightarrow R$ starts changing to $A \rightarrow B \rightarrow R$. This is because now we have a high reward so the algorithm will tend to more risk seeking methodology.
(Means more risk can be taken).