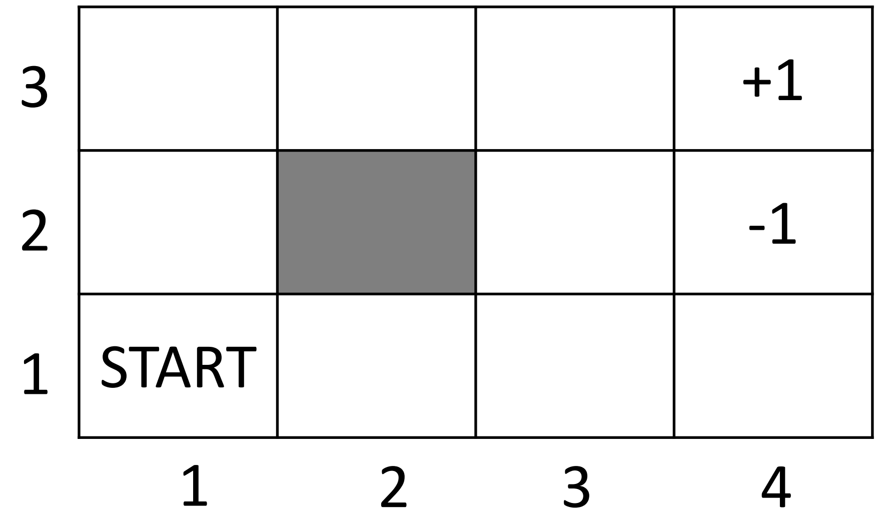
Markov Decision Processes
Part 2: Utilities of States, the Bellman Equation

CSE 4309 – Machine Learning
Vassilis Athitsos
Computer Science and Engineering Department
University of Texas at Arlington

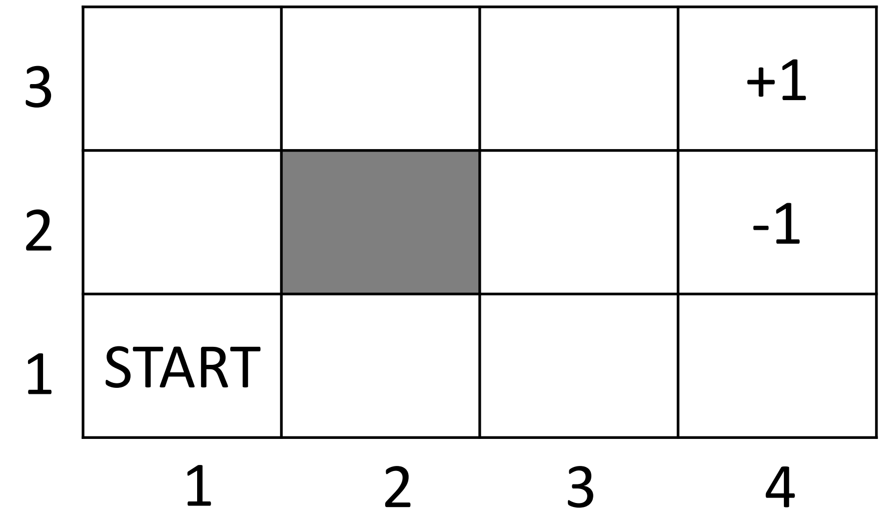
Review: MDPs



- A **<u>Markov Decision Process</u>** (MDP) is a sequential decision problem, with some additional assumptions.
- Assumption 1: **Markovian Transition Model**.
  - $p(s' \mid s, a, H) = p(s' \mid s, a)$
    - Given the last state, the history does not matter.
- Assumption 2: **Discounted Additive Rewards**.

$$U_h(s_0,\ s_1, \ldots, s_T) = \sum_{t=0}^{T} \gamma^t R(s_t)$$

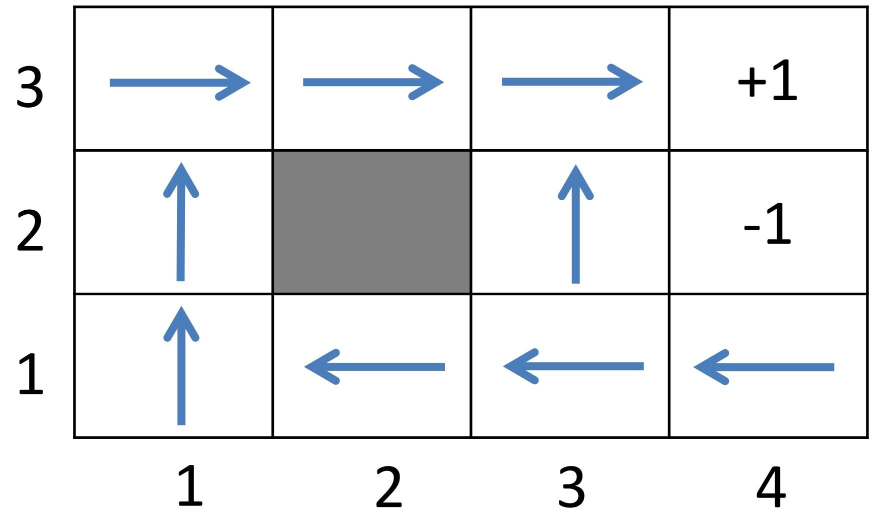
Review: Policy

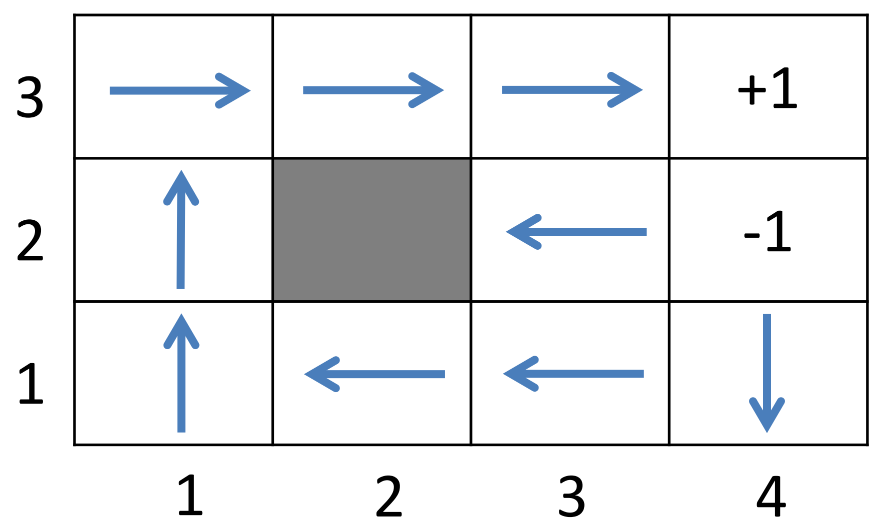


- When we have an MDP process, the problem that we typically want to solve is to find an optimal **policy**.
- A policy $\pi(s)$ is a function mapping states to actions.
  - When the agent is at state s, the policy tells the agent to perform action $\pi(s)$.
- An optimal policy π^* is a policy that maximizes the **expected utility**.
  - The expected utility of a policy π is the average utility attained per mission, when the agent carries out an infinite number of missions following that policy π.

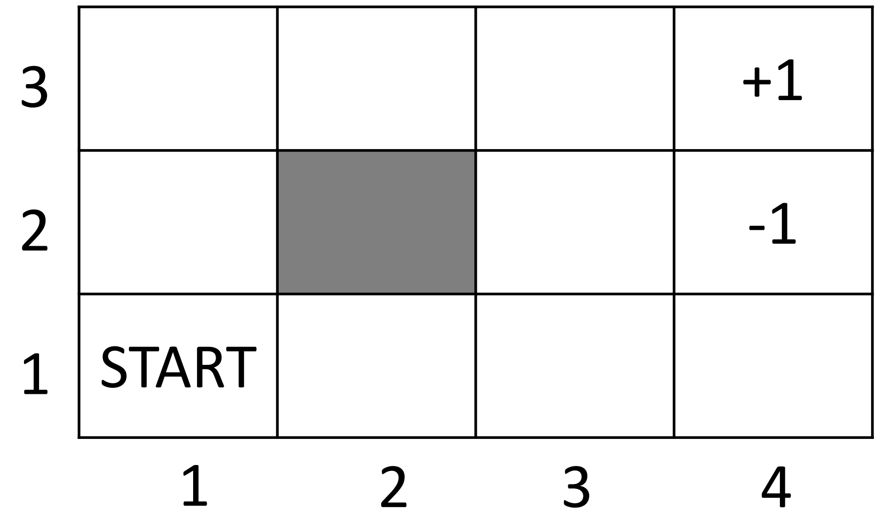
Policy Examples

- Top figure: the optimal policy for:
  - $R(s) = -0.04$ for non-terminal states s.
  - $\gamma = 1$.
- Bottom figure: the optimal policy for:
  - $R(s) = -0.02$ for non-terminal states s.
  - $\gamma = 1$.
- Changing $R(s)$ from -0.04 to -0.02 makes longer sequences less costly.





Utility of a State



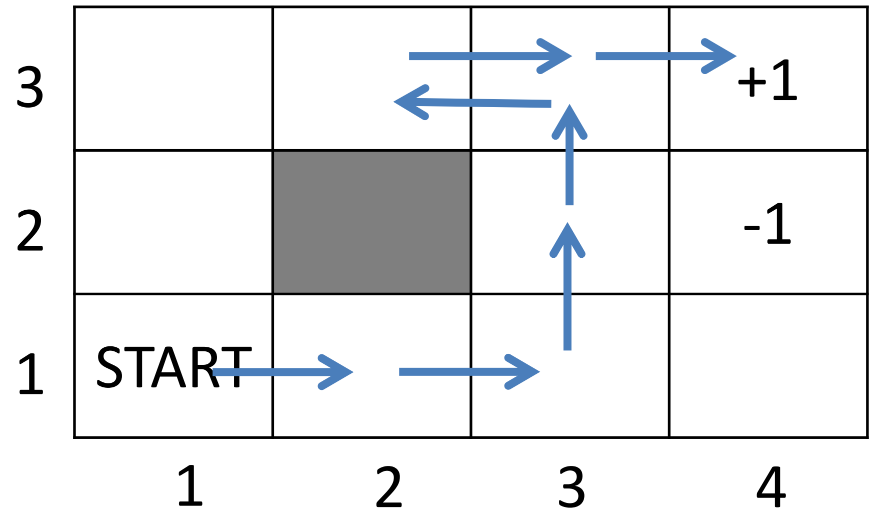
- In order to figure out how to compute the optimal policy π^*, we need to study some of its properties.
- We define the utility $U(s_0)$ of a state s_0 as the expected value $\mathrm{E}[U_h(s_0, s_1, \ldots, s_T)]$, measured over all possible sequences s_0, $s_1, \ldots, s_T$ that can happen if the agent follows policy π^*.
  - Obviously, we assume that the agent knows π^*, in order to follow that policy.
- If the agent follows a specific policy π^*, why are there multiple possible sequences of future states?
  - $\pi^*(s)$ tells us the action the agent will take at any state s, but, remember, the result of the action is **non-deterministic**.
  - The probability that action $\pi^*(s)$ will lead to state s' is modeled by the state transition function $p(s' \mid s, \pi^*(s))$

A Note on Notation

3				+1
2				-1
1	START			
	1	2	3	4

- Note that we have defined three different utility-related functions.
- $R(s_0)$ is the immediate reward obtained when the agent reaches state s_0.
- $U_h(s_0,\ s_1, \dots, s_T)$ is the (possibly discounted) sum of rewards of states $s_0,\ s_1, \dots, s_T$.
  - Thus, $U_h(s_0) = R(s_0)$, since $U_h(s_0) = \sum_{t=0}^{0} \gamma^t R(s_t)$
- $U(s_0)$ is the expected value $\mathrm{E}[U_h(s_0,\ s_1, \dots, s_T)]$, measured over **<u>all possible sequences</u>** $s_0,\ s_1, \dots, s_T$ that can happen if the agent is at state s_0 and **<u>the agent follows the optimal policy $\boldsymbol{\pi}^*$</u>**.

Utility of a Sequence



- Suppose that any non-terminal state yields a reward of -0.04.
- Suppose that $\gamma = 0.9$.
- Let's consider a state sequence $\mathbf{S}$ defined as:

$$\mathbf{S} = \big((1,1),\ (1,2), (1,3), (2,3), (3,3), (3,2), (3,3), (4,3)\big)$$

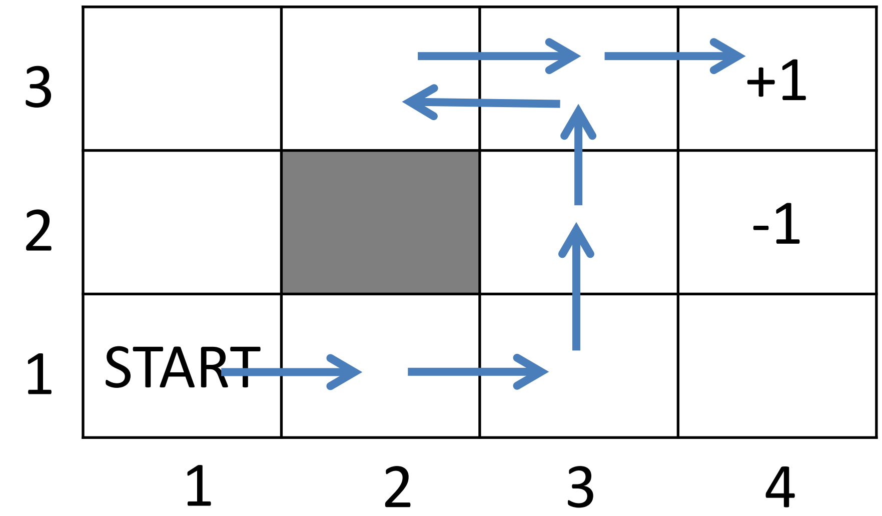
- How do we compute $U_h(\mathbf{S})$?

$$U_h(\mathbf{S}) = \sum_{t=0}^{T} \gamma^t R(s_t)$$

$$= 0.9^0 R(1,1) + 0.9^1 R(1,2) + 0.9^2 R(1,3) + 0.9^3 R(2,3) +$$
$$0.9^4 R(3,3) + 0.9^5 R(3,2) + 0.9^6 R(3,3) + 0.9^7 R(4,3)$$

$$= 1 * (-0.04) + 0.9 * (-0.04) + 0.81 * (-0.04) + 0.73 * (-0.04) +$$
$$0.66 * (-0.04) + 0.59 * (-0.04) + 0.53 * (-0.04) + 0.48 * 1$$

Utility of a Sequence



- Suppose that any non-terminal state yields a reward of -0.04.
- Suppose that $\gamma = 0.9$.
- Let's consider a state sequence $\mathbf{S}$ defined as:

$$\mathbf{S} = \big((1,1),\ (1,2), (1,3), (2,3), (3,3), (3,2), (3,3), (4,3)\big)$$

- How do we compute $U_h(\mathbf{S})$?

$$U_h(\mathbf{S}) = \sum_{t=0}^{T} \gamma^t R(s_t) = 0.27$$

Utility of a State

2	+1	
1	START	-1
	1	2

- What is the utility of state $(2,1)$ in this toy example?
- $\mathrm{U}(s_0) = \mathrm{E}[U_h(s_0,\ s_1, \ldots, s_T)]$, measured over all possible sequences we can get if:
  - We start from s_0.
  - We continue till we reach a terminal state.
  - We follow the optimal policy π^*.
- If we start with $s_0 = (2,1)$, what are all possible sequences $(s_0,\ s_1, \ldots, s_T)$?
- Since $(2,1)$ is a terminal state, the only possible sequence is $((2,1))$.
- Thus, $\mathrm{U}(2,1) = U_h((2,1)) = ???$

Utility of a State

2	+1	
1	START	-1
	1	2

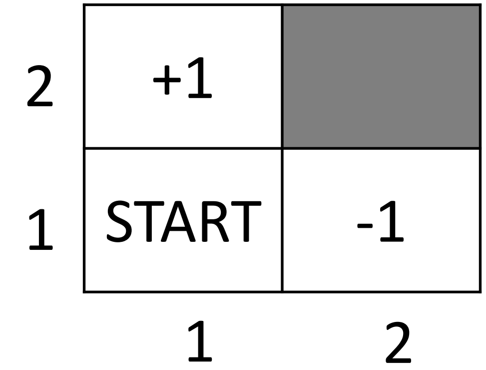
- What is the utility of state $(2,1)$ in this toy example?
- $\mathrm{U}(s_0) = \mathrm{E}[U_h(s_0,\ s_1, \ldots, s_T)]$, measured over all possible sequences we can get if:
  - We start from s_0.
  - We continue till we reach a terminal state.
  - We follow the optimal policy π^*.
- If we start with $s_0 = (2,1)$, what are all possible sequences $(s_0,\ s_1, \ldots, s_T)$?
- Since $(2,1)$ is a terminal state, the only possible sequence is $((2,1))$.
- Thus, $\mathrm{U}(2,1) = U_h((2,1)) = 1$.

Utility of a State

2	+1	
1	START	-1
	1	2

- What is the utility of state $(1,2)$ in this toy example?
- $\mathrm{U}(s_0) = \mathrm{E}[U_h(s_0,\ s_1, \ldots, s_T)]$, measured over all possible sequences we can get if:
  - We start from s_0.
  - We continue till we reach a terminal state.
  - We follow the optimal policy π^*.
- Since $(1,2)$ is a terminal state, the only possible sequence is $((1,2))$.
- Thus, $\mathrm{U}(1,2) = U_h((1,2)) = -1$.

Utility of a State



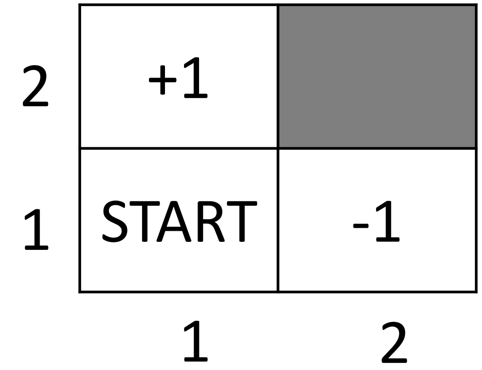
- What is the utility of state $(1{,}1)$?
- $\mathrm{U}(s_0) = \mathrm{E}[U_h(s_0,\ s_1, \ldots, s_T)]$, measured over all possible sequences we can get if:
  - We start from s_0.
  - We continue till we reach a terminal state.
  - We follow the optimal policy π^*.
- $(1{,}1)$ is not a terminal state.
- How many possible sequences are there?

Utility of a State

2	+1	
1	START	-1
	1	2

- What is the utility of state $(1,1)$?
- $\mathrm{U}(s_0) = \mathrm{E}[U_h(s_0,\ s_1, \ldots, s_T)]$, measured over all possible sequences we can get if:
  - We start from s_0.
  - We continue till we reach a terminal state.
  - We follow the optimal policy π^*.
- $(1,1)$ is not a terminal state.
- There are infinitely many possible sequences. Assuming $\gamma = 0.9$:
  - $\big((1,1), (2,1)\big)$, with utility $U_h = -0.04 + 0.9 * 1 = 0.86$
  - $\big((1,1), (1,2)\big)$, with utility $U_h = -0.04 + 0.9 * (-1) = -0.94$
  - $\big((1,1), (1,1), (2,1)\big)$, with $U_h = -0.04 + 0.9 * (-0.04) + 0.81 * 1 = 0.84$
  - $\big((1,1), (1,1), (1,2)\big)$, $U_h = -0.04 + 0.9 * (-0.04) + 0.81 * (-1) = -0.89$
  - …

Utility of a State



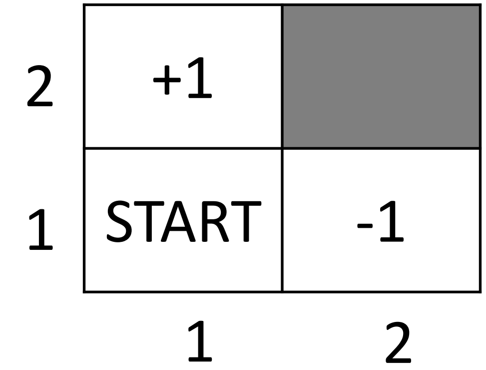
- What is the utility of state $(1,1)$?
- $\mathrm{U}(s_0) = \mathrm{E}[U_h(s_0,\ s_1, \ldots, s_T)]$.
- There are infinitely many possible sequences. Assuming $\gamma = 0.9$:
  - $\big((1,1),(2,1)\big)$, with utility $U_h = -0.04 + 0.9 * 1 = 0.86$
  - $\big((1,1),(1,2)\big)$, with utility $U_h = -0.04 + 0.9 * (-1) = -0.94$
  - $\big((1,1),(1,1),(2,1)\big)$, with $U_h = -0.04 + 0.9 * (-0.04) + 0.81 * 1 = 0.84$
  - $\big((1,1),(1,1),(1,2)\big)$, $U_h = -0.04 + 0.9 * (-0.04) + 0.81 * (-1) = -0.89$
  - $\big((1,1),(1,1),(1,1),(2,1)\big)$
  - $\big((1,1),(1,1),(1,1),(1,2)\big)$
  - $\big((1,1),(1,1),(1,1),(1,1),(2,1)\big)$
  - $\big((1,1),(1,1),(1,1),(1,1),(1,2)\big)$
  - ….

Utility of a State

2	+1	
1	START	-1
	1	2

- What is the utility of state $(1,1)$?
- $\mathrm{U}(s_0) = \mathrm{E}[U_h(s_0,\ s_1, \ldots, s_T)]$, measured over all possible sequences we can get if:
  - We start from s_0.
  - We continue till we reach a terminal state.
  - We follow the optimal policy π^*.
- There are infinitely many possible sequences. Assuming $\gamma = 0.9$:
  - $\big((1,1), (2,1)\big)$, with utility $U_h = -0.04 + 0.9 * 1 = 0.86$
  - $\big((1,1), (1,2)\big)$, with utility $U_h = -0.04 + 0.9 * (-1) = -0.94$
  - $\big((1,1), (1,1), (2,1)\big)$, with $U_h = -0.04 + 0.9 * (-0.04) + 0.81 * 1 = 0.84$
  - $\big((1,1), (1,1), (1,2)\big)$, $U_h = -0.04 + 0.9 * (-0.04) + 0.81 * (-1) = -0.89$
  - …
- How can we measure the expected value of U_h over this infinite set of sequences?

Utility of a State



- What is the utility of state $(1,1)$?
- $\mathrm{U}(s_0) = \mathrm{E}[U_h(s_0,\ s_1, \ldots, s_T)]$, measured over all possible sequences we can get if:
  - We start from s_0.
  - We continue till we reach a terminal state.
  - We follow the optimal policy π^*.
- There are infinitely many possible sequences.
- $\mathrm{E}[U_h(s_0,\ s_1, \ldots, s_T)]$ is a weighted average, where the weight of each state sequence is the probability of that sequence, **assuming that we are following the optimal policy $\boldsymbol{\pi}^*$**.
- What is the optimal policy $\boldsymbol{\pi}^*$?
  - It is the one that maximizes $\mathrm{U}(s)$ for all states s.
- It looks like a chicken-and-egg problem: we must know $\boldsymbol{\pi}^*$ to compute $\mathrm{U}(s)$, and we must know values $\mathrm{U}(s)$ to compute $\boldsymbol{\pi}^*$.

Utility of a State

2	+1	
1	START	-1
	1	2

- What is the utility of state $(1{,}1)$?
- $\mathrm{U}(s_0) = \mathrm{E}[U_h(s_0,\ s_1, \ldots, s_T)]$.
- Suppose that, for state $(1{,}1)$, the optimal action is "up".
  - We will prove that "up" is indeed optimal, a bit later.
- If the agent follows the optimal policy then, after one "up" action:
  - With probability 0.8 the agent gets to state $(2{,}1)$.
    $$U_h\big((1{,}1), (2{,}1)\big) = -0.04 + 0.9 * 1 = 0.86$$
  - With probability 0.1 the agent gets to state $(1{,}2)$.
    $$U_h\big((1{,}1), (1{,}2)\big) = -0.04 + 0.9 * (-1) = -0.94$$
  - With probability 0.1, the agent stays at state $(1{,}1)$.
- So: $\mathrm{U}(1{,}1) = \mathrm{E}\big[U_h\big((1{,}1),\ s_1, \ldots, s_T\big)\big]$
  $$= 0.8 * 0.86 + 0.1 * (-0.94) + 0.1 * X$$
- In the above, X is the expected utility if $s_0 = s_1 = (1{,}1)$.
  - Let's see how to compute X.

Utility of a State

2	+1	
1	START	-1
	1	2

- What is the utility of state $(1,1)$?
- $\mathrm{U}(s_0) = \mathrm{E}[U_h(s_0, s_1, \ldots, s_T)]$.
- Suppose that $s_0 = s_1 = (1,1)$.
- What is the expected utility in that case?
- $\mathrm{E}\left[U_h\left((1,1), (1,1), s_2, \ldots, s_T\right)\right]$ can be decomposed as:
  - The reward for state s_0, which is known: $\mathrm{R}(1,1) = -0.04$
  - The expected value of the rewards for states $s_1 = (1,1), s_2, \ldots, s_T$, which will be $\mathrm{E}[\gamma \mathrm{R}(1,1) + \gamma^2 \mathrm{R}(s_2) + \gamma^3 \mathrm{R}(s_3) + \cdots + \gamma^T \mathrm{R}(s_T)]$.
- So: $\mathrm{E}\left[U_h\left((1,1), (1,1), s_2, \ldots, s_T\right)\right] =$
  $-0.04 + \mathrm{E}[\gamma \mathrm{R}(1,1) + \gamma^2 \mathrm{R}(s_2) + \gamma^3 \mathrm{R}(s_3) + \cdots + \gamma^T \mathrm{R}(s_T)] =$
  $-0.04 + \gamma \mathrm{E}[\mathrm{R}(1,1) + \gamma \mathrm{R}(s_2) + \gamma^2 \mathrm{R}(s_3) + \cdots + \gamma^{T-1} \mathrm{R}(s_T)]$
- The expression highlighted in red is the expected utility over all sequences starting at state $(1,1)$, which is the definition of $\mathrm{U}(1,1)$.

Utility of a State

2	+1	
1	START	-1
	1	2

- What is the utility of state $(1,1)$?
- $\mathrm{U}(s_0) = \mathrm{E}[U_h(s_0,\ s_1, \ldots, s_T)]$.
- Suppose that $s_0 = s_1 = (1,1)$.
- What is the expected utility in that case?
- $\mathrm{E}\big[U_h\big((1,1),(1,1),s_2,\ldots,s_T\big)\big]$ can be decomposed as:
  - The reward for state s_0, which is known: $\mathrm{R}(1,1) = -0.04$
  - The expected value of the rewards for states $s_1 = (1,1), s_2, \ldots, s_T$, which will be $\mathrm{E}[\gamma\mathrm{R}(1,1) +\ \gamma^2\mathrm{R}(s_2) + \gamma^3\mathrm{R}(s_3) + \cdots + \gamma^T\mathrm{R}(s_T)]$.
- So: $\mathrm{E}\big[U_h\big((1,1),(1,1),s_2,\ldots,s_T\big)\big] =$

$$-0.04 + \mathrm{E}[\gamma\mathrm{R}(1,1) +\ \gamma^2\mathrm{R}(s_2) + \gamma^3\mathrm{R}(s_3) + \cdots + \gamma^T\mathrm{R}(s_T)] =$$

$$-0.04 + \gamma\mathrm{E}[\mathrm{R}(1,1) + \gamma\mathrm{R}(s_2) + \gamma^2\mathrm{R}(s_3) + \cdots + \gamma^{T-1}\mathrm{R}(s_T)] =$$

$$-0.04 + \gamma\mathrm{U}(1,1)$$

Utility of a State

2	+1	
1	START	-1
	1	2

- What is the utility of state $(1,1)$?
- $\mathrm{U}(s_0) = \mathrm{E}[U_h(s_0,\ s_1, \dots, s_T)]$.
- If we combine the results from the previous slides, we get:

$$\mathrm{U}(1,1) = \mathrm{E}\big[U_h\big((1,1),\ s_1, \dots, s_T\big)\big]$$
$$= 0.8 * 0.86 + 0.1 * (-0.94) + 0.1 * \mathrm{E}\big[U_h\big((1,1),(1,1), s_2, \dots, s_T\big)\big]$$
$$= 0.8 * 0.86 + 0.1 * (-0.94) + 0.1 * (-0.04 + \gamma \mathrm{U}(1,1))$$

- This is an equation with one unknown, $\mathrm{U}(1,1)$. We can solve as:

$$\mathrm{U}(1,1) = 0.8 * 0.86 + 0.1 * (-0.94) + 0.1 * \big(-0.04 + \gamma \mathrm{U}(1,1)\big) \Rightarrow$$

$$\mathrm{U}(1,1) = 0.594 - 0.004 + 0.1 \ * 0.9 * \mathrm{U}(1,1) \Rightarrow$$

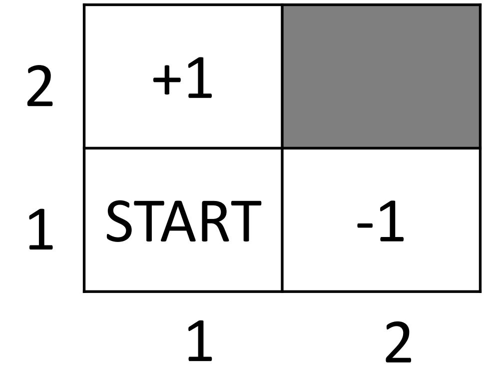
$$0.91 * \mathrm{U}(1,1) = 0.590 \Rightarrow \mathbf{U(1,1) = 0.648}$$

Utility of a State

2	+1	
1	START	-1
	1	2

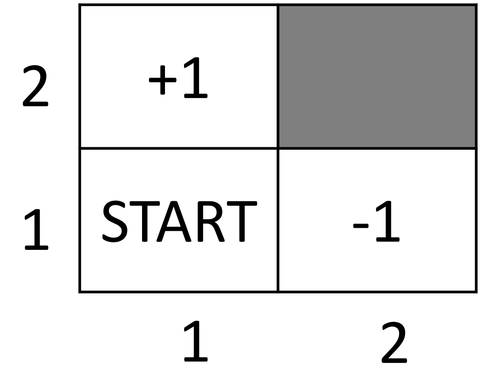
- What is the utility of state $(1,1)$?
- We have shown that, if the optimal action for state $(1,1)$ is "up", then $\mathrm{U}(1,1) = 0.648$.
- Using the exact same approach, we can measure $\mathrm{U}(1,1)$ under the other three assumptions:
  - That the optimal action for state $(1,1)$ is "down".
  - That the optimal action for state $(1,1)$ is "left".
  - That the optimal action for state $(1,1)$ is "right".
- If we compute the four values of $\mathrm{U}(1,1)$, obtained under each of the four assumptions, then what can we conclude?

Utility of a State



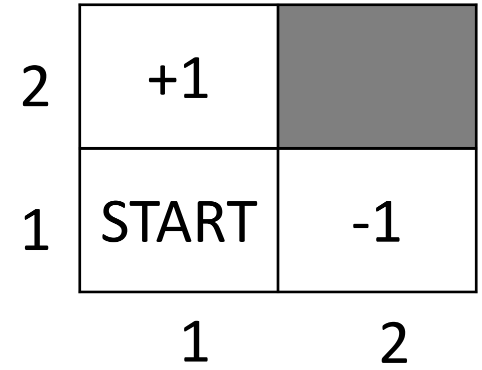
- What is the utility of state $(1,1)$?
- We have shown that, if the optimal action for state $(1,1)$ is "up", then $\mathrm{U}(1,1) = 0.648$.
- Using the exact same approach, we can measure $\mathrm{U}(1,1)$ under the other three assumptions:
  – That the optimal action for state $(1,1)$ is "down".
  – That the optimal action for state $(1,1)$ is "left".
  – That the optimal action for state $(1,1)$ is "right".
- If we compute the four values of $\mathrm{U}(1,1)$, obtained under each of the four assumptions, then what can we conclude?
  – The optimal action for $(1,1)$ is the action that leads to the highest of the four values.
  – The true value of $\mathrm{U}(1,1)$ is the highest of those four values.
- If we do the calculations, "up" is the optimal action.

Utility of a State



- What is the utility of state $(1,1)$?
  - In other words, what is the expected total reward between now and the end of the mission, if the current position is $(1,1)$?
- What is $\pi^*(1,1)$?
  - In other words, what is the optimal action to take at state $(1,1)$?
- We computed that:
  - $\mathrm{U}(1,1) = 0.648$.
  - $\pi^*(1,1) = \text{"up"}$.
- This problem was as simplified as possible, and it still took a significant amount of calculations to solve.
  - We even skipped most of the calculations, for the hypotheses that the action is "down", "left", and "right".
- Our next goal is to identify algorithms for solving such problems.

Utility of a State



- We want to identify general methods for computing:
  - The utility of all states.
  - The optimal policy π^*, which specifies for each state s the optimal action $\pi^*(s)$.
- To do that, we will revisit our solution for state $(1,1)$, and we will reformulate that solution in a way that is easier to generalize.

Utility of a State

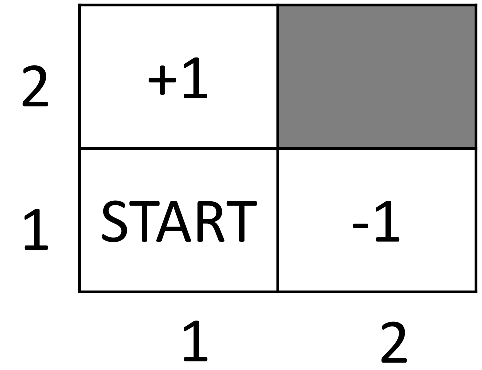
2	+1	
1	START	-1
	1	2

- We computed $\mathrm{U}(1{,}1)$ by:
  - Computing, for each possible action a that we can take at state $(1{,}1)$, the value of $\mathrm{U}(1{,}1)$ under the assumption that that action is optimal.
  - Choosing the maximum of those values as the true value of $\mathrm{U}(1{,}1)$.
- We can generalize this approach.
- First, some notation:
  - Define $A(s)$ to be the set of all actions that the agent can take at state s.
  - Define $\mathrm{U}(s, a)$ as the utility of state s **under the assumption** that $\pi^*(\mathrm{s}) = a$, i.e, the assumption that the best action at state s is a.
- Then:

$$\boxed{\mathrm{U}(s) = \max_{a \in A(s)} \{\mathrm{U}(s, a)\}}$$

$$\boxed{\pi^*(s) = \operatorname*{argmax}_{a \in A(s)} \{\mathrm{U}(s, a)\}}$$

Utility of a State



$$\mathrm{U}(s) = \max_{a \in A(s)} \{\mathrm{U}(s, a)\}$$

$$\pi^*(s) = \operatorname*{argmax}_{a \in A(s)} \{\mathrm{U}(s, a)\}$$

- To compute $\mathrm{U}((1,1), \text{"up"})$, i.e., the value of $\mathrm{U}(1,1)$ under the assumption that $\pi^*(1,1) = \text{"up"}$, we considered all possible outcomes of the "up" action:
  - With probability 0.8 the agent gets to state $(2,1)$.
  - With probability 0.1 the agent gets to state $(1,2)$.
  - With probability 0.1, the agent stays at state $(1,1)$.
- We computed the expected utility for each of those outcomes.

Utility of a State

2	+1	
1	START	-1
	1	2

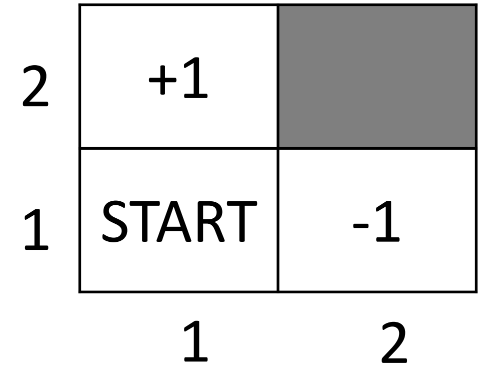
$$\mathrm{U}(s) = \max_{a \in A(s)} \{\mathrm{U}(s,a)\}$$

$$\pi^*(s) = \underset{a \in A(s)}{\operatorname{argmax}} \{\mathrm{U}(s,a)\}$$

- To compute $\mathrm{U}((1,1), \text{"up"})$, i.e., the value of $\mathrm{U}(1,1)$ under the assumption that $\pi^*(1,1) = \text{"up"}$, we considered all possible outcomes of the "up" action:
- We computed the expected utility for each of those outcomes.
- $\mathrm{U}((1,1), \text{"up"})$ was the weighted sum of the expected utility for each outcome, using as weights the probabilities of the outcomes.
- Thus:

$$\mathrm{U}(s,a) = \sum_{s'} \{p(s' \mid s,a) \mathrm{E}[U_h(s,\ s', \ldots, s_T)]\}$$

Utility of a State



$$\mathrm{U}(s) = \max_{a \in A(s)} \{\mathrm{U}(s,a)\}$$

$$\pi^*(s) = \underset{a \in A(s)}{\mathrm{argmax}} \{\mathrm{U}(s,a)\}$$

$$\mathrm{U}(s,a) = \sum_{s'} \{p(s' \mid s,a) \mathrm{E}[U_h(s,\ s', \dots, s_T)]\}$$

- Furthermore, we can decompose $\mathrm{E}[U_h(s,\ s', \dots, s_T)]$ as:

$$\mathrm{E}[U_h(s,\ s', \dots, s_T)] = R(s) + \gamma \mathrm{E}[U_h(s', \dots, s_T)] = R(s) + \gamma U(s').$$

- Therefore: $$\mathrm{U}(s,a) = R(s) + \gamma \sum_{s'} \{p(s' \mid s,a) U(s')\}$$

The Bellman Equation

2	+1	
1	START	-1
	1	2

$$\mathrm{U}(s) = \max_{a \in A(s)} \{\mathrm{U}(s,a)\}$$

$$\pi^*(s) = \underset{a \in A(s)}{\operatorname{argmax}} \{\mathrm{U}(s,a)\}$$

$$\mathrm{U}(s,a) = R(s) + \gamma \sum_{s'} [p(s' | s,a) U(s')]$$

- Combining these equations together, we get:

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

- This equation is called the <u>**Bellman equation**</u>.

The Bellman Equation

2	+1	
1	START	-1
	1	2

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

- For each state s, we get a Bellman equation.
- If our environment has N states, we need to solve a system of N Bellman equations.
- In this system of equations, there is a total of N unknowns:
  - The N values $\mathrm{U}(s)$.
- There is an iterative algorithm for solving this system of equations, called the **<u>value iteration algorithm</u>**. This is what we will study next.