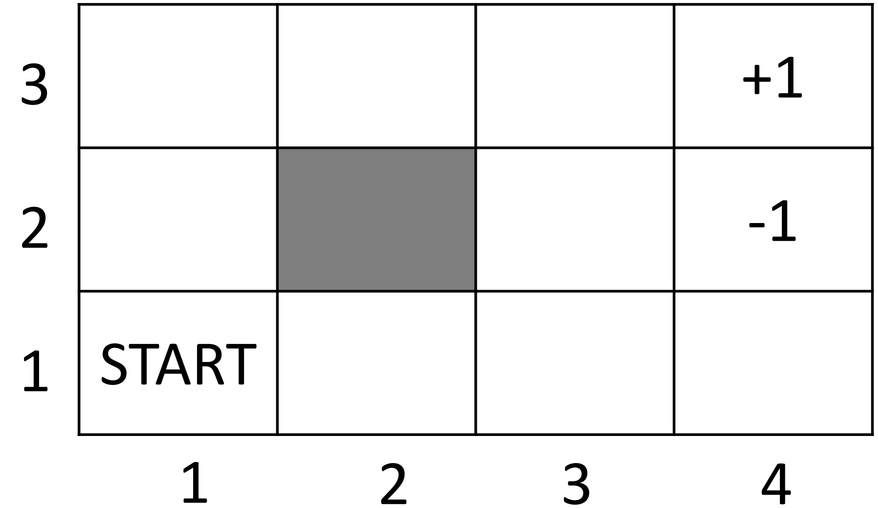
Markov Decision Processes

Part 1: Basic Definitions

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

A Sequential Decision Problem

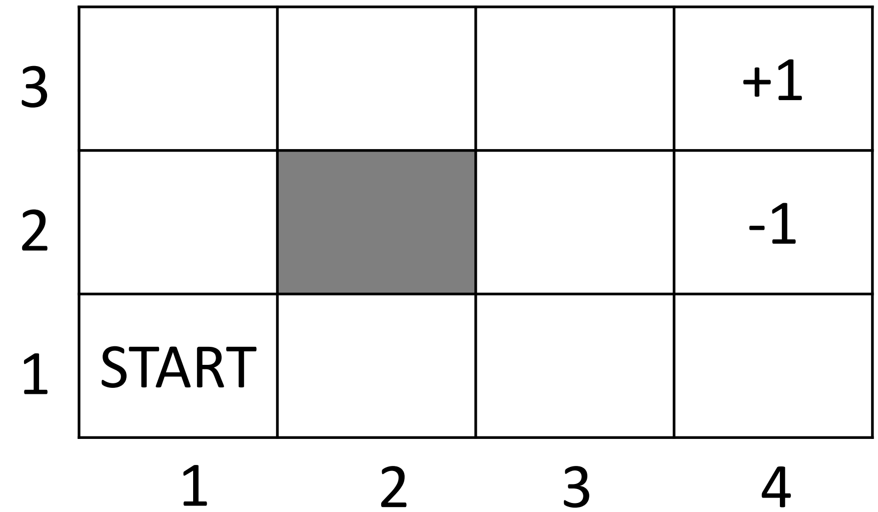


This example is taken from:

S. Russell and P. Norvig,
"Artificial Intelligence: A Modern Approach",
third edition (2009), Prentice Hall.

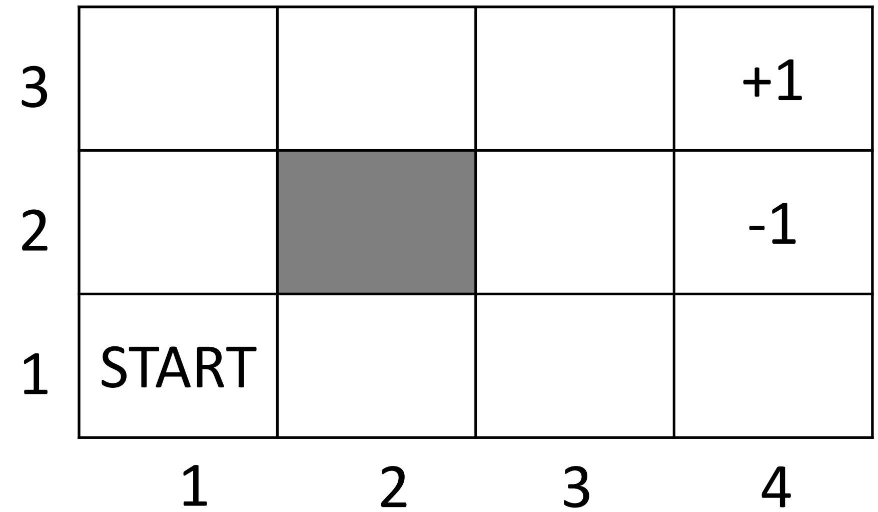
- We have an environment that is a 3×4 grid.
- We have an agent, that starts at position $(1,1)$.
- There are (at most) four possible actions: go left, right, up, or down.
- Position $(2,2)$ cannot be reached.
- Positions are denoted as $(\mathrm{row}, \mathrm{col})$.

A Sequential Decision Problem



- Positions $(2,4)$ and $(3,4)$ are terminal.
- A **mission** is a sequence of actions, that starts with the agent at the START position, and ends with the agent at a terminal position.
  - If the agent reaches position $(3,4)$, the reward is $+1$.
  - If the agent reaches position $(2,4)$, the reward is -1 (so it is actually a penalty).
- The agent wants to maximize the total rewards gained during its mission.

A Deterministic Case



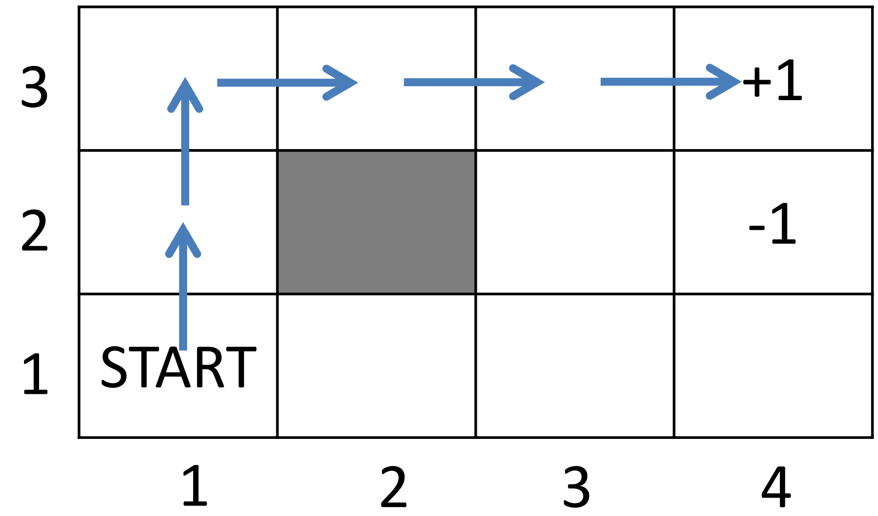
- Under some conditions, the solution for reward maximization is easy to find.
- Suppose that each action always succeeds:
  – The "go left" action takes you one position to the left.
  – The "go right" action takes you one position to the right.
  – The "go up" action takes you one position upwards.
  – The "go down" action takes you one position downwards.
- This situation is called **<u>deterministic</u>**.
  – A **<u>deterministic environment</u>** is an environment where the result of any action is known in advance.
  – A **<u>non-deterministic environment</u>** is an environment where the result of any action is not known in advance.

A Deterministic Case

3				+1
2		■		-1
1	START			
	1	2	3	4

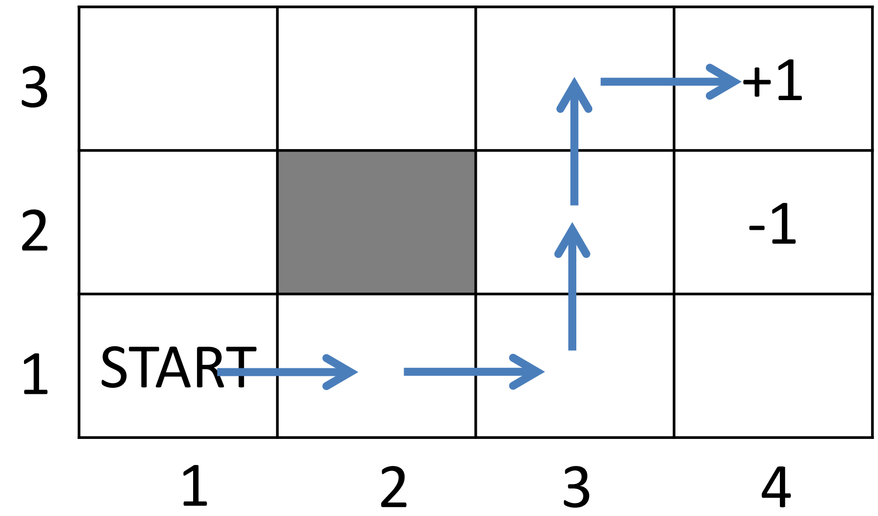
- Under some conditions, the solution for reward maximization is easy to find.
- Suppose that each action always succeeds:
  - The "go left" action takes you one position to the left.
  - The "go right" action takes you one position to the right.
  - The "go up" action takes you one position upwards.
  - The "go down" action takes you one position downwards.
- Suppose that any non-terminal state yields a reward of -0.04.
- Then, what is the optimal sequence of actions?

A Deterministic Case



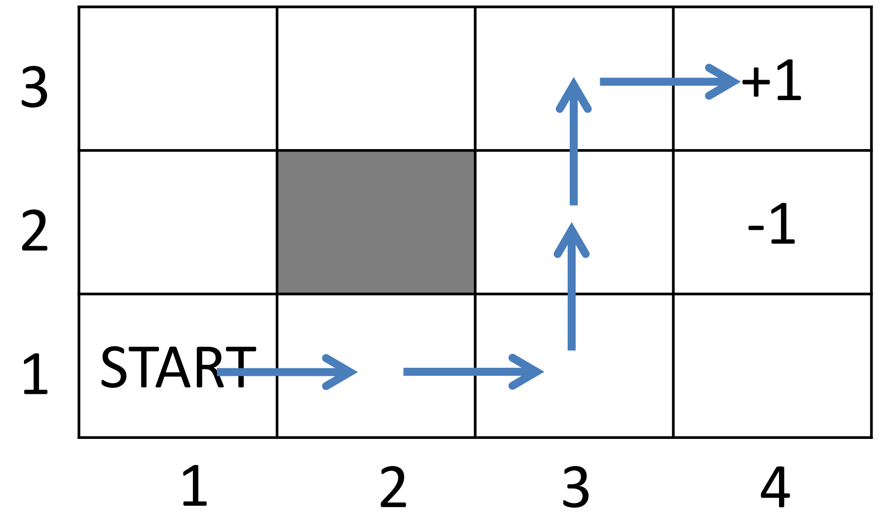
- Under some conditions, the solution for reward maximization is easy to find.
- Suppose that each action always succeeds:
  - The "go left" action takes you one position to the left.
  - The "go right" action takes you one position to the right.
  - The "go up" action takes you one position upwards.
  - The "go down" action takes you one position downwards.
- Suppose that any non-terminal state yields a reward of -0.04.
- Then, what is the optimal sequence of actions?
  - Up, up, right, right, right gets the agent from START to position $(3{,}4)$.
  - Total rewards: $1 \ - 5 * .04 = 0.8$ (five non-terminal states, including START).

A Deterministic Case



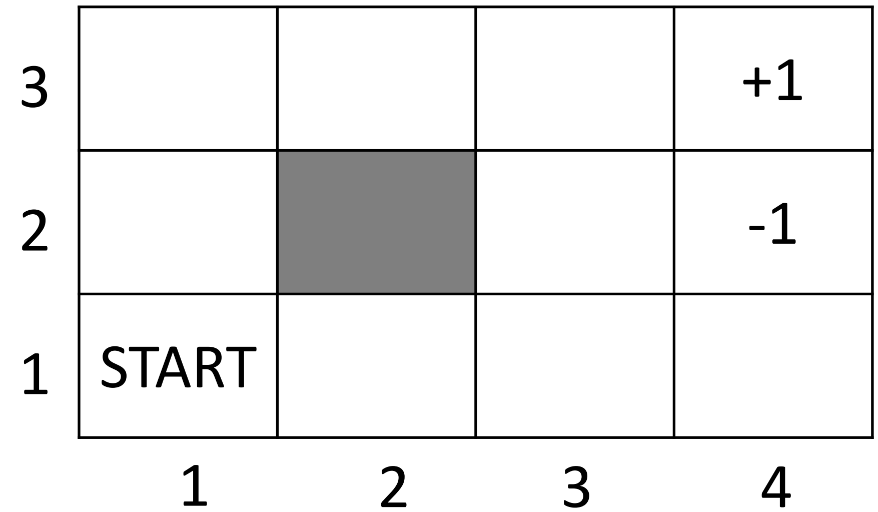
- Under some conditions, the solution for reward maximization is easy to find.
- Suppose that each action always succeeds:
  - The "go left" action takes you one position to the left.
  - The "go right" action takes you one position to the right.
  - The "go up" action takes you one position upwards.
  - The "go down" action takes you one position downwards.
- Suppose that any non-terminal state yields a reward of -0.04.
- The optimal sequence is not unique.
  - Right, right, up, up, right is also optimal.
  - Total rewards: $1\ -5*.04=0.8$ (five non-terminal states, including START).

A Deterministic Case



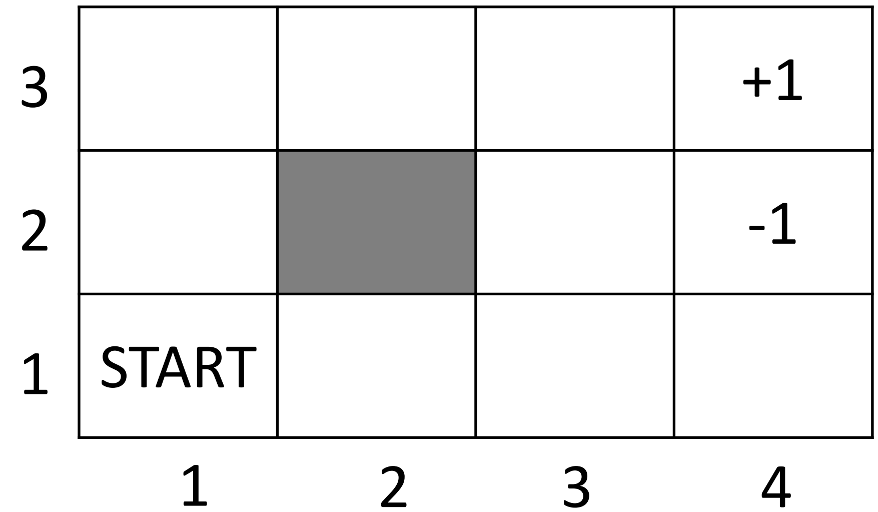
- Under some conditions, the solution for reward maximization is easy to find.
- Suppose that each action always succeeds:
  - The "go left" action takes you one position to the left.
  - The "go right" action takes you one position to the right.
  - The "go up" action takes you one position upwards.
  - The "go down" action takes you one position downwards.
- Suppose that any non-terminal state yields a reward of -0.04.
- The optimal sequence can be found using well-known algorithms such as **breadth-first search**.

A Non-Deterministic Case



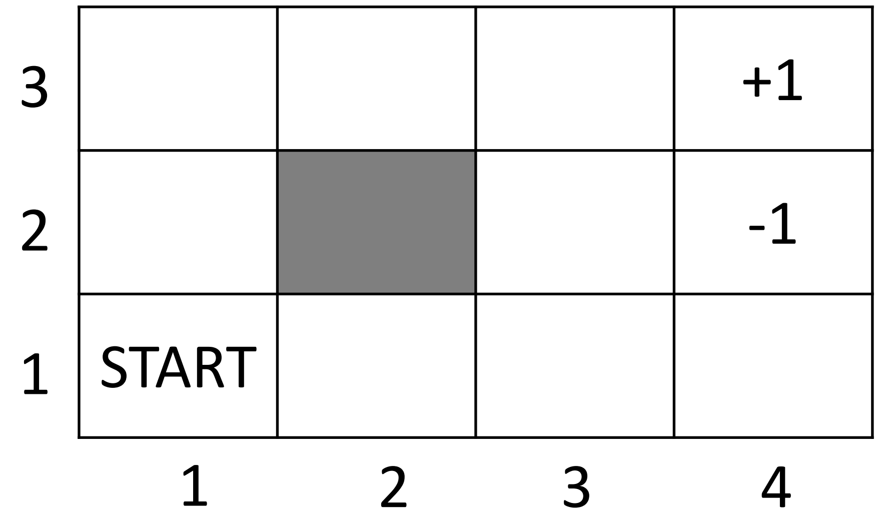
- Under some conditions, life gets more complicated.
- Suppose that each action:
  - Succeeds with probability 0.8.
  - Has a 0.2 probability of moving to a direction that differs by 90 degrees from the intended direction.
- For example: the "go up" action:
  - Has a probability of 0.8 to take the agent one position upwards.
  - Has a probability of 0.1 to take the agent one position to the left.
  - Has a probability of 0.1 to take the agent one position to the right.

A Non-Deterministic Case



- Under some conditions, life gets more complicated.
- Suppose that each action:
  - Succeeds with probability 0.8.
  - Has a 0.2 probability of moving to a direction that differs by 90 degrees from the intended direction.
- Suppose that bumping into the wall leads to not moving.
- For example:
  - The agent is at position $(1{,}1)$.
  - The agent executes the "go up" action.
  - Due to bad luck, the action moves the agent to the left.
  - The agent hits the wall, and remains at position $(1{,}1)$.

Sequential Decision Problems



- Under some conditions, life gets more complicated.
- Suppose that each action:
  - Succeeds with probability 0.8.
  - Has a 0.2 probability of moving to a direction that differs by 90 degrees from the intended direction.
- Suppose that bumping into the wall leads to not moving.
- In that case, choosing the best action to take at each position is a more complicated problem.
- A **<u>sequential decision problem</u>** consists of choosing the best sequence of actions, so as to maximize the total rewards.

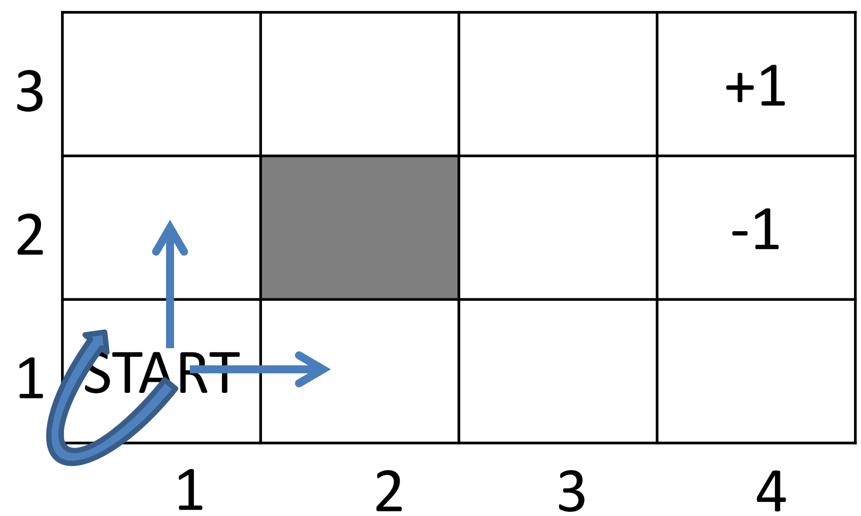
Markov Decision Processes (MDPs)

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

- A <u>**Markov Decision Process**</u> (MDP) is a sequential decision problem, with some additional assumptions.
- Assumption 1: **Markovian Transition Model**.
  - The probability $p(s' \mid s, a, H)$ is the probability of ending up in state s', given:
    - The previous state s, where the agent was taking the last action.
    - The last action a.
    - The **history** H of all prior actions and states since the start of the mission.
  - In a Markovian transition model, $p(s' \mid s, a, H) = p(s' \mid s, a)$
    - Given the last state, the history does not matter.

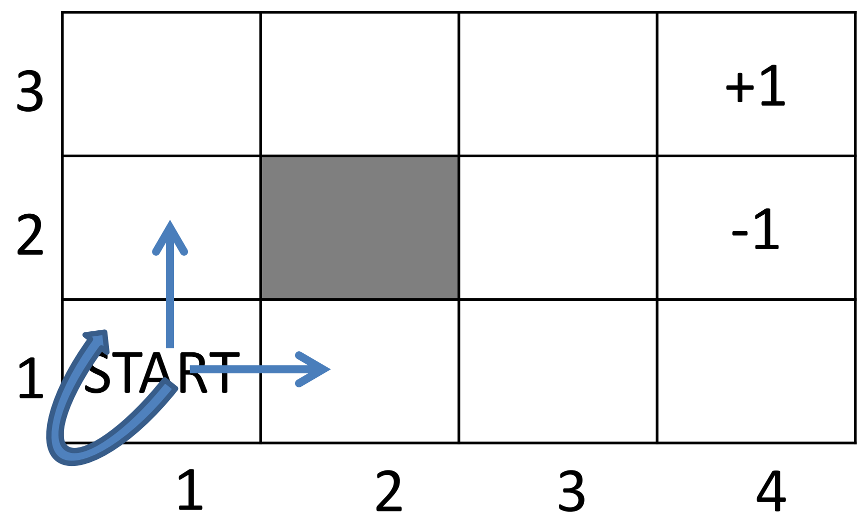
A Transition Model Example

- Suppose that each action:
  - Succeeds with probability 0.8.
  - Has a 0.2 probability of moving to a direction that differs by 90 degrees from the intended direction.
- Suppose that bumping into the wall leads to not moving.
- $p((1,1) \mid (1,1), \text{"left"}) = ???$
- $p((2,1) \mid (1,1), \text{"left"}) = ???$
- $p((1,2) \mid (1,1), \text{"left"}) = ???$



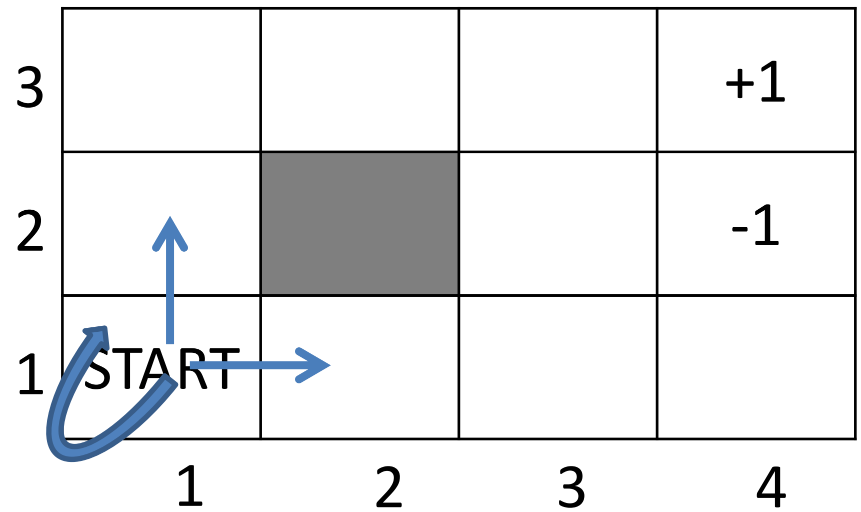
A Transition Model Example

- Suppose that each action:
  - Succeeds with probability 0.8.
  - Has a 0.2 probability of moving to a direction that differs by 90 degrees from the intended direction.
- Suppose that bumping into the wall leads to not moving.
- $p((1,1) \mid (1,1), \text{"left"}) = 0.9$
  - 0.8 chance of going left and hitting the wall.
  - 0.1 chance of going down and hitting the wall.
- $p((2,1) \mid (1,1), \text{"left"}) = 0.1$
- $p((1,2) \mid (1,1), \text{"left"}) = 0$
  - If you try to go left, you never end up going right.



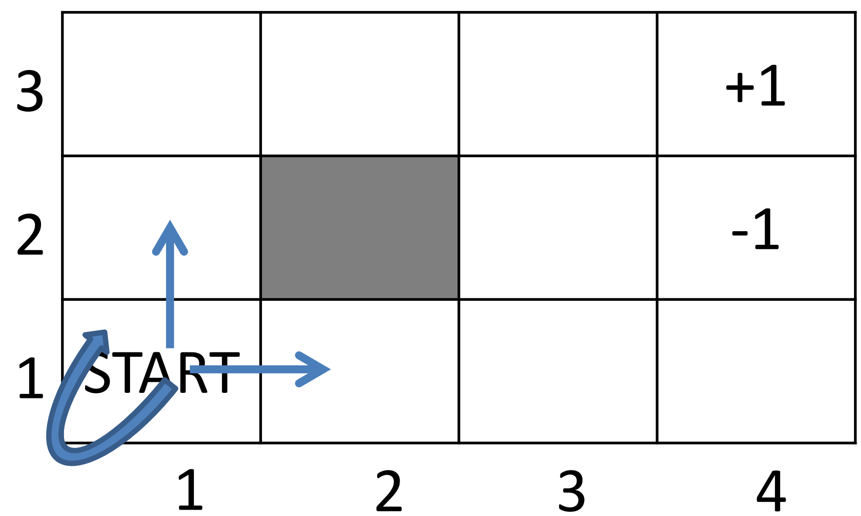
A Transition Model Example

- Suppose that each action:
  - Succeeds with probability 0.8.
  - Has a 0.2 probability of moving to a direction that differs by 90 degrees from the intended direction.
- Suppose that bumping into the wall leads to not moving.
- $p((1,1) \mid (1,1), \text{"right"}) = ???$
- $p((2,1) \mid (1,1), \text{"right"}) = ???$
- $p((1,2) \mid (1,1), \text{"right"}) = ???$



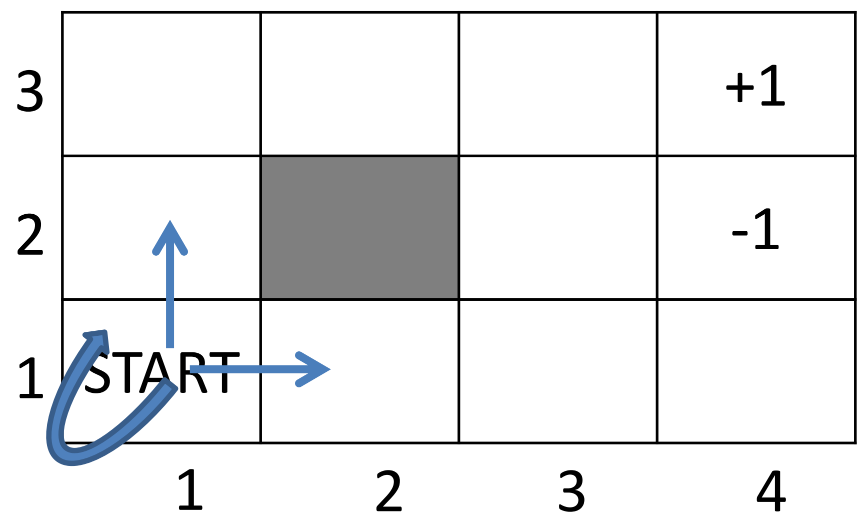
A Transition Model Example

- Suppose that each action:
  - Succeeds with probability 0.8.
  - Has a 0.2 probability of moving to a direction that differs by 90 degrees from the intended direction.
- Suppose that bumping into the wall leads to not moving.
- $p((1,1) \mid (1,1), \text{"right"}) = 0.1$
  - 0.1 chance of going down and hitting the wall.
- $p((2,1) \mid (1,1), \text{"right"}) = 0.1$
- $p((1,2) \mid (1,1), \text{"right"}) = 0.8$



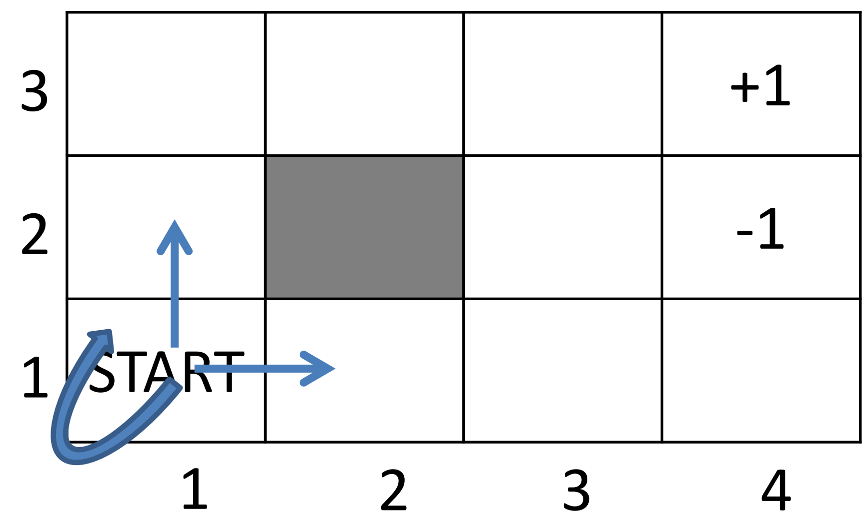
A Transition Model Example

- Suppose that each action:
  - Succeeds with probability 0.8.
  - Has a 0.2 probability of moving to a direction that differs by 90 degrees from the intended direction.
- Suppose that bumping into the wall leads to not moving.
- $p((1,1) \mid (1,1), \text{"up"}) = ???$
- $p((2,1) \mid (1,1), \text{"up"}) = ???$
- $p((1,2) \mid (1,1), \text{"up"}) = ???$



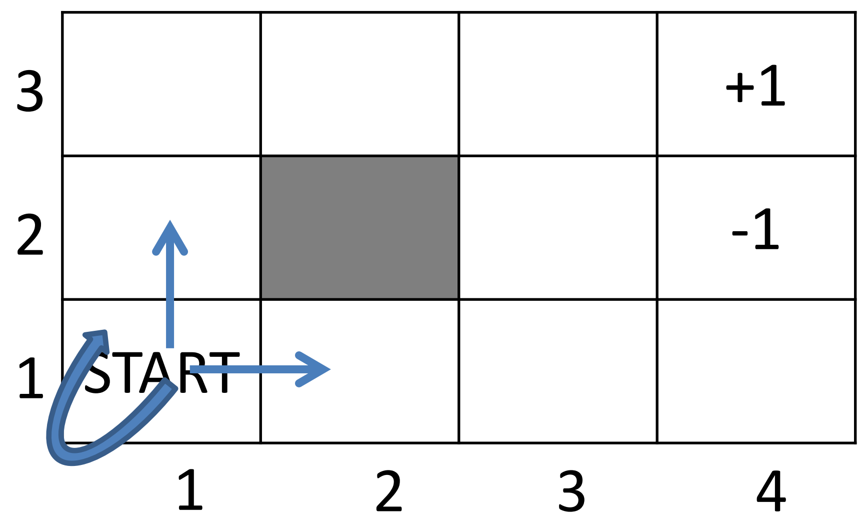
A Transition Model Example

- Suppose that each action:
  - Succeeds with probability 0.8.
  - Has a 0.2 probability of moving to a direction that differs by 90 degrees from the intended direction.
- Suppose that bumping into the wall leads to not moving.
- $p((1,1) \mid (1,1), \text{"up"}) = 0.1$
- $p((2,1) \mid (1,1), \text{"up"}) = 0.8$
- $p((1,2) \mid (1,1), \text{"up"}) = 0.1$



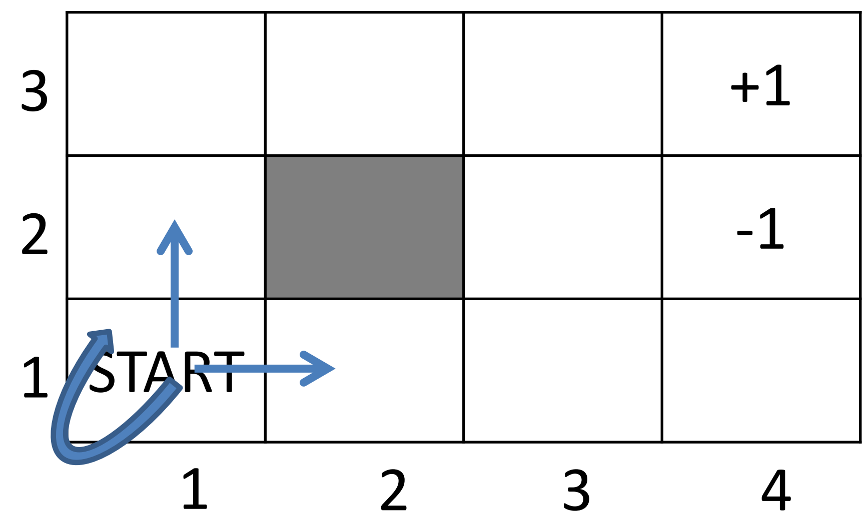
A Transition Model Example

- Suppose that each action:
  - Succeeds with probability 0.8.
  - Has a 0.2 probability of moving to a direction that differs by 90 degrees from the intended direction.
- Suppose that bumping into the wall leads to not moving.
- $p((1{,}1) \mid (1{,}1), \text{"down"}) = ???$
- $p((2{,}1) \mid (1{,}1), \text{"down"}) = ???$
- $p((1{,}2) \mid (1{,}1), \text{"down"}) = ???$



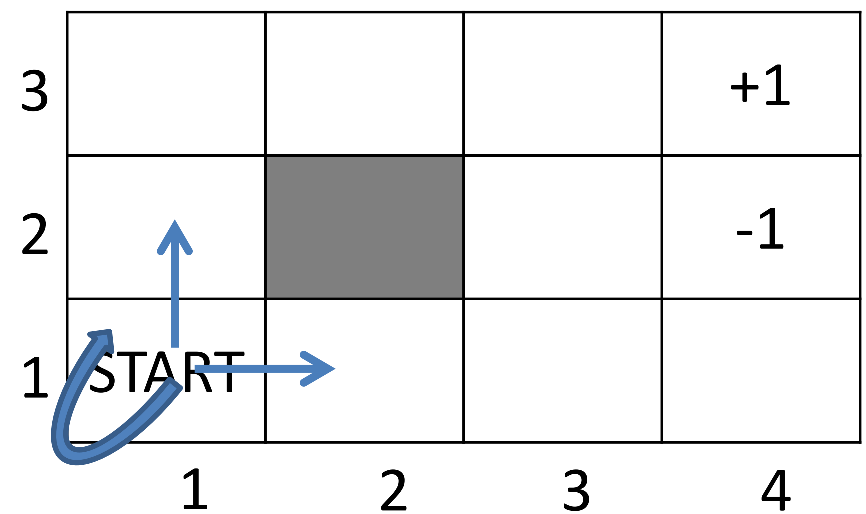
A Transition Model Example

- Suppose that each action:
  - Succeeds with probability 0.8.
  - Has a 0.2 probability of moving to a direction that differs by 90 degrees from the intended direction.
- Suppose that bumping into the wall leads to not moving.
- $p((1,1) \mid (1,1), \text{"down"}) = 0.9$
  - 0.8 chance of going down and hitting the wall.
  - 0.1 chance of going left and hitting the wall.
- $p((2,1) \mid (1,1), \text{"down"}) = 0$
  - If you try to go down, you never end up going up.
- $p((1,2) \mid (1,1), \text{"down"}) = 0.1$

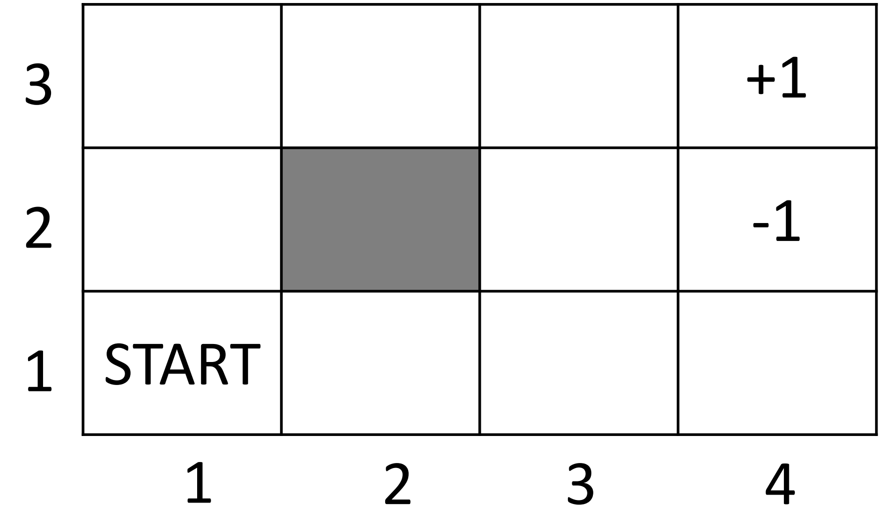


A Transition Model Example

- Suppose that each action:
  - Succeeds with probability 0.8.
  - Has a 0.2 probability of moving to a direction that differs by 90 degrees from the intended direction.
- Suppose that bumping into the wall leads to not moving.
- In a similar way, we can define all probabilities $p(s' \mid s, a)$ for:
  - Every one of the 11 legal values for state s.
  - Every one of the 2 to 4 legal values for neighbor s'.
  - Every one of the 4 legal values for action a.



Markov Decision Processes (MDPs)



- Assumption 2:
  Discounted Additive Rewards.
  - The **<u>utility</u>** U_h of a state sequence $s_0, s_1, \ldots, s_T$ is:

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

- In the above equation:
  - $R(s)$ is the **reward** function, mapping each state s to a reward.
  - γ is called the **discount factor**, $0 \leq \gamma \leq 1$.

Discounted Additive Rewards

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

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

- Suppose that $\gamma = 1$. Then:

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

- Therefore, when $\gamma = 1$, the utility function is **additive**.
  - It is simply the sum of the rewards of all states in the sequence.

Discounted Additive Rewards

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

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

- When $\gamma < 1$, the above formula indicates that the agent prefers immediate rewards over future rewards.
- The agent is at state s_0, considering what to do next.
- Sequence $s_1, \ldots, s_T$ is a possible sequence of future states.
- As t increases, γ^t decreases exponentially towards 0.
  - Thus, rewards coming far into the future (large t) are heavily discounted, with factor γ^t that quickly gets close to 0.

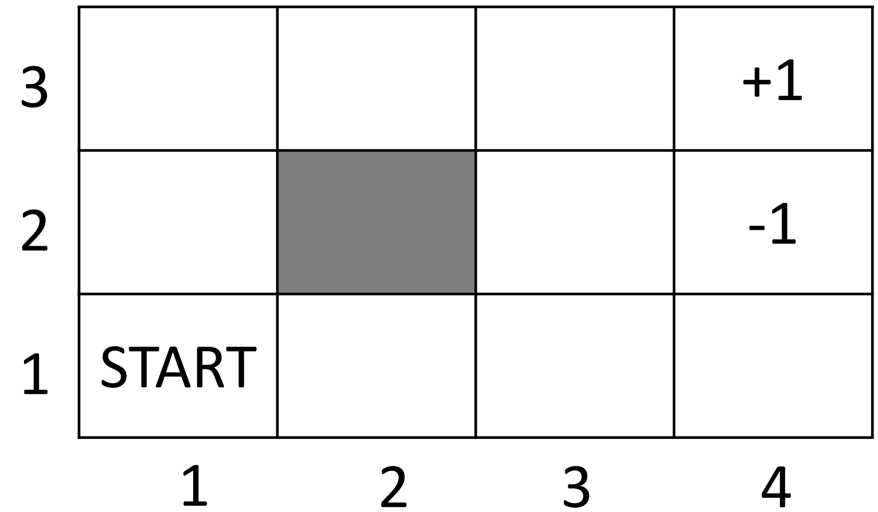
Discounted Additive Rewards

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

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

- This type of utility is called **discounted additive rewards**, since:
  - The utility is **additive**, it is a (weighted) summation of rewards attained at individual states.
  - The reward at each state s_t is **discounted** by factor γ^t.
- When $\gamma = 1$, then we simply have **additive rewards**.

Discounted Additive Rewards



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

- When does it make sense to use $\gamma < 1$, so that future rewards get discounted?
- Discounted rewards are (unfortunately?) good models of human behavior.
  - Slacking now is often preferable, versus acing the exam later.
  - The reward for slacking is relatively low but immediate.
  - The reward for acing the exam is higher, but more remote.

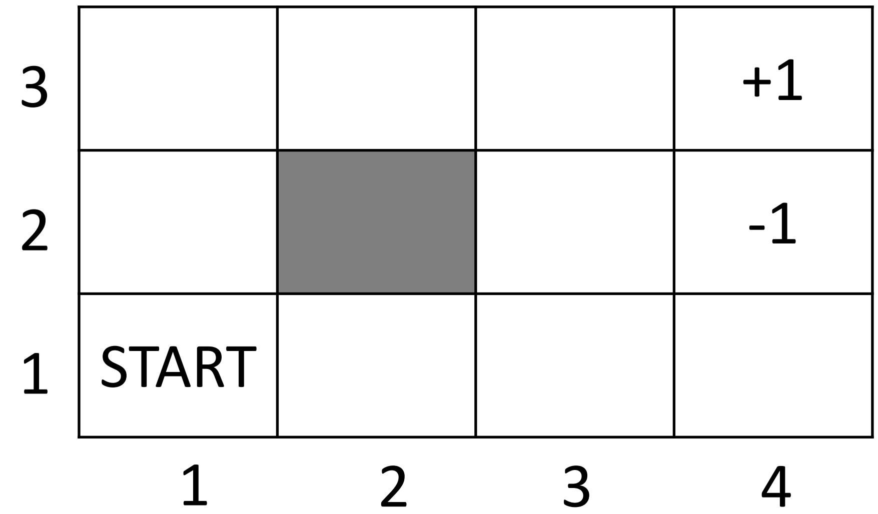
Discounted Additive Rewards

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

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

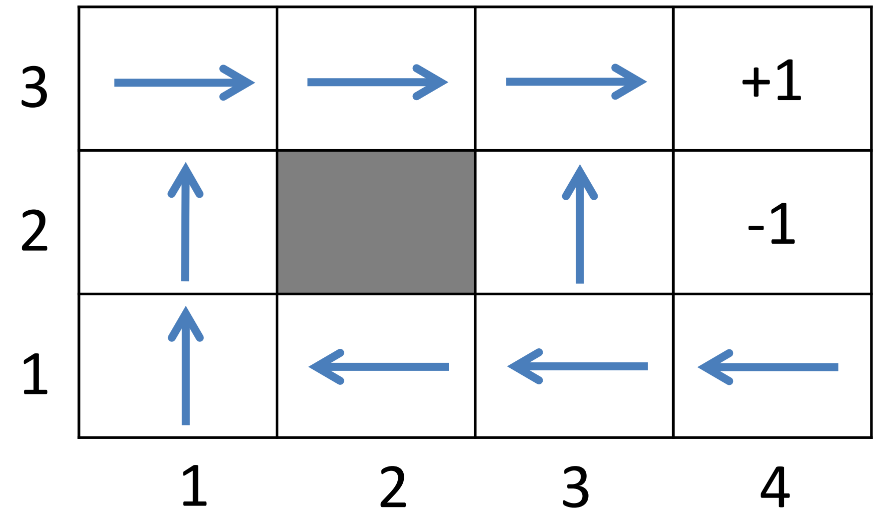
- When does it make sense to use $\gamma < 1$, so that future rewards get discounted?
- Discounted rewards are also a way to get an agent to focus on the near term.
  - We often want our intelligent agents to achieve results within a specific time window.
  - In that case, discounted rewards de-emphasize the contribution of states reached beyond that time window.

The MDP Problem



- When we have an MDP process, the problem that we typically want to solve is to find an optimal <u>**policy**</u>.
- 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.

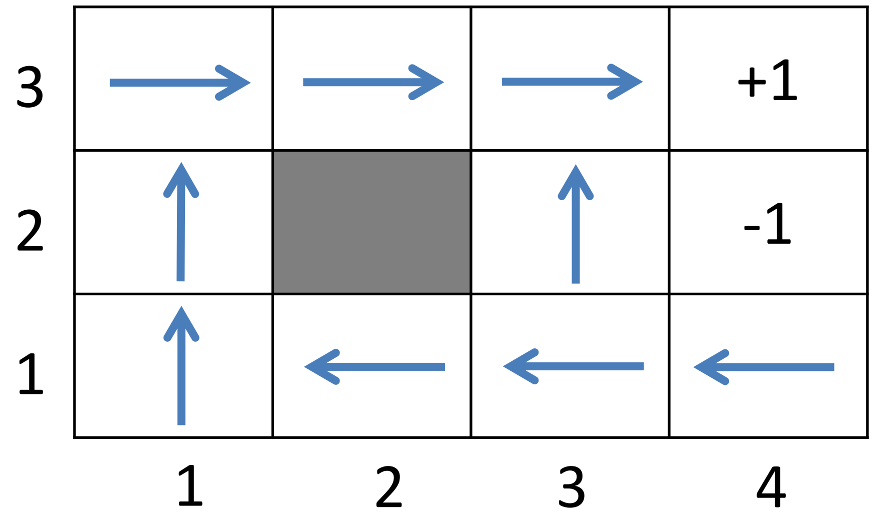
Policy Examples

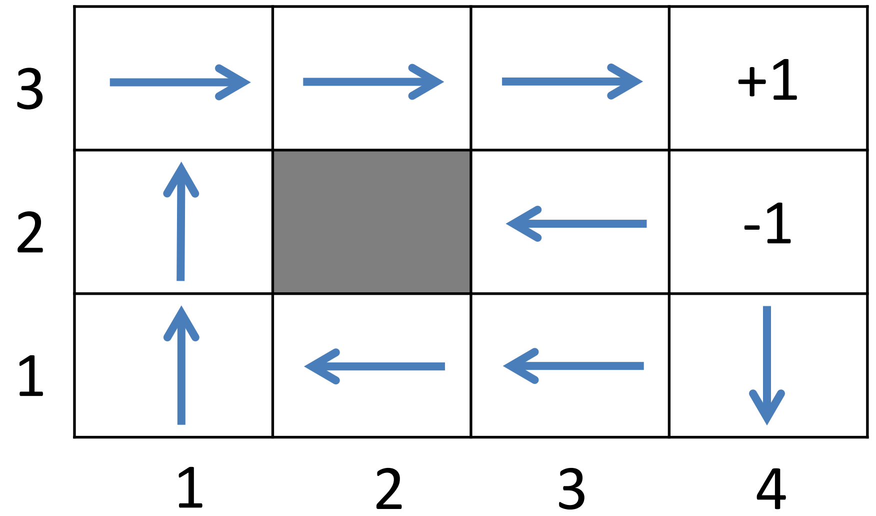


- A policy $\pi(s)$ is a function mapping states to actions.
- An optimal policy π^* is a policy that maximizes the **expected utility**.
- The figure shows an example policy, that happens to be optimal when:
  - $R(s) = -0.04$ for non-terminal states s.
  - $\gamma = 1$.

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.





Policy Examples

- Top figure: the optimal policy for:
  - $R(s) = -0.04$ for non-terminal s.
  - $\gamma = 1$.
- Bottom figure: the optimal policy for:
  - $R(s) = -0.1$ for non-terminal s.
  - $\gamma = 1$.
- Changing $R(s)$ from -0.04 to -0.1 makes longer sequences more costly.
  - It is worth taking risks to reach the +1 state as fast as possible.