Fundamentals of Artificial Intelligence
Exercise 11: Making Complex Decisions (over time)

Jonathan Külz

Technical University of Munich

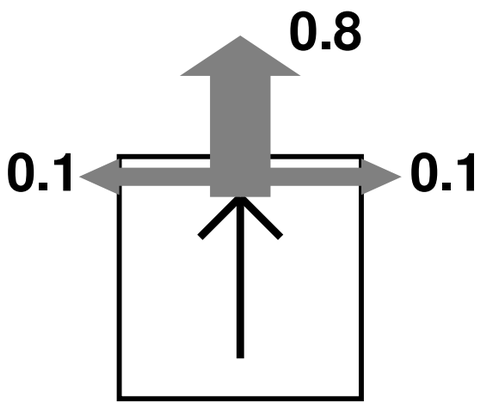
February 02nd, 2024

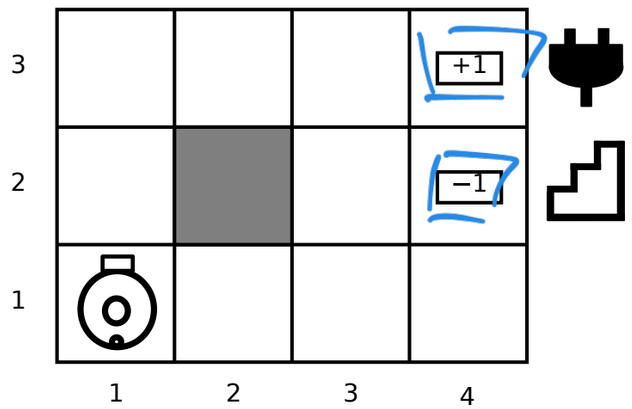
Summary - Rational Decisions Over Time

$P(s' \mid s, a)$

- 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





$\gamma = 1$

- 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 $\mathcal{S}_T = \{s_{charge}, s_{stairs}\}$)

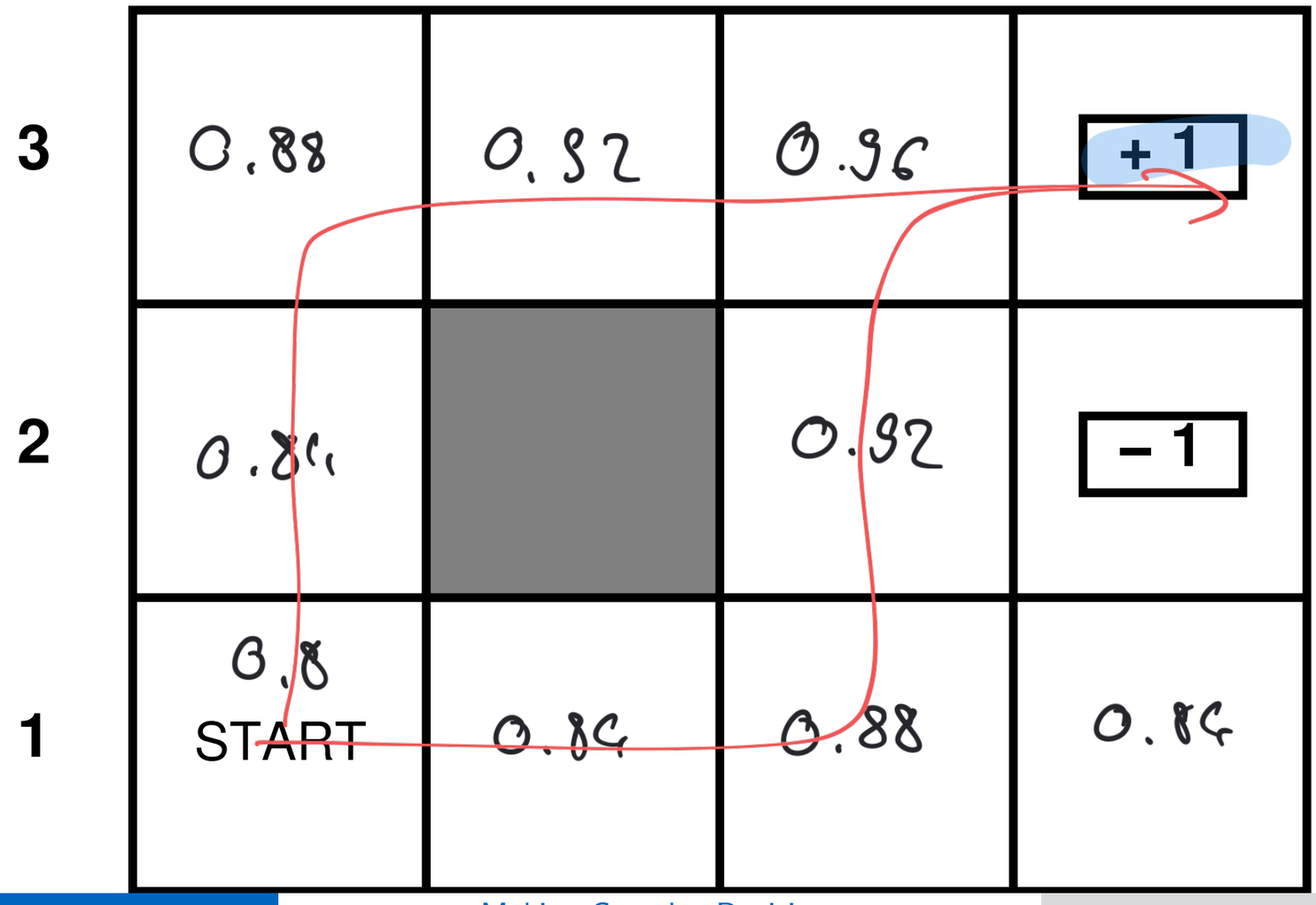
$$R(s, a, s') = R(s) = \begin{cases} 1 & \text{if } s = s_{charge} \\ -1 & \text{if } s = s_{stairs} \\ -0.04 & \forall s \notin \mathcal{S}_T \end{cases}$$

Problem 11.1: Roomba Problem

$$U(s) = R(s) + \gamma \cdot \max_a \sum P(s' \mid s, a) U(s')$$

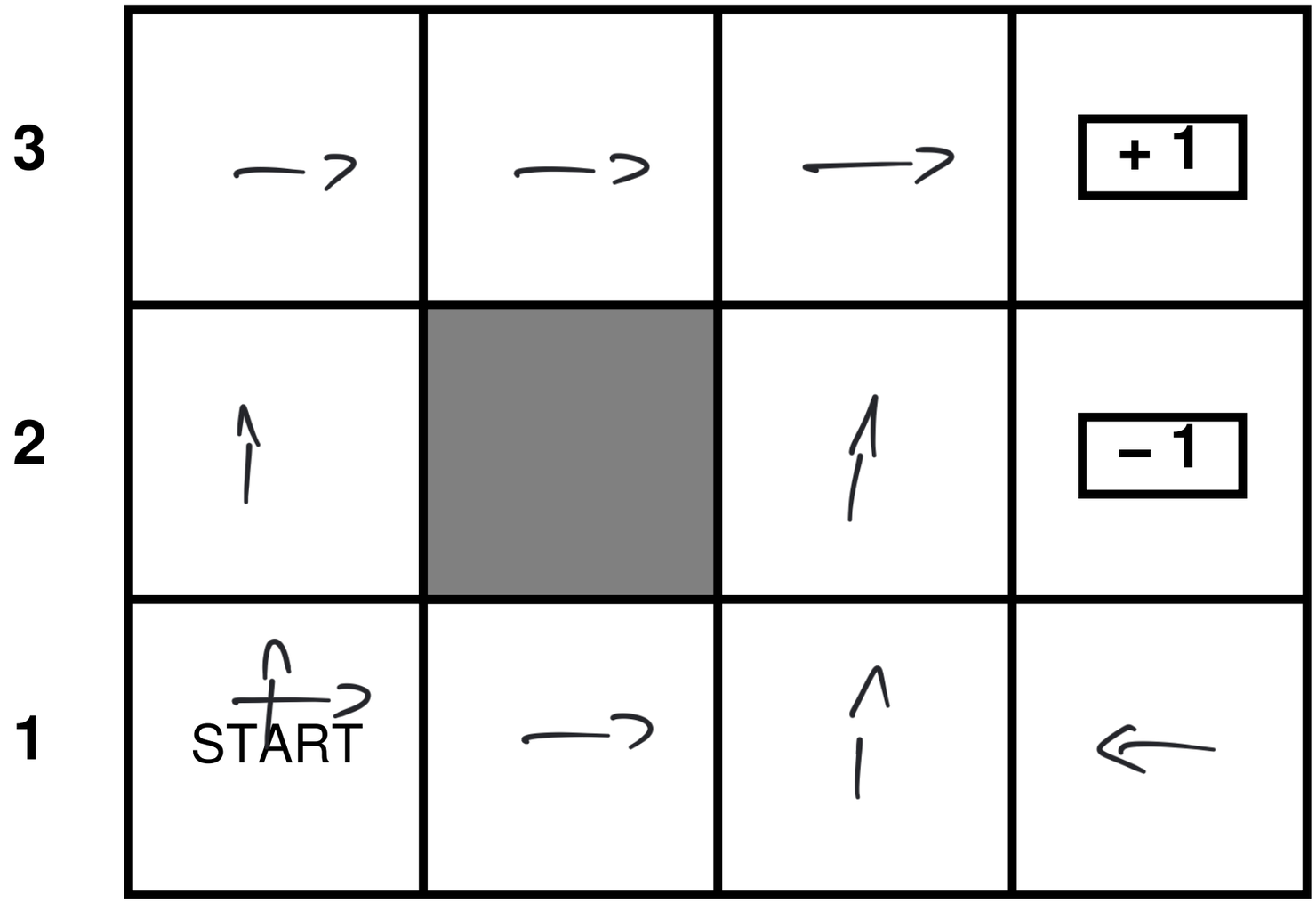
Problem 11.1.1 Assuming the transition probability as **deterministic** and the discount factor as 1. Find the **value** of all states.

$$R(s) = \begin{cases} -0.04 \\ -1 \\ 1 \end{cases}$$



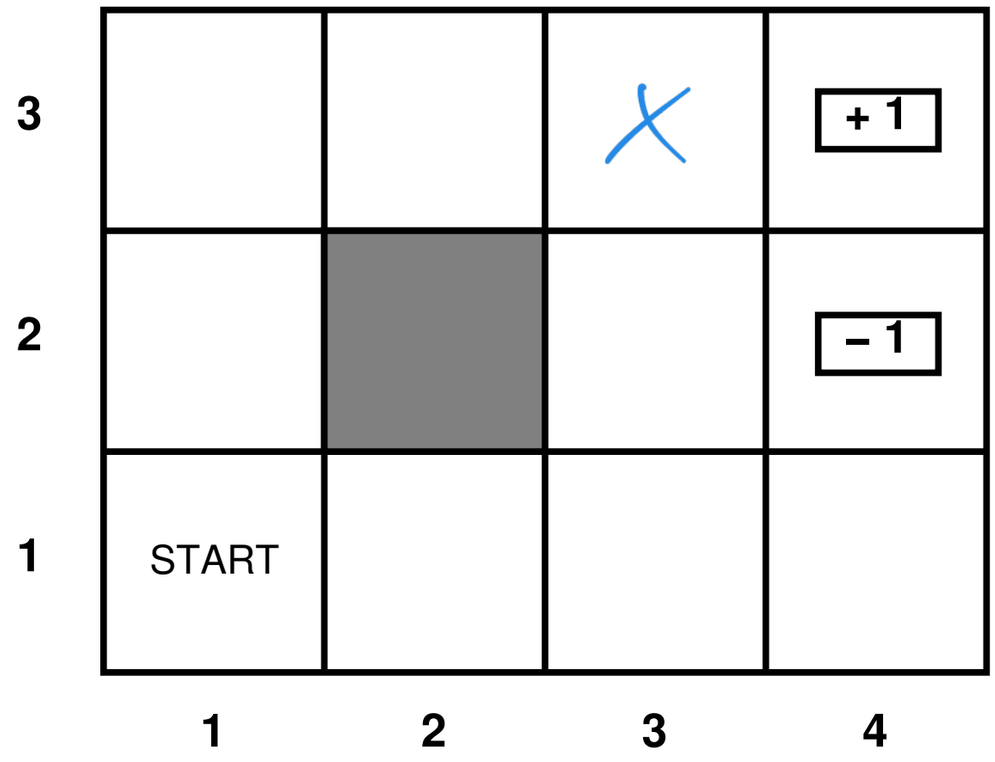
Problem 11.1: Roomba Problem

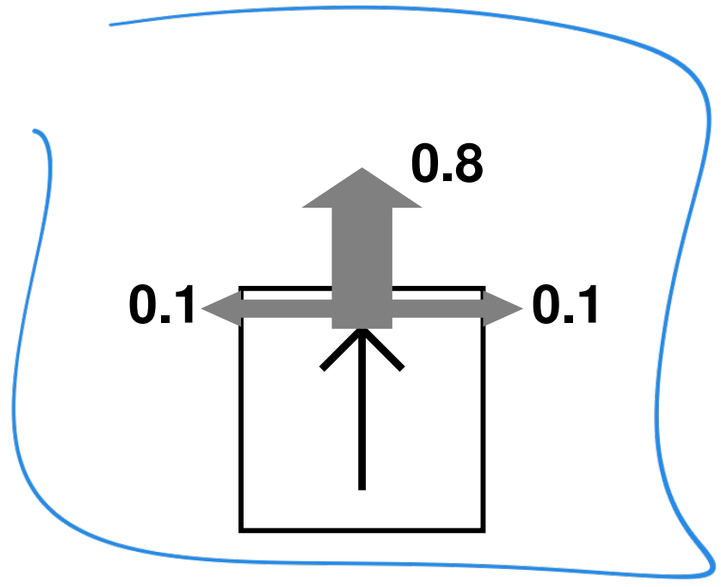
Problem 11.1.2 Show the corresponding **policy**.



Problem 11.1: Roomba Problem

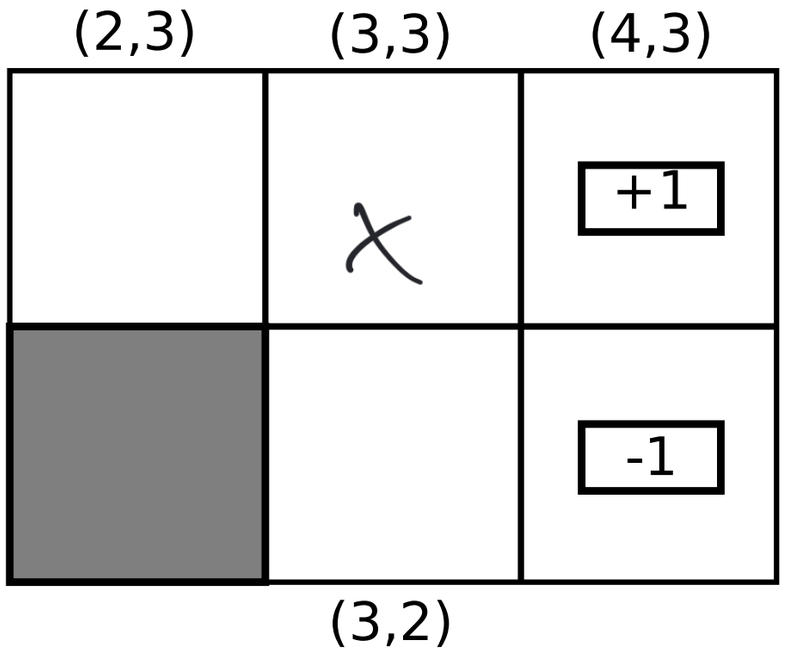
Problem 11.1.3 Assume that the transition probability is stochastic. Calculate the value of $U(3,3)$ using the **value iteration** algorithm for 2 iterations. Assume that all initial utilities are zero and $U^1(1,3) = -0.04$, $U^1(2,3) = -0.04$ and $U^1(3,2) = -0.04$.





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 $U^1(3,2) = -0.04$.



Lecture: $U(s) = \max_{a \in A(s)} \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma U(s')\right]$

Bellman equation (if reward depends on state only)

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

$$U_{\pi_t}(s) = \max_a \mathbb{E}\left[\sum_t \gamma^t R_t \mid a, s_t\right]$$

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

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

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

$$U^1(3,3) = 0.760 (\text{Right})$$

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 $U(3,3)$

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

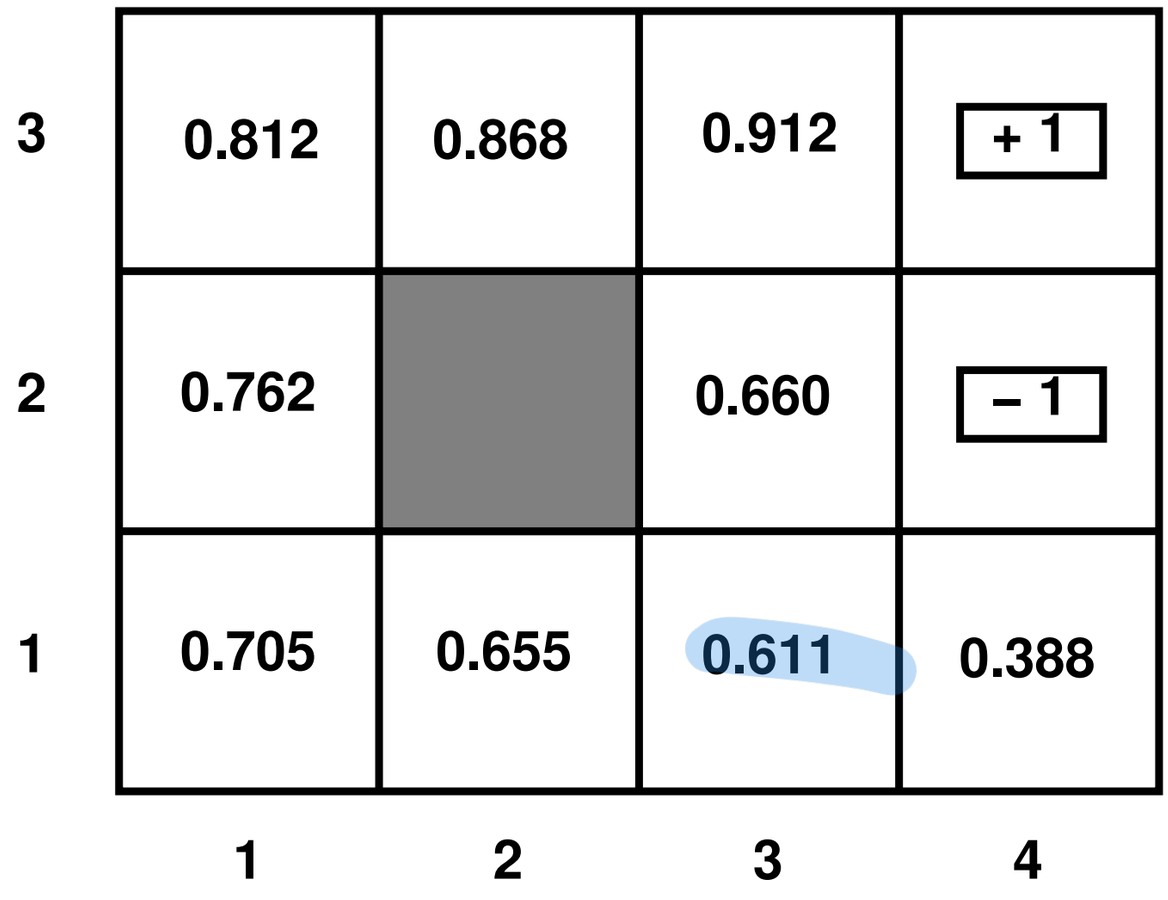
Iteration 2

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

$U^2(3,3) = 0.832$ (Right)

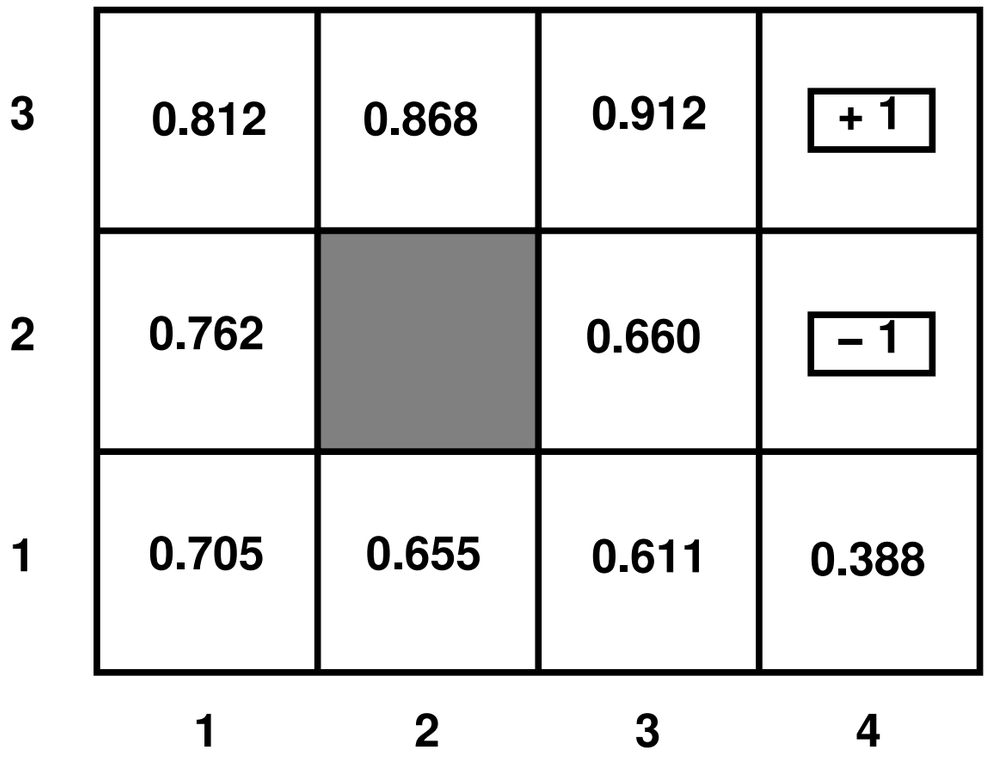
Problem 11.1: Roomba Problem

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



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.



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

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

$= \arg\max_a$ [$0.1\,U(3,2) + 0.8\,U(4,1) + 0.1\,U(3,1)$, ← r
$0.1\,U(3,2) + 0.8\,U(2,1) + 0.1\,U(3,1)$, ← l
$0.1\,U(2,1) + 0.8\,U(3,2) + 0.1\,U(4,1)$, ← u
$0.1\,U(4,1) + 0.8\,U(3,1) + 0.1\,U(2,1)$] ← d

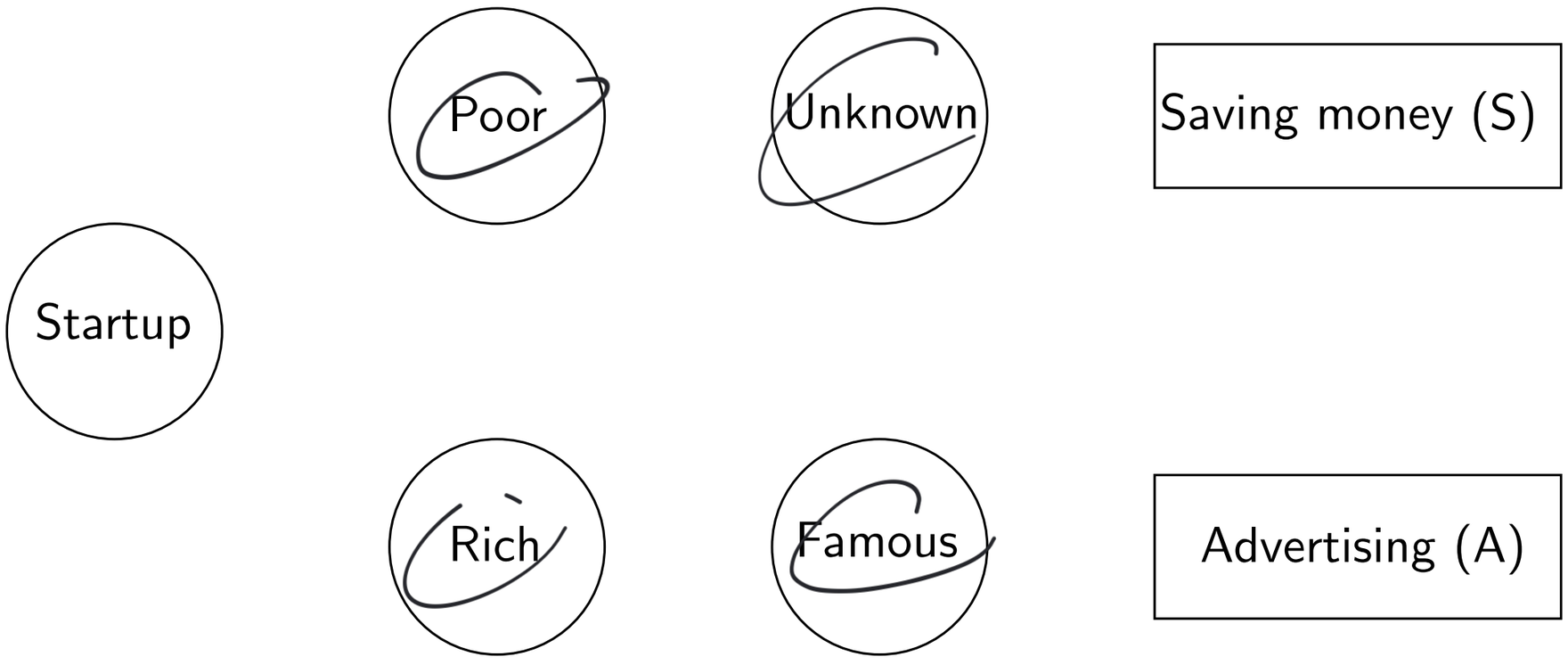
$= \arg\max_a$ [0.4375, r
0.6511, l
0.6323, u
0.5931] d

$\pi^*(s) = l$

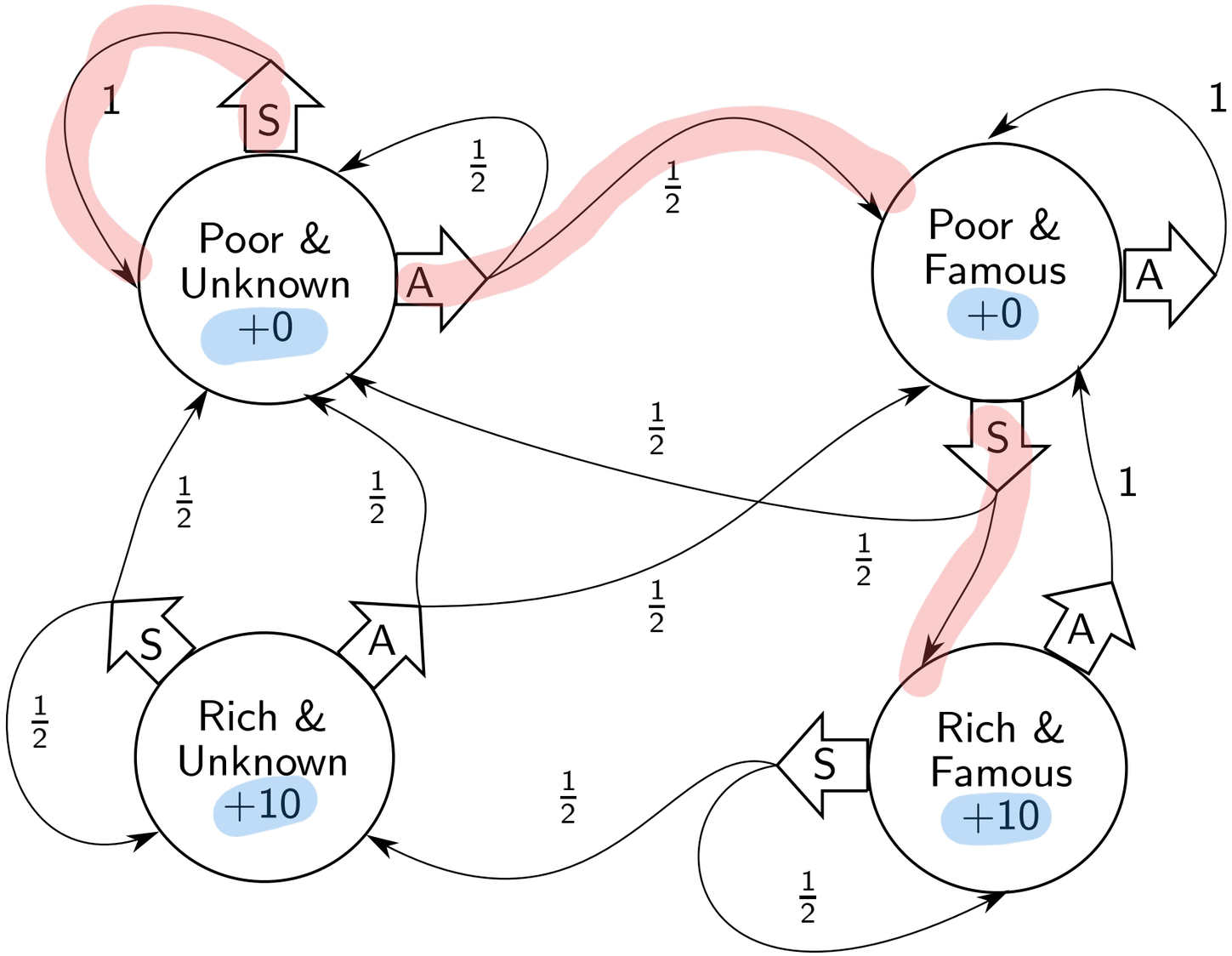
3	0.812	0.868	0.912	+1
2	0.762		0.660	−1
1	0.705	0.655	0.611	0.388
	1	2	3	4

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.

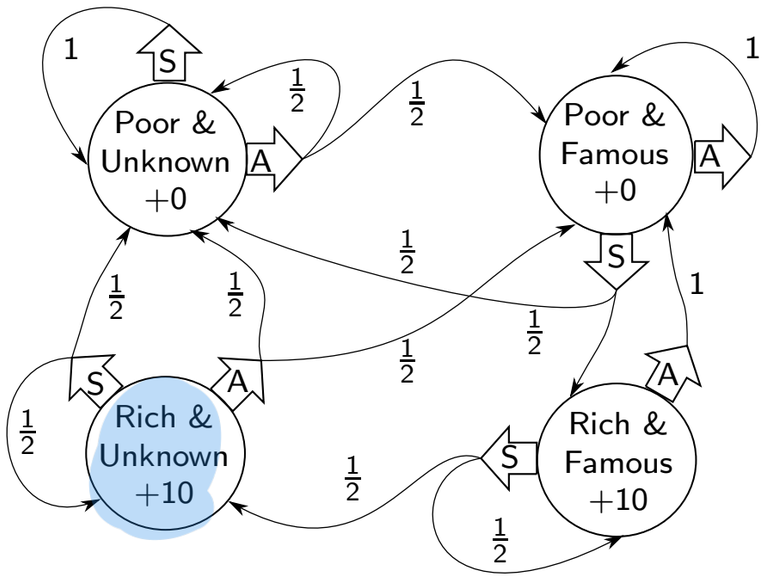


Problem 11.2: Startup Dilemma



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:

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

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:

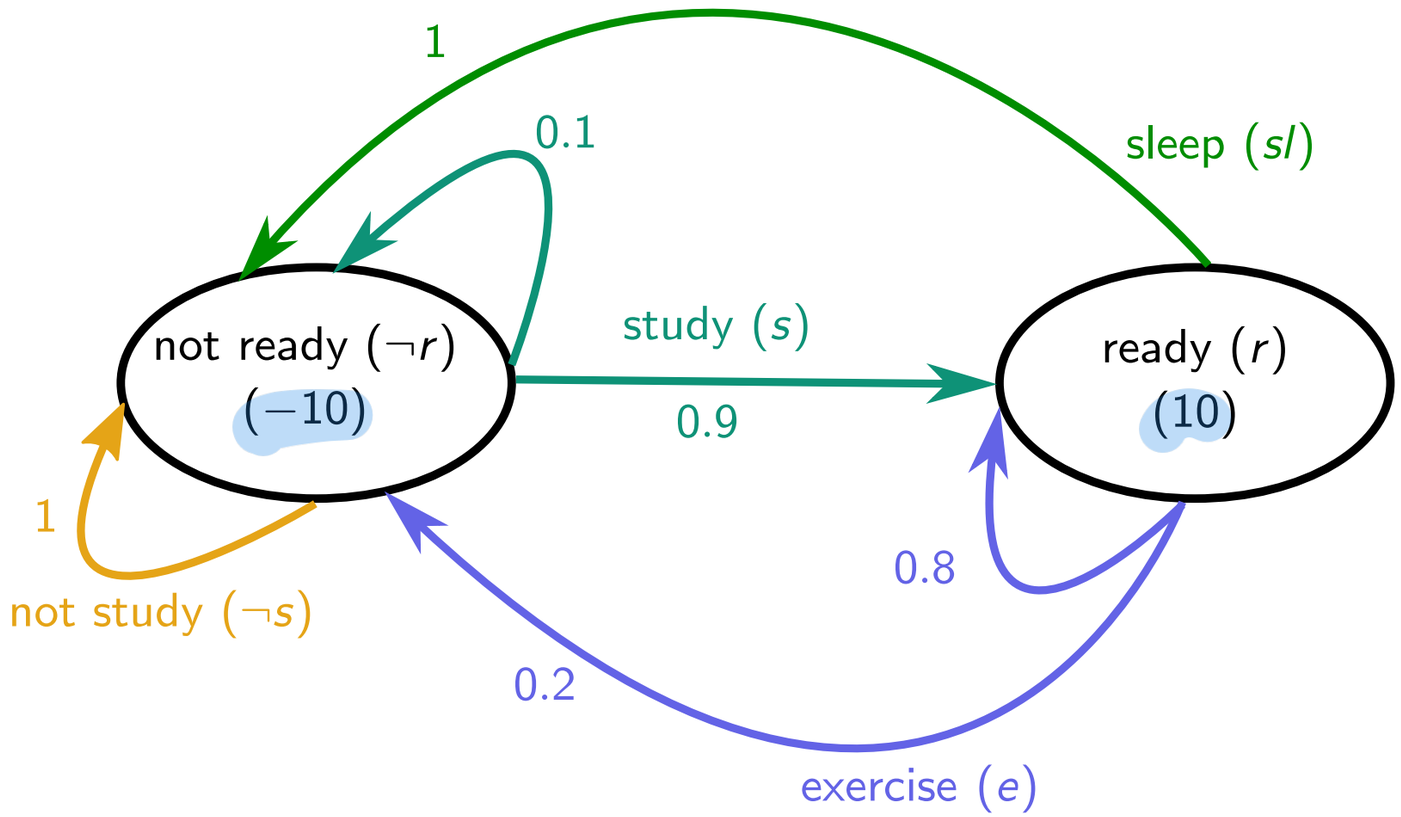
$$\begin{aligned} U^2(r,u) = \quad & R(r,u) \quad + \quad \gamma \max \quad [P((p,u)|(r,u),A) \cdot U^1(p,u) \qquad (A) \\ & + P((p,f)|(r,u),A) \cdot U^1(p,f), \\ & P((p,u)|(r,u),S) \cdot U^1(p,u) \qquad (S) \\ & + P((r,u)|(r,u),S) \cdot U^1(r,u)], \end{aligned}$$

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

$$U^2(r,u) = \quad 14.5$$

Problem 11.3: AI Exam

$$R(s) = \begin{cases} 10, & s = r \\ -10, & s = \neg r \end{cases}$$



Problem 11.3: AI Exam

$\pi_1(s)$

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 $U_0^{\pi}(r)$ and $U_0^{\pi}(\neg r)$:

$$U^{\pi}(r) = R(r) + \gamma \left[P(r|r,e) U_0^{\pi}(r) + P(\neg r|r,e) U_0^{\pi}(\neg r) \right]$$

$$= 10 + 0.9 \left[0.8\, U_0^{\pi}(r) + 0.2\, U_0^{\pi}(\neg r) \right]$$

$$10 = 0.28\, U_0^{\pi_0}(r) - 0.18\, U_0^{\pi}(\neg r)$$

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 $U_i(s) = R(s) + \gamma \sum_{s'} P(s'|s, \pi_i(s)) U_i(s')$

Compute $U_0(r)$ and $U_0(\neg r)$: $U_0^{\pi}(\neg r) = R(\neg r) + \gamma \left[P(r|\neg r, s) U_0^{\pi}(r) + P(\neg r|\neg r, s) U_0^{\pi}(\neg r) \right]$

$$= -10 + 0.9 \left[0.9\, U_0^{\pi}(r) + 0.1\, U_0^{\pi}(\neg r) \right]$$

$$-10 = 0.91\, U_0^{\pi}(\neg r) - 0.81\, U_0^{\pi}(r)$$

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 $\quad U_i(s) = R(s) + \gamma \sum_{s'} P(s'|s, \pi_i(s)) U_i(s')$

Summarize the linear equations:

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

Solution:

$$\begin{array}{rcl} U_0^{\pi}(r) & = & 66.7, \\ U_0^{\pi}(\neg r) & = & 48.4. \end{array}$$

$$V(s) = \max_{a \in A} \left(\sum_{s'} P(s' \mid s,a) \left[R(s,a,s') + \gamma V(s') \right] \right)$$

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

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

$$= \max_{a \in A} \left(\sum_{s'} P(s' \mid s,a) R(s) + \ldots \right)$$

$$= 1 \cdot R(s) + \max_{a \in A} \sum_{s'} \gamma P(s' \mid a,s) V(s')$$

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

Compute $\pi_1(\neg r)$: $\arg_a \max \left[P(r|\neg r, s)\, U_0^{\pi}(r) + P(\neg r|\neg r, s)\, U_0^{\pi}(\neg r) \; ; \; P(\neg r|\neg r, \neg s)\, U_0^{\pi}(\neg r) \right]$

$= \arg_a \max \left[0.9 \cdot 66.7 + 0.1 \cdot 48.4 \; ; \; 1 \cdot 48.4 \right]$

$= [64.87 \; ; \; 48.4] = s$

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

Compute $\pi_1(r)$:

$$\pi_1(r) = \arg\max_{a \in \{e, sl\}} \left[P(\neg r \mid r, sl)\, U_0(\neg r);\ P(r \mid r, e)\, U_0(r) + P(\neg r \mid r, e)\, U_0(\neg r) \right]$$

$$= \arg\max_a \left[1 \cdot 48.4;\ 0.8 \cdot 66.7 + 0.2 \cdot 48.4 \right]$$

$$= \arg\max_a \left[48.4;\ 63.04 \right] = e$$