Grundlagen der Künstlichen Intelligenz

Exercise 11: Making Complex Decisions — Winter Semester 2022/23

Eivind Meyer

Problem 11.1:

(Taken from **Russell and Norvig 2010**) Suppose that the Roomba is situated in the 4x3 environment shown in Figure 1(a). Beginning in the start state, it must choose an action at each step. The interaction with the enviroment terminates when the Roomba reaches one of the terminal states shown in Figure 1(a): the state where the charger is situated is marked +1 and the state where the stairs are located is marked -1. The actions for each state are *Right, Left, Up* and *Down*. The particular model of the stochastic motion that we adopt is illustrated in Figure 1(b). The reward is given as -0.04 for all states except the terminal states and the value of γ is given as 1.

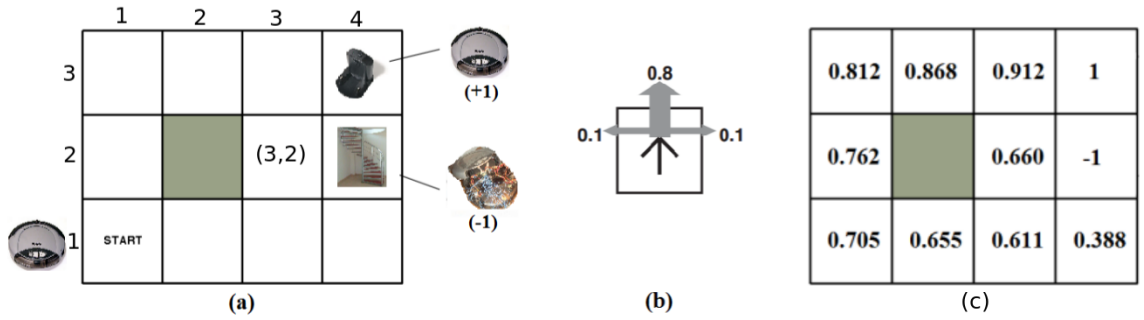


Figure 1: Problem 11.1

Problem 11.1.1: Assuming the transition probability as deterministic and the discount factor as 1 find the value of all states.

Problem 11.1.2: Show the corresponding policy.

Problem 11.1.3: Assuming the transition probability as stochastic and calculate the value of $U(3,3)$ using the value iteration algorithm for 2 iterations. Assume that all unknown initial utilities are zero and $U^1(1,3) = -0.04$, $U^1(2,3) = -0.04$ and $U^1(3,2) = -0.04$.

Additional task: try to compute the values of $U(1,3)$, $U(2,3)$, $U(3,2)$ *for 2 iterations using the value iteration algorithm. Assume that* $U^1(1,2) = -0.04$ *holds.*

Problem 11.1.4: Compute the optimal policy of state $(3,1)$ after convergence. The utilities after convergence are shown in Fig. 1 (c).

Problem 11.2:

(Taken from **Andrew W. Moore**) Assume that you run a startup company. In every decision period, you must choose between Saving money (S) or Advertising (A). If you advertise, you may become famous (f) (50%) but because of spending money you may become poor (p). If you save money, you may become rich (r) with probability 50% but you may become also unknown (u) because you don't advertise.

Problem 11.1:

(Taken from **Russell and Norvig 2010**) Suppose that the Roomba is situated in the 4x3 environment shown in Figure 1(a). Beginning in the start state, it must choose an action at each step. The interaction with the enviroment terminates when the Roomba reaches one of the terminal states shown in Figure 1(a): the state where the charger is situated is marked +1 and the state where the stairs are located is marked -1. The actions for each state are *Right*, *Left*, *Up* and *Down*. The particular model of the stochastic motion that we adopt is illustrated in Figure 1(b). The reward is given as -0.04 for all states except the terminal states and the value of γ is given as 1.

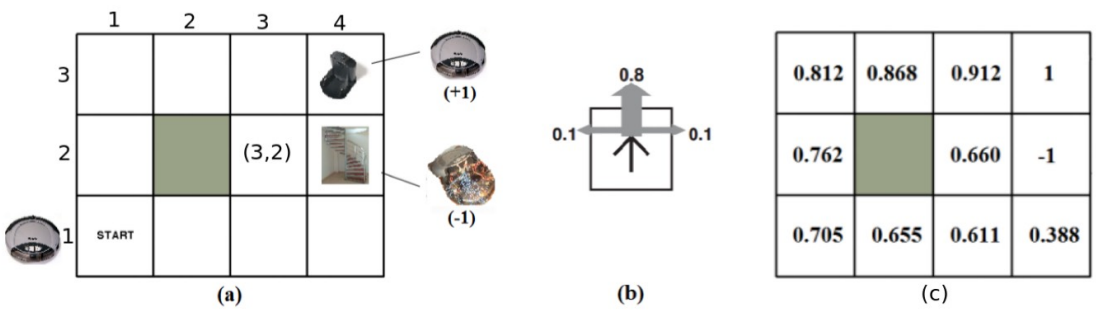


Figure 1: Problem 11.1

Problem 11.1.1: Assuming the transition probability as deterministic and the discount factor as 1 find the value of all states.

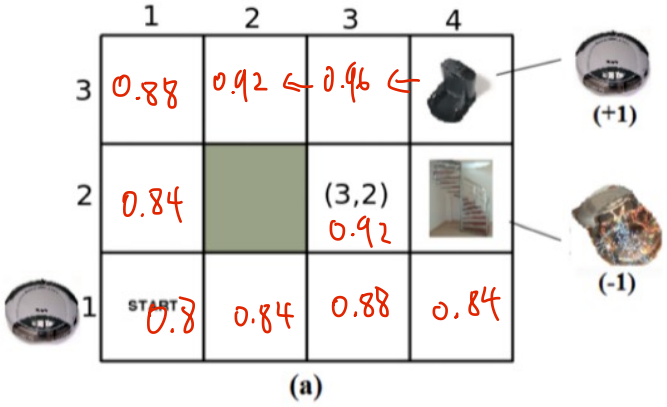
State : $s \in S$ Action: $a \in \{left, right, up, down\}$

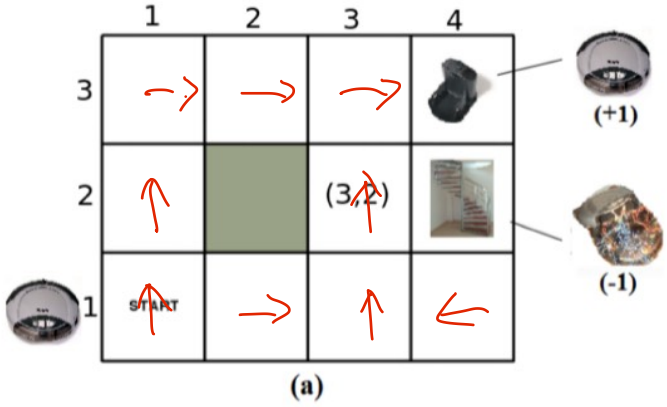
Modle : $P(s' | s, a)$

Reward : $R(s', s, a) = R(s) = \begin{cases} -0.04 & \text{move} \\ 1 & s = \text{charge} \\ -1 & s = \text{stairs} \end{cases}$

11.1.1 Bellman Value Update

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





Problem 11.1.3: Assuming the transition probability as stochastic and calculate the value of $U(3,3)$ using the value iteration algorithm for 2 iterations. Assume that all unknown initial utilities are zero and $U^1(1,3) = -0.04$, $U^1(2,3) = -0.04$ and $U^1(3,2) = -0.04$.

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

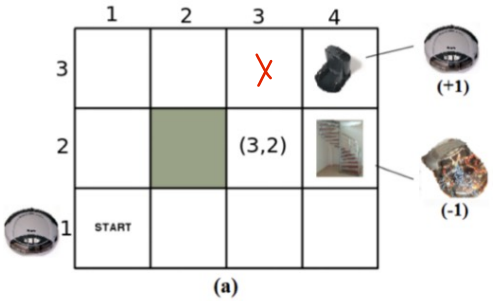


Figure 1: Problem 11.1

Iteration 1

$$U'(3,3) = R(3,3) + \gamma \max [$$

right $\Rightarrow P((3,3)|(3,3),r) \cdot U^0(3,3) + P((4,3)|(3,3),r) \cdot U^0(4,3) + P((3,2)|(3,3),r) \cdot U^0(3,2)$

$0.8 = 0.1 \cdot 0 + 0.8 \cdot 1 + 0.1 \cdot 0$

left $\Rightarrow P((3,2)|(3,3),l) \cdot U^0(3,2) + P((2,3)|(3,3) \cdot l) \cdot U^0(2,3) + P((3,3)|(3,3) \cdot l) \cdot U^0(3,3)$

$0 = 0.1 \cdot 0 + 0.8 \cdot 0 + 0.1 \cdot 0$

up $\Rightarrow P((2,3)|(3,3),u) \cdot U^0(2,3) + P((3,3)|(3,3) \cdot u) \cdot U^0(3,3) + P((4,3)|(3,3),u) \cdot U^0(4,3)$

$0.1 = 0.1 \cdot 0 + 0.8 \cdot 0 + 0.1 \cdot 1$

down $\Rightarrow P((4,3)|(3,3),d) \cdot U^0(4,3) + P((3,2)|(3,3) \cdot d) \cdot U^0(3,2) + P((2,3)|(3,3),d) \cdot U^0(2,3)$

$0.1 = 0.1 \cdot 1 + 0.8 \cdot 0 + 0.1 \cdot 0$

$U'(3,3) = -0.04 + 1 \cdot 0.8 = 0.76$ (Right)

Iteration 2

$$U^2(3,3) = R(3,3) + \gamma \max [$$

right $\Rightarrow P((3,3)|(3,3),r) \cdot U'(3,3) + P((4,3)|(3,3),r) \cdot U'(4,3) + P((3,2)|(3,3),r) \cdot U'(3,2)$

$0.872 = 0.1 \cdot 0.76 + 0.8 \cdot 1 + 0.1 \cdot -0.04$

left $\Rightarrow P((3,2)|(3,3),l) \cdot U'(3,2) + P((2,3)|(3,3) \cdot l) \cdot U^0(2,3) + P((3,3)|(3,3) \cdot l) \cdot U^0(3,3)$

$0.04 = 0.1 \cdot -0.04 + 0.8 \cdot -0.04 + 0.1 \cdot 0.76$

up $\Rightarrow P((2,3)|(3,3),u) \cdot U^0(2,3) + P((3,3)|(3,3) \cdot u) \cdot U^0(3,3) + P((4,3)|(3,3),u) \cdot U^0(4,3)$

$0.764 = 0.1 \cdot -0.04 + 0.8 \cdot 0.76 + 0.1 \cdot 1$

down $\Rightarrow P((4,3)|(3,3),d) \cdot U^0(4,3) + P((3,2)|(3,3) \cdot d) \cdot U^0(3,2) + P((2,3)|(3,3),d) \cdot U^0(2,3)$

$0.004 = 0.1 \cdot 1 + 0.8 \cdot -0.04 + 0.1 \cdot -0.04$

$U^2(3,3) = -0.04 + 1 \times 0.872 = 0.832$ (Right)

Problem 11.1.4: Compute the optimal policy of state (3, 1) after convergence. The utilities after convergence are shown in Fig. 1 (c).

3	0.812	0.868	0.912	1
2	0.762		0.660	-1
1	0.705	0.655	0.611	0.388
	1	2	(c) 3	

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

$\arg\max$ [right $\rightarrow 0.1 \cdot U(3,2) + 0.8 \cdot U(4,1) + 0.1 \cdot U(3,1) = 0.4375$

left $\rightarrow 0.1\ U(3,1) + 0.8 \cdot U(2,1) + 0.1 \cdot U(3,2) = 0.6511$

up $\rightarrow 0.1\ U(2,1) + 0.8 \cdot U(3,2) + 0.1\ U(4,1) = 0.6323$

down $\rightarrow 0.1\ U(4,1) + 0.8\ U(3,1) + 0.1\ U(2,1) = 0.5911$

$\pi^*(s) =$ left

Problem 11.2:

(Taken from **Andrew W. Moore**) Assume that you run a startup company. In every decision period, you must choose between Saving money (S) or Advertising (A). If you advertise, you may become famous (f) (50%) but because of spending money you may become poor (p). If you save money, you may become rich (r) with probability 50% but you may become also unknown (u) because you don't advertise.

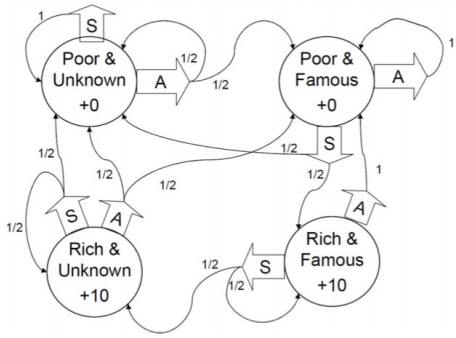


Figure 2: Problem 11.2

Problem 11.2.1: Calculate the utility value for state $U(r, u)$ for 2 iterations using value iteration. Assume that the discount factor is 0.9 and that all initial utilities are zero. Furthermore use $U^1(p, f) = 0,\ U^1(p, u) = 0$.

$\gamma = 0.9$

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

Iteration 1

$U'(r,u) = R(r,u) + \gamma \cdot \max$ [

$A \Rightarrow P((p,u) \mid (r,u), A) \cdot U^0(p,u) + P((p,f) \mid (r,u), A) \cdot U^0(p,f)$,

$S \Rightarrow P((p,u) \mid (r,u), S) \cdot U^0(p,u) + P((r,u) \mid (r,u), S) \cdot U^0(r,u)$]

$= 10 + 0.9 \cdot 0 = 10$

Iteration 2

$U^2(r,u) = R(r,u) + \gamma \cdot \max$ [

$A \Rightarrow P((p,u) \mid (r,u), A) \cdot U^1(p,u) + P((p,f) \mid (r,u), A) \cdot U^1(p,f)$,

$S \Rightarrow P((p,u) \mid (r,u), S) \cdot U^1(p,u) + P((r,u) \mid (r,u), S) \cdot U^1(r,u)$]

$= 10 + 0.9 \cdot 5 = 14.5$

$\frac{1}{2} \cdot 10$

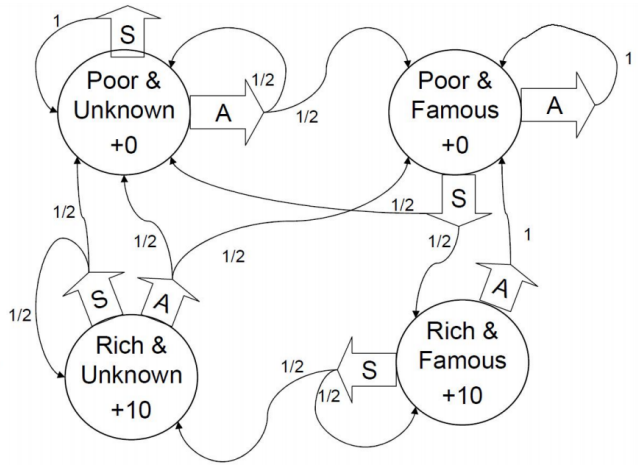


Figure 2: Problem 11.2

Problem 11.2.1: Calculate the utility value for state $U(r, u)$ for 2 iterations using value iteration. Assume that the discount factor is 0.9 and that all initial utilities are zero. Furthermore use $U^1(p, f) = 0,\ U^1(p, u) = 0$.

Additional task: Calculate the utility values for all other states for 2 iterations using value iteration.

Problem 11.3:

Assume that you have an Artificial Intelligence exam tomorrow and you can be either ready (r) or not ready ($\neg r$) for the exam (see Fig. 3). Apply the policy iteration algorithm for one iteration in order to determine the policies $\pi_1(\neg r)$ and $\pi_1(r)$. Assume that the discount factor is $\gamma = 0.9$ and the initial policies are $\pi_0(\neg r) = s$ and $\pi_0(r) = e$. The rewards for $\neg r$ and r are -10 and 10, respectively.

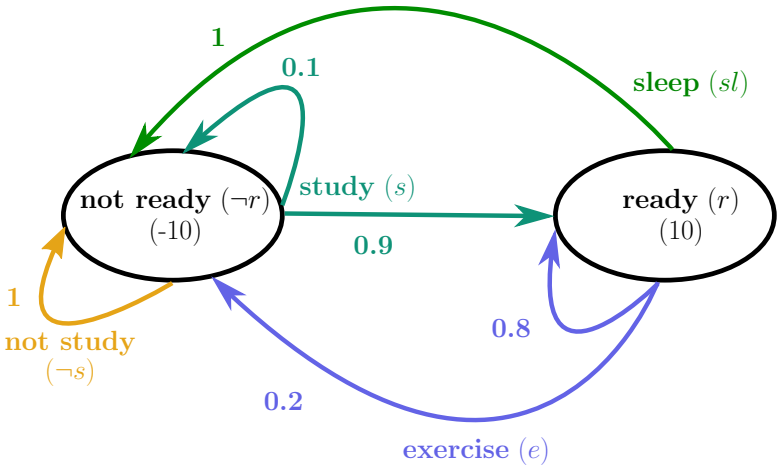


Figure 3: Problem 11.3

Problem 11.3:

Assume that you have an Artificial Intelligence exam tomorrow and you can be either ready (r) or not ready ($\neg r$) for the exam (see Fig. 3). Apply the policy iteration algorithm for one iteration in order to determine the policies $\pi_1(\neg r)$ and $\pi_1(r)$. Assume that the discount factor is $\gamma = 0.9$ and the initial policies are $\pi_0(\neg r) = s$ and $\pi_0(r) = e$. The rewards for $\neg r$ and r are -10 and 10, respectively.

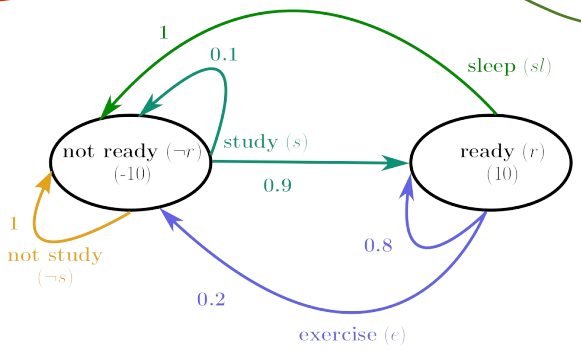


Figure 3: Problem 11.3

$\pi_0(r) = e$

$\pi_0(\neg r) = s$

policy Evaluation $V_i^{\pi}(s) = R(s) + \gamma \cdot \sum_{s'} p(s'|s,a) \cdot V_i^{\pi}(s')$

$V_0^{\pi_1}(r) = R(r) + \gamma \cdot [$

$\pi_0 \rightarrow p(\neg r | r, \pi_0) \cdot V_0^{\pi}(\neg r) + p(r|r,\pi_0) \cdot V_0^{\pi}(r)]$

$0.2 \cdot V_0^{\pi}(\neg r) + 0.8 \cdot V_0^{\pi}(r)$

$\Rightarrow V_0^{\pi}(r) = 10 + 0.9[0.2\, V_0^{\pi}(\neg r) + 0.8\, V_0^{\pi}(r)]$

$= 10 + 0.18\, V_0^{\pi}(\neg r) + 0.72\, V_0^{\pi}(r)$

$0.28\, V_0^{\pi}(r) - 0.18\, V_0^{\pi}(\neg r) = 10$

$V_0^{\pi}(\neg r) = R(\neg r) + \gamma \cdot [P(r|\neg r, \pi_0) \cdot V_0^{\pi}(r) + P(\neg r | \neg r, \pi_0) \cdot V_0^{\pi}(\neg r)]$

$= -10 + 0.9[0.9 \cdot V_0^{\pi}(r) + 0.1\, V_0^{\pi}(\neg r)]$

$= -10 + 0.81\, V_0^{\pi}(r) + 0.09\, V_0^{\pi}(\neg r)$

$0.91\, V_0^{\pi}(\neg r) - 0.81\, V_0^{\pi}(r) = -10$

$V_0^{\pi}(\neg r) = 48.62$

$V_0^{\pi}(r) = 66.97$

policy improvement

$$\pi_{i+1}(s) = \underset{a \in A(s)}{\operatorname{argmax}} \sum_{s'} P(s'|s,a)\, U_i(s')$$

$\pi_1(\neg r) = \operatorname{argmax}$ [

study → $P(r|\neg r, s) \cdot U_0(r) + P(\neg r|\neg r, s) \cdot U_0(\neg r)$

$0.9 \cdot 66.97 + 0.1 \cdot 48.62 = 65.14$

hot study → $P(\neg r|\neg r, \neg s) \cdot U_0(\neg r)$

$1 \cdot 48.62 = 48.62$

$\pi_1(\neg r) =$ Study

$\pi_1(r) = \operatorname{argmax}$ [

exercise → $P(r|r,e) \cdot U_0(r) + P(\neg r|r,e) \cdot U_0(\neg r)$

$0.8 \cdot 66.97 + 0.2 \cdot 48.62 = 63.3$

Sleep → $P(\neg r|r,s) \cdot U_0(\neg r)$

$1 \cdot 48.62 = 48.62$

$\pi_1(r) =$ exercise

References

Russell and Norvig (2010), Artificial Intelligence: A Modern Approach. Prentice Hall
Andrew W. Moore, Markov Decision Processes, Tutorial Slides. https://www.autonlab.org/_media/tutorials/mdp09.pdf