Fundamentals of Artificial Intelligence
Exercise 11: Making Complex Decisions

Jonathan Külz

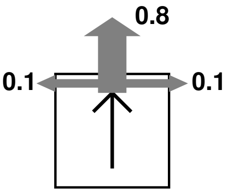
Technical University of Munich

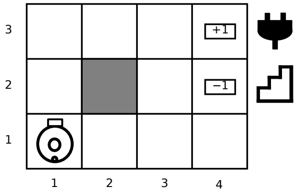
February 02nd, 2024

Summary - Rational Decisions Over Time

- Sequential decision problems in uncertain discrete environments can be modeled as **Markov decision processes (MDPs)**
- The utility of a state sequence is the sum of all the rewards over the sequence, possibly discounted over time.
- The optimal solution of an MDP is a **policy** that associates a decision with every state that the agent might reach. A solution can be obtained by **value iteration**.
- **Policy iteration** usually converges faster, since a policy might already be optimal without knowing the exact utilities of each state.

Problem 11.1: Roomba Problem



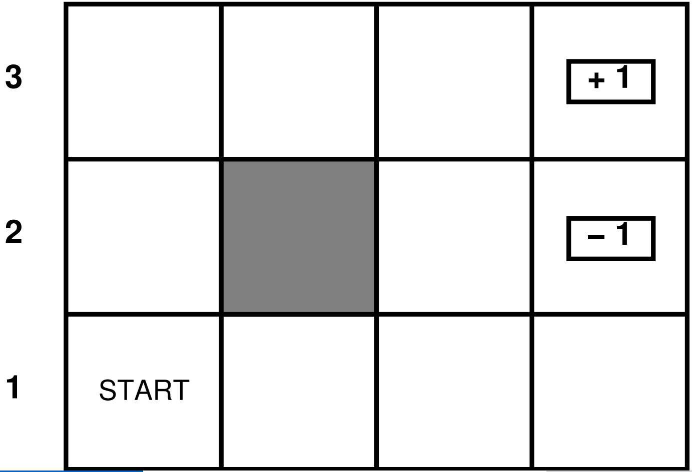


- States $s \in S$, actions $a \in A = \{Up, Down, Left, Right\}$.
- **Model** $P(s'|s,a)$ = probability that a in s leads to s'.
- **Reward function** (with terminal states $\mathcal{S}_T = \{s_{charge}, s_{stairs}\}$)

$$R(s,a,s') = R(s) = \begin{cases} 1 & \text{if } s = s_{charge} \\ -1 & \text{if } s = s_{stairs} \\ -0.04 & \forall s \notin \mathcal{S}_T \end{cases}$$

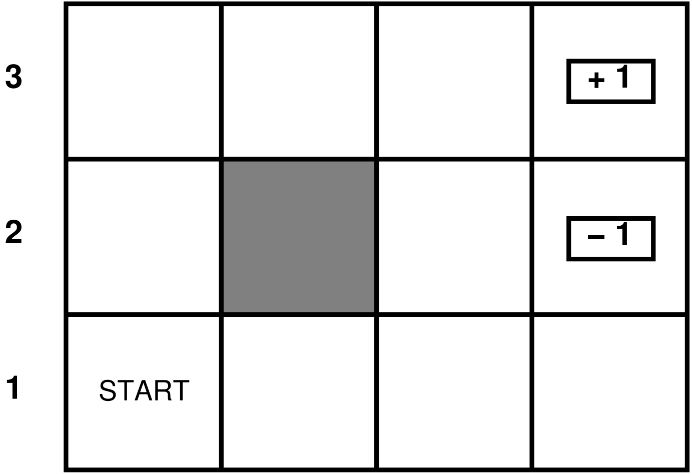
Problem 11.1: Roomba Problem

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



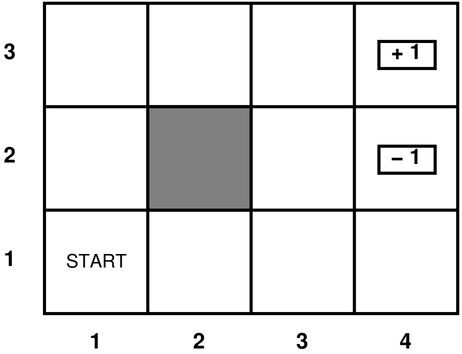
Problem 11.1: Roomba Problem

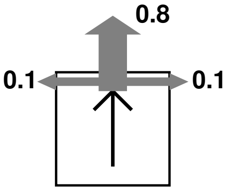
Problem 11.1.2 Show the corresponding **policy**.



Problem 11.1: Roomba Problem

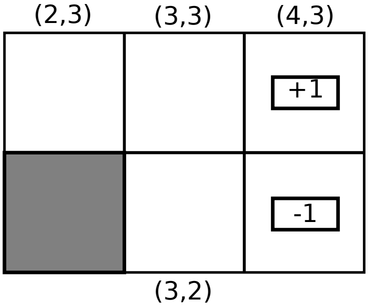
Problem 11.1.3 Assume that the transition probability is stochastic. Calculate the value of $U(3,3)$ using the **value iteration** algorithm for 2 iterations. Assume that all initial utilities are zero and $U^1(1,3) = -0.04$, $U^1(2,3) = -0.04$ and $U^1(3,2) = -0.04$.





Problem 11.1: Roomba Problem

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



Bellman equation (if reward depends on state only)

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

Problem 11.1: Roomba Problem

Problem 11.1.3 Compute $U(3,3)$

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

Iteration 1 $U^1(3,3) = R(3,3) + \gamma \max\Big[$

$$\begin{aligned}
&P((3,3)|(3,3),r)\cdot U^0(3,3) && + P((4,3)|(3,3),r)\cdot U^0(4,3) && + P((3,2)|(3,3),r)\cdot U^0(3,2)), && \text{(Right)}\\
&P((3,3)|(3,3),l)\cdot U^0(3,3) && + P((2,3)|(3,3),l)\cdot U^0(2,3) && + P((3,2)|(3,3),l)\cdot U^0(3,2)), && \text{(Left)}\\
&P((2,3)|(3,3),u)\cdot U^0(2,3) && + P((3,3)|(3,3),u)\cdot U^0(3,3) && + P((4,3)|(3,3),u)\cdot U^0(4,3)), && \text{(Up)}\\
&P((2,3)|(3,3),d)\cdot U^0(2,3) && + P((3,2)|(3,3),d)\cdot U^0(3,2) && + P((4,3)|(3,3),d)\cdot U^0(4,3))\Big] && \text{(Down)}
\end{aligned}$$

$$\begin{aligned}
U^1(3,3) = \quad & -0.04+ \quad \max \quad && [(0.1\cdot 0 + 0.8\cdot 1 + 0.1\cdot 0), && \text{(Right)}\\
& && (0.1\cdot 0 + 0.8\cdot 0 + 0.1\cdot 0), && \text{(Left)}\\
& && (0.1\cdot 0 + 0.8\cdot 0 + 0.1\cdot 1), && \text{(Up)}\\
& && (0.1\cdot 0 + 0.8\cdot 0 + 0.1\cdot 1)] && \text{(Down)}\\
U^1(3,3) = \quad & 0.760\text{(Right)}
\end{aligned}$$

Problem 11.1: Roomba Problem

Problem 11.1.3 Compute $U(3,3)$

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

Iteration 2

$$\begin{aligned}
&P((3,3)|(3,3),r) \cdot U^1(3,3) &&+ P((4,3)|(3,3),r) \cdot U^1(4,3) &&+ P((3,2)|(3,3),r) \cdot U^1(3,2)), &&\text{(Right)}\\
&P((3,3)|(3,3),l) \cdot U^1(3,3) &&+ P((2,3)|(3,3),l) \cdot U^1(2,3) &&+ P((3,2)|(3,3),l) \cdot U^1(3,2)), &&\text{(Left)}\\
&P((2,3)|(3,3),u) \cdot U^1(2,3) &&+ P((3,3)|(3,3),u) \cdot U^1(3,3) &&+ P((4,3)|(3,3),u) \cdot U^1(4,3)), &&\text{(Up)}\\
&P((2,3)|(3,3),d) \cdot U^1(2,3) &&+ P((3,2)|(3,3),d) \cdot U^1(3,2) &&+ P((4,3)|(3,3),d) \cdot U^1(4,3))] &&\text{(Down)}
\end{aligned}$$

Problem 11.1: Roomba Problem

Problem 11.1.3 Compute $U(3,3)$

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

Iteration 2

$$
\begin{aligned}
U^2(3,3) = -0.04 \ + \ \max \quad & [(0.1 \cdot (0.760) + 0.8 \cdot (1) + 0.1 \cdot (-0.04)), && \text{(Right)} \\
& (0.1 \cdot 0.76 + 0.8 \cdot (-0.04) + 0.1 \cdot (-0.04)), && \text{(Left)} \\
& (0.1 \cdot (-0.04) + 0.8 \cdot (0.760) + 0.1 \cdot 1), && \text{(Up)} \\
& (0.1 \cdot (-0.04) + 0.8 \cdot (-0.04) + 0.1 \cdot 1)] && \text{(Down)}
\end{aligned}
$$

$$
U^2(3,3) = 0.832 \qquad \text{(Right)}
$$

Problem 11.1: Roomba Problem

Problem 11.1.4 Compute the **optimal policy** of state $(3, 1)$ after convergence. The utilities after convergence are given.

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	3	4

Problem 11.1: Roomba Problem

Problem 11.1.4 Compute the **optimal policy** of state $(3, 1)$ after convergence. The utilities after convergence are given.

	1	2	3	4
3	0.812	0.868	0.912	+ 1
2	0.762		0.660	– 1
1	0.705	0.655	0.611	0.388

Optimal Policy

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

Problem 11.1: Roomba Problem

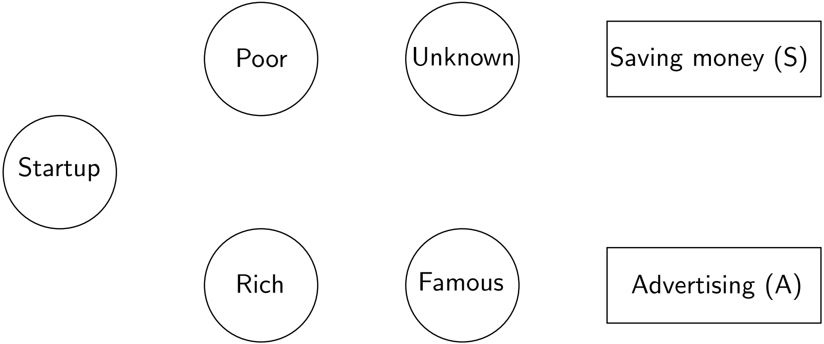
Problem 11.1.4

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

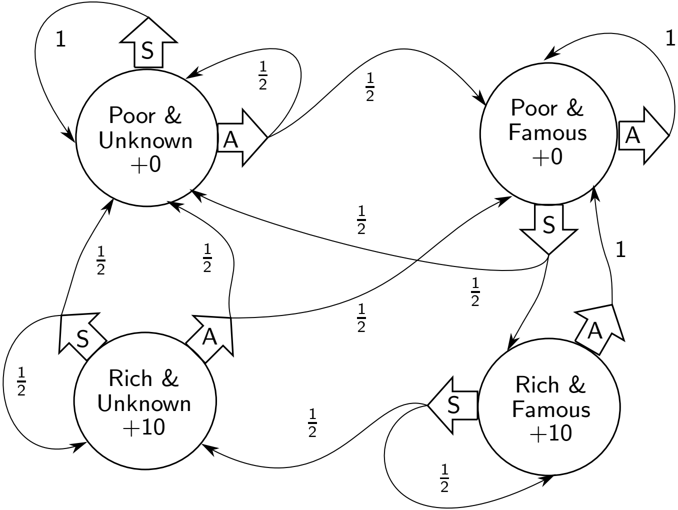
3	**0.812**	**0.868**	**0.912**	**+1**
2	**0.762**		**0.660**	**–1**
1	**0.705**	**0.655**	**0.611**	**0.388**
	1	**2**	**3**	**4**

Problem 11.2: Startup Dilemma

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.

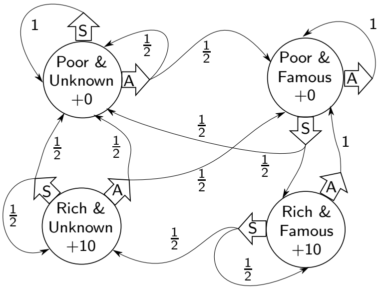


Problem 11.2: Startup Dilemma



Problem 11.2: Startup Dilemma

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 states are zero. Furthermore use $U^1(p,f) = 0,\ U^1(p,u) = 0$.



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

Problem 11.2: Startup Dilemma

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 states are zero. Furthermore use $U^1(p,f) = 0,\ U^1(p,u) = 0$.

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

Iteration 2:

$$\begin{aligned} U^2(r,u) = \quad R(r,u) \quad + \quad \gamma \max \quad & [P((p,u)|(r,u),A) \cdot U^1(p,u) & \text{(A)} \\ & +P((p,f)|(r,u),A) \cdot U^1(p,f), & \\ & P((p,u)|(r,u),S) \cdot U^1(p,u) & \text{(S)} \\ & +P((r,u)|(r,u),S) \cdot U^1(r,u)], & \end{aligned}$$

Problem 11.2: Startup Dilemma

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 states are zero. Furthermore use $U^1(p, f) = 0,\ U^1(p, u) = 0.$

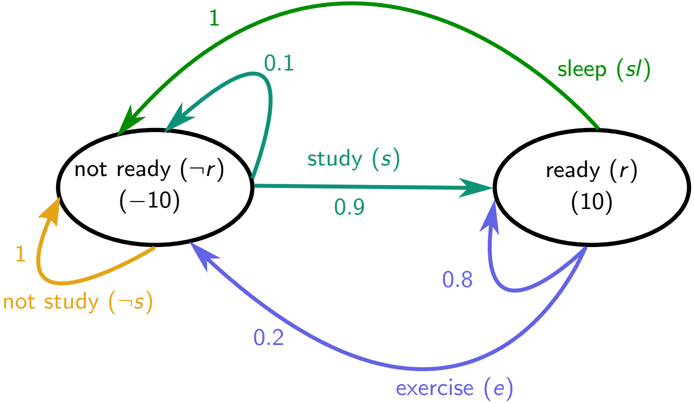
$$U(s) = R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s, a)U(s')$$

Iteration 2:

$$\begin{aligned} U^2(r,u) = \quad & R(r,u) \quad + \quad \gamma \max \quad [P((p,u)|(r,u),A) \cdot U^1(p,u) && \text{(A)} \\ & \qquad\qquad\qquad\qquad + P((p,f)|(r,u),A) \cdot U^1(p,f), && \\ & \qquad\qquad\qquad\qquad P((p,u)|(r,u),S) \cdot U^1(p,u) && \text{(S)} \\ & \qquad\qquad\qquad\qquad + P((r,u)|(r,u),S) \cdot U^1(r,u)], && \end{aligned}$$

$$\begin{aligned} U^2(r,u) = \quad & 10 \quad + \quad 0.9 \max \quad [0.5 \cdot 0 + 0.5 \cdot 0, && \text{(A)} \\ & \qquad\qquad\qquad\qquad 0.5 \cdot 0 + 0.5 \cdot 10], && \text{(S)} \\ U^2(r,u) = \quad & 14.5 \end{aligned}$$

Problem 11.3: AI Exam



Problem 11.3: AI Exam

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.

Problem 11.3: AI Exam

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.

Policy iteration

- **Policy evaluation**: Given a policy π_i, calculate $U_i = U^{\pi_i}$, the utility of each state if π_i were to be executed.
- **Policy improvement**: Calculate a new policy π_{i+1} using a one-step look-ahead based on U_i using $\pi_{i+1}(s) = \arg\max_{a \in A(s)} \sum_{s'} P(s'|s,a)U(s')$.

Problem 11.3: AI Exam

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.

Step 1. Policy evaluation $$U_i(s) = R(s) + \gamma \sum_{s'} P(s'|s, \pi_i(s)) U_i(s')$$

Compute $U_0(r)$ and $U_0(\neg r)$:

Problem 11.3: AI Exam

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.

Step 1. Policy evaluation $\quad U_i(s) = R(s) + \gamma \sum_{s'} P(s'|s, \pi_i(s))U_i(s')$

Compute $U_0(r)$ and $U_0(\neg r)$:

Problem 11.3: AI Exam

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.

Step 1. Policy evaluation $U_i(s) = R(s) + \gamma \sum_{s'} P(s'|s, \pi_i(s)) U_i(s')$

Summarize the linear equations:

$$\begin{aligned} 0.91 \cdot U_0(\neg r) - 0.81 \cdot U_0(r) &= -10 \\ (-0.18) \cdot U_0(\neg r) + 0.28 \cdot U_0(r) &= 10 \end{aligned}$$

Solution:

$$\begin{aligned} U_0(r) &= 66.7, \\ U_0(\neg r) &= 48.4. \end{aligned}$$

Problem 11.3: AI Exam

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.

Step 2. Policy improvement $\pi_{i+1}(s) = \underset{a \in A(s)}{\arg\max} \sum_{s'} P(s'|s, a)\, U_i(s')$

Compute $\pi_1(\neg r)$:

Problem 11.3: AI Exam

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.

Step 2. Policy improvement $\quad \pi_{i+1}(s) = \underset{a \in A(s)}{\arg\max} \sum_{s'} P(s'|s, a)\, U_i(s')$

Compute $\pi_1(r)$: