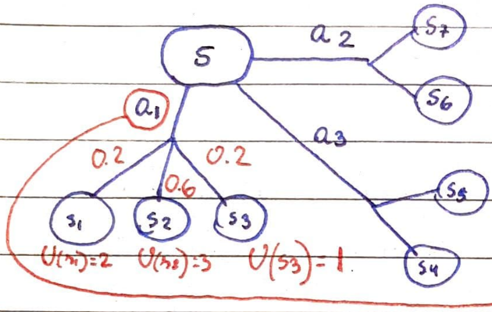
SUMMARY

Expected Utility of X; where X is a random variables that can take multiple values K

$$E(X) = \sum_{K} K \, Pr(X = K)$$

→ we don't know to which state S_a an action a will lead us → Expected utility



$$E(u(a)) = \sum_{s' \in S_a} u(s') \cdot Pr(S_a = s')$$

Example → expected utility of action a_1:

$$E(u(a_1)) = \underset{2}{u(s_1)} \times \underset{0.2}{Pr(S_a = s_1)} + \underset{3}{u(s_2)} \times \underset{0.6}{Pr(S_a = s_2)} + \ldots$$

$$\ldots \underset{1}{u(s_3)} \times \underset{0.2}{Pr(S_a = s_3)}$$

- We repeat the process for every a And we choose the a with highest expected utility....

Alternative methods to choose an action:

maximin

$$a^* = \arg\max_{a \in A} \left\{ \min_{s' \in S_a} u(s') \right\}$$

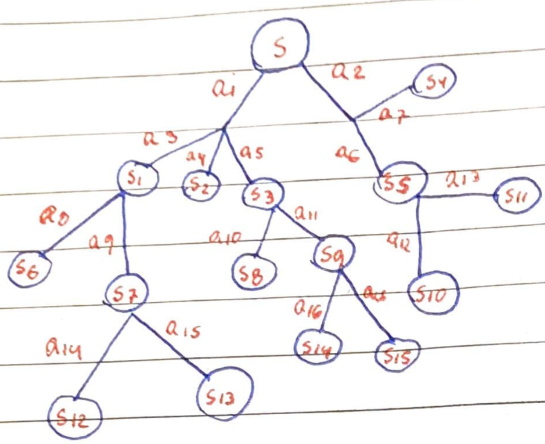
Among the minimum utilities of each action, we choose the action with the highest minimum utility

maximax

$$a^* = \arg\max_{a \in A} \left\{ \max_{s' \in S_a} u(s') \right\}$$

Among the maximum utilities of each action, we choose the action with the highest maximum utility

Sequential decision making



We cannot use a simple expected utility for a single action ....

As This action might be good in the short term ......

But what about in the long term with a sequence of actions?

SEQUENTIAL DECISION MAKING (SUMMARY II)

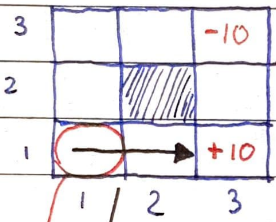
Basic framework: Markov decision process (MDP)

- Set of states $s \in S$
- Set of Actions $A(s)$ in each state
- transition model $P(s' \mid s, a)$

s' → next state; s → current state; a → current action

- A reward function

- SOLUTION of MDPs → $\underbrace{\text{a policy } \pi}_{\text{we would prefer the optimum policy } \pi^*}$



States: (1,1), (2,1), (3,1), (1,2),
(1,3), (2,3), (3,3), (3,2)

policy π → is the optimum policy → π^*

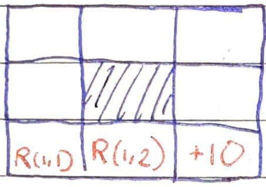
Action: Stop, North, South, East, West

Example transition model:

$P(\underset{\text{future state}}{(1,2)} \mid \underset{\text{current state}}{(1,1)}, \underset{\text{indicated action}}{\text{North}}) = 0.8$

SUMMARY III

Each state has a Rexard assigned To it → Rexard as cost of moving is a frequent approach



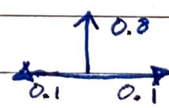
Terminal state final Rexard

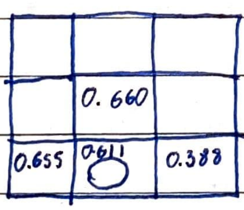
(1,1) → Agent moves to (1,2) → Receives a Rexard (which can be negative).

As with Previous (non-sequential) decision making processes, we need to know the utilities. Then, calculate the expected utility of each action. Hoxeber, This time we have To include The future actions we might take.....

→ Image we have the right utilities of each state. Then To decide on action we need to calculate the expected utility →

Transition model





$E(u(north)) = 0.8 \times 0.660 + 0.1 \times 0.655 + 0.1 \times 0.388$

$E(u(east)) = 0.8 \times 655 + 0.1 \times 0.660 + \underbrace{0.1 \times 0.611}$

The agents stays in the same place

$E(u(west)) = 0.8 \times 0.388 + 0.1 \times 0.660 + 0.1 \times 0.611$

Then, we choose the action with highest utility

SUMMARY IV

How to calculate the utilies that make the agent follow an optimal policy π^*

→ OPTIMAL POLICY (given the right utilities) $\pi^*(s) = \underset{a \in A(s)}{\arg\max} \sum_{s'} P(s'|s,a) \underbrace{U^{\pi^*}(s')}$

We need to calculate this

→ We can use the $\boxed{\text{Bellman equation}}$:

$$U(s) = R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a)\, U(s')$$

discount factor

We will have to solve as many bellman equations as states in the map...

This is hard to solve in big scenarios

Example:

solve this is really hard

$$\begin{cases} U(1,1) = R(1,1) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a)\, U(s') \\ U(1,2) = R(1,2) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a)\, U(s') \\ U(1,3) = R(1,3) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a)\, U(s') \\ \vdots \end{cases}$$

SUMMARY V

We can get the Utilities with a computational approach called / labelled as

VALUE ITERATION

$$\rightarrow \underbrace{U_{i+1}(s)}_{\text{Utility in the next iteration}} \leftarrow R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s' \mid s, a)\, U_i(s')$$

approximation

Repeat it several times until the value of U does not change

1

a) Exnected Utility

Non - sequential decision making

a_1 = office a_2 = going out

P(work | office) = 0.5

P(distracted | office) = 0.3

P(colleague | office) = 0.2

P(coffe | out) = 0.95

P(spill | out) = 0.05

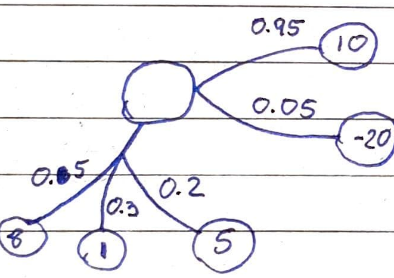
U(work) = 8

U(distracted) = 1

U(colleague) = 5

U(coffe) = 10

U(spill) = -20



• E(U(office)) =

0.5 × 8 + 0.3 × 1 + 0.2 × 5 = 5.3

• E(U(coffe)) =

0.95 × 10 + 0.05 × -20 = 8.5

E(U(coffe)) > E(U(office)) → we choose coffe

b) Which action should we choose

coffe expected utility is higher

c) Maximax and maximin

maximin → $a^* \underset{a \in A}{\operatorname{argmax}} \{ \underset{s' \in S_a}{\min} u(s') \}$

→ Among the minimum utilities of each action, we choose the action with the highest minimum utility maximax

$$a^* = \arg\max_{a \in A} \{ \max_{s' \in S_a} U(s') \}$$

Among the maximum utilities, we choose the action with the highest maximum utilities

$\boxed{\text{maximin:}}$

$$\arg\max_{s \in \{office, out} \{ \min^{office}(8, 1, 5), \min^{out}(10, -20) \}$$

min office

min out

$$\arg\max_{s \in \{office, out} \{ \min office (1), \min out (-20) \}$$

maximin chooses office

maximax

$$\arg\max_{s \in \{office, out\}} \{ \max^{office}(8, 1, 5), \max^{out}(10, -20) \}$$

max office

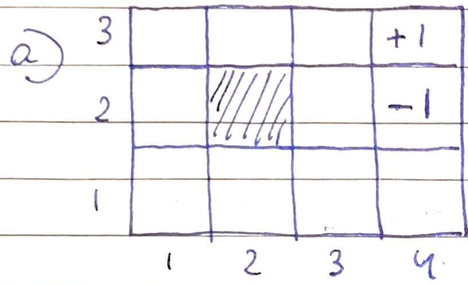
max out

$$\arg\max_{s \in \{office, out} \{ \max office (8), \max out (10) \}$$

maximax chooses out

Exercise 2

a)



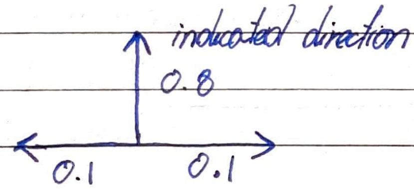
States: (1,1), (1,2), (1,3), (2,1), (2,3), (3,1), (3,2), (3,3), (4,1) (4,2), (4,3)

Initial state: (1,1)

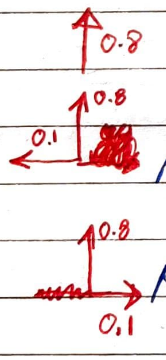
Actions: up, down, left, right

$$R(s) \begin{cases} 1 & \text{for } s = (4,3) \rightarrow \text{terminal state} \\ -1 & \text{for } s = (4,2) \rightarrow \text{terminal state} \\ -0.04 & \text{for } s \neq (4,3) \text{ and } (4,2) \end{cases}$$

Transition model



Example:



$P((1,2) \mid (1,1), Up) = 0.8$

$P((1,1) \mid (1,1), Up) = 0.1$

$P((2,1) \mid (1,1), Up) = 0.1$

b) Normal Bellman equation → $U(s) =$

$$\rightarrow U(s) = R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s' \mid s, a)\, U(s')$$

Deterministic Bellman (No Probability)

$$\rightarrow U(s) = R(s) + \gamma \max_{a \in A(s)} U(s')$$

we know which state is going to be the result of every of the agents actions

c) Deterministic Value Iteration

$$\underbrace{U_{i+1}(s) \leftarrow R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a) \times U(s')}_{\text{stochastic version}}$$

$$\underbrace{U_{i+1}(s) \leftarrow R(s) + \gamma \max_{a \in A(s)} U(s')}_{\text{deterministic}}$$

→ $\gamma = 1$ and $U(s) = 0$

1st iteration:

3	-0.04	-0.04	0.96	+1
2	-0.04	//////	-0.04	-1
1	-0.04	-0.04	-0.04	-0.04
	1	2	3	4

Examples

$$\underbrace{U_{i+1}(1,1) = -0.04 + 1 \times \max}_{-0.04} \begin{cases} \text{Up: } 0 \\ \text{East: } 0 \\ \text{South: } 0 \\ \text{West: } 0 \end{cases}$$

$$\underbrace{U_{i+1}(3,3) = -0.04 + 1 \times \max}_{0.96} \begin{cases} \text{Up: } 0 \\ \text{East: } 1 \\ \text{South: } 0 \\ \text{West: } 0 \end{cases}$$

2nd Round / Iteration

3	-0.08	0.92	0.96	+1
2	-0.08	//////	0.92	-1
1	-0.08	-0.08	-0.08	-0.08
	1	2	3	4

Examples

$$\underbrace{U_{i+1}(2,3) = -0.04 + 1 \times \max \begin{cases} \text{Up: } 0.04 \\ \text{East: } 0.96 \\ \text{West: } -0.04 \\ \text{South: } -0.04 \end{cases}}_{0.92}$$

2nd Continuation

5th iteration

0.88	0.92	0.96	
0.84		0.92	
0.8	0.84	0.8	

d) Calculate the optimal policy

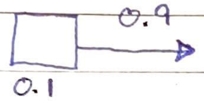
→ Remember: we have to calculate the expected u to know the best policy

However, we are in a deterministic environment. So, is just the utility

0.88 →	0.92 →	0.96 →	+1
0.84 ↑		0.92 ↑	0.92
0.8 ↑ →	0.84 →	0.8 ↑	0.84 ←

} Best policy

3



Non deterministic Bellman:

$$U_{i+1}(s) = R(s) + \underset{0.9}{\gamma} \max_{a \in A(s)} \sum_{s'} P(s'/s,a)\, U(s')$$

3	0	0	0	1
2	0	//////	0	-1
1	0	0	0	0
	1	2	3	4

Example

$U_{i+1}(3,3) \Rightarrow$

$$\underbrace{-0.04 + 0.9 \max \begin{cases} Up: 0 \times 0.9 + 0 \times 0.1 = 0 \\ Left: 0 \times 0.9 + 0 \times 0.1 = 0 \\ Right = 1 \times 0.9 + 0 \times 0.1 = 0.9 \\ Down = 0 \times 0.9 + 0 \times 0.1 = 0 \end{cases}}_{0.77}$$

Repeat this for all states. Then, in the next iteration repeat this for every state. Repeat this process until U does not change much.

3	0.87	0.91	0.96	+1
2	0.82	//////	0.91	-1
1	0.78	0.82	0.87	0.82
	1	2	3	4

} final Utilities

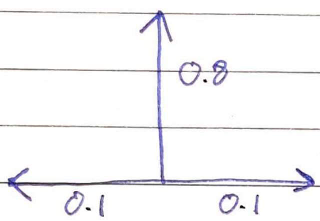
To know the best policy / action.....
We need to calculate the expected utility and choose the maximum one

$$\pi^*(s) = \arg\max_{a \in A} \sum_{s'} P(s'/s,a)\, U^*(s')$$

3 Example:

$$\pi^*(3,2) = \arg\max \begin{cases} \text{Up: } 0.96\times0.9+0.1\times0.91 = 0.955 \\ \text{Down: } 0.87\times0.9+0.1\times0.91 = 0.874 \\ \text{Left: } 0.91\times0.9+0.1\times0.1 = 0.91 \\ \text{Right: } -1\times0.9+0.1\times0.91 = -0.809 \end{cases}$$

4 → New transition model



a) 1 Value iteration for (3,1)

$$U_{i+1}(s) = R(s) + \gamma \underset{s\in A(s)}{\arg\max} \sum_{s'} P(s'|a,s)\cdot U(s')$$

→ $U_{i+1}(3,3) = -0.04 + \gamma \max\{0.8\times0.66\ldots\ldots$

3	0.812	0.868	0.918	1
2	0.762		0.660	-1
1	0.705	0.655	0.611	0.388
	1	2	3	4

$$\max \begin{cases} \text{Up: } 0.8\times0.660+0.1\times0.388+0.655\times0.1 = 0.6323 \\ \text{Down: } 0.8\times0.611+0.1\times0.655+0.1\times0.388 = 0.593 \\ \boxed{\text{Left: } 0.8\times0.655+0.1\times0.611+0.1\times0.660 = 0.6511} \\ \text{Right: } 0.8\times0.388+0.1\times0.611+0.1\times0.660 = 0.4375 \end{cases}$$

$U_{i+1}(3,3) = -0.04 + 1\times0.6511 = \boxed{0.6111}$

The value has not changed → convergence

4

b) Expected Utility in (1,3)

Up: $0.8 \times 0.660 + 0.1 \times 0.655 + 0.1 \times 0.388 = 0.6323$
Down: $0.8 \times 0.611 + 0.1 \times 0.655 + 0.1 \times 0.388 = 0.5931$
Left: $0.8 \times 0.655 + 0.1 \times 0.660 + 0.1 \times 0.611 = \boxed{0.6511}$
Righ: $0.8 \times 0.388 + 0.1 \times 0.660 + 0.1 \times 0.611 = 0.4375$

c) maximax and maximin

- Example in (3,1)

maximax

Up: max is 0.660
Down: max is 0.655
Left: max is 0.660
Right: max is 0.660

among these ones the max is 0.660. We will need to randomly choose one

maximin

Up: min is 0.388
Down: min is 0.388
Left: min is 0.611
Right: min is 0.388

among these ones the max is 0.611. therefore the agent will go left