

6CCS3AIN & 7CCSMAIN, 2018, Tutorial 04 Answers (Version 1.2)

1. (a) We have a decision between staying in the office and going out. Both options will lead to one or more states.

Staying in the office means that I will either work, get distracted, or talk with colleague. These states have the following utilities:

$$\begin{aligned}U(\text{work}) &= 8 \\U(\text{distracted}) &= 1 \\U(\text{colleague}) &= 5\end{aligned}$$

and the probabilities of these happening, given I stay in the office are:

$$\begin{aligned}P(\text{work}|\text{office}) &= 0.5 \\P(\text{distracted}|\text{office}) &= 0.3 \\P(\text{colleague}|\text{office}) &= 0.2\end{aligned}$$

Thus the expected utility of staying in the office is:

$$\begin{aligned}EU(\text{office}) &= 0.5 \times 8 + 0.3 \times 1 + 0.2 \times 5 \\&= 5.3\end{aligned}$$

Going out can result in the states coffee and spill, with utilities:

$$\begin{aligned}U(\text{coffee}) &= 10 \\U(\text{spill}) &= -20\end{aligned}$$

and the relevant probabilities are:

$$\begin{aligned}P(\text{coffee}|\text{out}) &= 0.95 \\P(\text{spill}|\text{out}) &= 0.05\end{aligned}$$

Thus $EU(\text{out}) = 8.5$.

- (b) The principle of maximum expected utility says to pick the option with the greatest expected utility, in this case that is to go out.
- (c) The maximin principle says to pick the options based on the utility of their worst outcomes, and to pick the option with the largest utility for its worst outcome.

Formally we have:

$$\begin{aligned}a &= \arg \max_{s \in \{\text{office}, \text{out}\}} \{ \min(U(s)) \} \\&= \arg \max_{s \in \{\text{office}, \text{out}\}} \{ \min^{\text{office}}(8, 1, 5), \min^{\text{out}}(10, -20) \} \\&= \arg \max_{s \in \{\text{office}, \text{out}\}} \{ 1^{\text{office}}, -20^{\text{out}} \} \\&= \text{office}\end{aligned}$$

so the maximin principle would say I should stay in the office.

The maximax principle says to pick the option based on the best outcome of each action, and to pick the one with the largest utility for the best outcome, in other words:

$$a = \arg \max_{s \in \{\text{office}, \text{out}\}} \{ \max(U(s)) \}$$

which will pick out.

2. The grid world is:

3				+1
2				-1
1				
	1	2	3	4

(a) The formal description of this as an MDP is the following:

- States: $(1, 1), (1, 2), (1, 3), \dots (2, 1), (2, 3), \dots (4, 3)$
Note that the location $(2, 2)$, where there is an obstacle, does not correspond to a state.
- Initial state: $(1, 1)$.
- Actions: Up, Down, Left Right for all states.
- Reward:

$$R(s) = \begin{cases} 1 & \text{for } s = (4, 3) \\ -1 & \text{for } s = (4, 2) \\ -0.04 & \text{otherwise} \end{cases}$$

- Transition model, $P(s'|s, a)$:

$$\begin{aligned} P((1, 2)|(1, 1), Up) &= 0.8 \\ P((1, 1)|(1, 1), Up) &= 0.1 \\ P((2, 1)|(1, 1), Up) &= 0.1 \\ P((1, 1)|(1, 1), Down) &= 0.9 \\ &\vdots \end{aligned}$$

(b) For a deterministic model, we no longer have to worry about the expected utility of an action, because we know it succeeds, and we have:

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

where s' is the state that results from action a .

(c) We will assume that $\gamma = 1$, and set $U(s)$ to 0 initially. Then, after the first round of value iteration using the above formula, we get:

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

After a second round we get:

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

so after 3 more rounds we will end up with:

3	0.88	0.92	0.96	+1
2	0.84		0.92	-1
1	0.8	0.84	0.88	0.84
	1	2	3	4

when the values will change no more.

See the spreadsheet (on KEATS) for detail on the calculations.

- (d) At each state, the optimum policy is to pick the action with the highest expected utility. Since actions are deterministic in this case, that equates to deciding to move to the state with the highest utility (make sure you understand why this is the case). As a result, we have:

3	→	→	→	
2	↑		↑	
1	↗	→	↑	←
	1	2	3	4

Note that we assume the agent does not move once it has reached (4, 3) or (4, 2) as in the lecture.

3. (a) My spreadsheet tells me that when the values have converged, we have (rounding to two decimal places):

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

For $\gamma = 1$.

The calculation for the first iteration, for (3, 3) is:

$$U(s) = -0.04 + \gamma \max \begin{pmatrix} (1 \times 0) & Up \\ (0.9 \times 0 + 0.1 \times 0) & Down \\ (0.9 \times 0 + 0.1 \times 0) & Left \\ (0.9 \times 1 + 0.1 \times 0) & Right \end{pmatrix}$$

and so on for the rest of the states.

See the spreadsheet (on KEATS) for the rest of intermediate calculations.

- (b) The computation of the best action in each state parallels the calculation above. For (3, 3) we want to establish the maximum of the expected values for the 4 possible actions:

$$\begin{pmatrix} (1 \times 0.96) & Up \\ (0.9 \times 0.91 + 0.1 \times 0.96) & Down \\ (0.9 \times 0.91 + 0.1 \times 0.96) & Left \\ (0.9 \times 1 + 0.1 \times 0.96) & Right \end{pmatrix}$$

so Right will be the best action.

Repeating this for all the states, we quickly find that the optimum policy is the same as in the deterministic case (though the expected utility of the policy will be lower since the agent will spend more time in lower utility states, and will take more steps to get to the goal).

What that suggests is that for a motion model like the one we have here, the deterministic version of value iteration is a pretty good approximation of the non-deterministic version (though the difference in the utilities is obvious from the figures).

4. (a)

$$U((3,1)) = -0.04 + \gamma \max \begin{pmatrix} 0.8 \times 0.660 + 0.1 \times 0.655 + 0.1 \times 0.388 & Up \\ 0.8 \times 0.611 + 0.1 \times 0.655 + 0.1 \times 0.388 & Down \\ 0.8 \times 0.655 + 0.1 \times 0.660 + 0.1 \times 0.611 & Left \\ 0.8 \times 0.388 + 0.1 \times 0.660 + 0.1 \times 0.611 & Right \end{pmatrix}$$

which reduces to:

$$U((3,1)) = -0.04 + \gamma \max \begin{pmatrix} 0.6323 & Up \\ 0.5931 & Down \\ 0.6511 & Left \\ 0.4375 & Right \end{pmatrix}$$

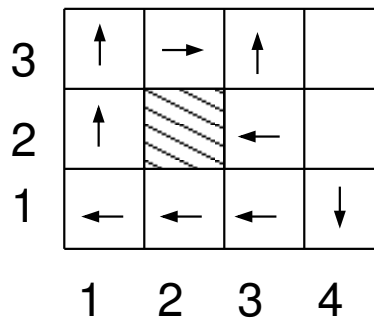
which comes to 0.611.

This is the same as the utility of the state. Which is exactly what we would expect from the Bellman equation once we have the state utility that is the one under the optimum policy (which is what happens when value iteration terminates).

- (b) The numbers in the equation above are the expected utilities of the actions. So the action with the maximum expected utility is Left.

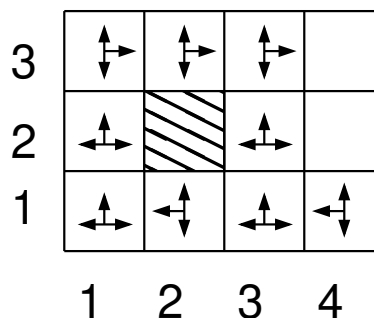
Note that this is not the same as moving towards the state with the maximum utility.

- (c) i. For the maximin policy, we look at the worst outcome of each action, and pick the action which maximises this worst outcome. We get:



The policy makes sure that in every state the agent has no chance of entering (4,2) while making sure that it will eventually get to (4,3).

- ii. For the maximax policy, we look at the best outcome of each action and pick the action which maximises this. Because every action has three possible outcomes, there are always three actions which can lead to the maximum utility. Since maximax ignores the probability of outcomes, it treats all of these actions as equally good:



The only action that is not selected by maximax is the one that points away from the state with the maximum utility.

Of course, an agent can't pick three actions, so any agent that wanted to employ maximax in this environment would have to pick one of the three actions.

5. There is no question 5.