FPT Education
FPT UNIVERSITY

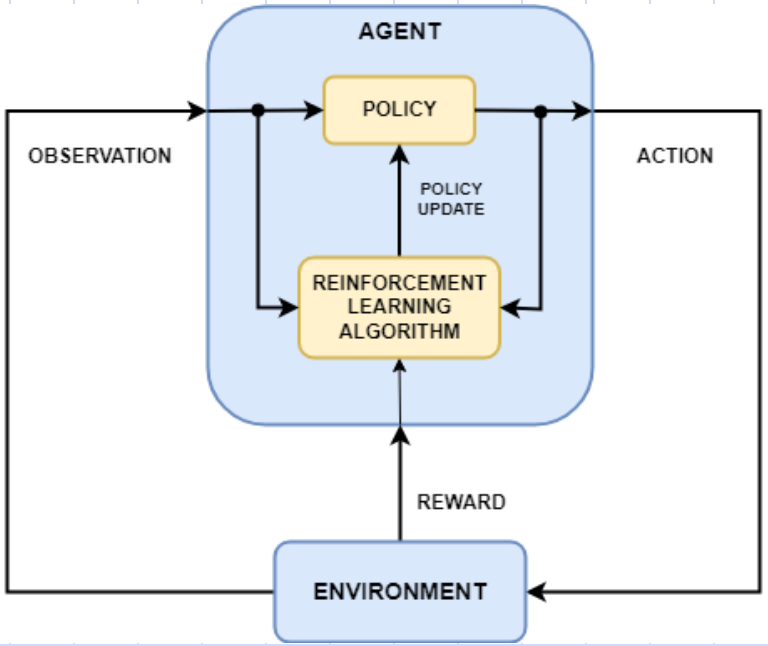
Policies and Value Functions

Objectives

- Recognize that a policy is a distribution over actions for each state,
- Describe the roles of the state value and action value functions in reinforcement learning
- Examples of policies and value functions for a given MDP

Policies

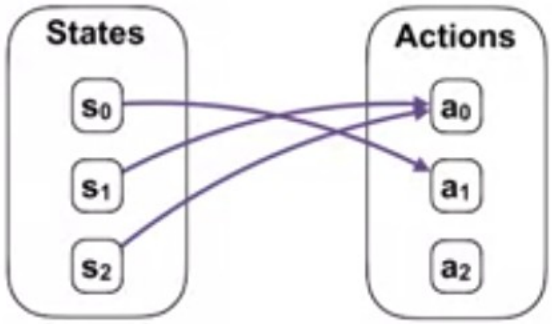
- A policy in RL defines the mapping from states of the environment to actions that an agent should take.
- It essentially det e agent within its enviror



Policy types

- Deterministic Policy:
  - In this type of policy, for each state, the agent selects a single action with certainty.
  - Mathematically, it can be represented as $\pi(s) = a$, where π is the policy, s is the state, and a is the action.

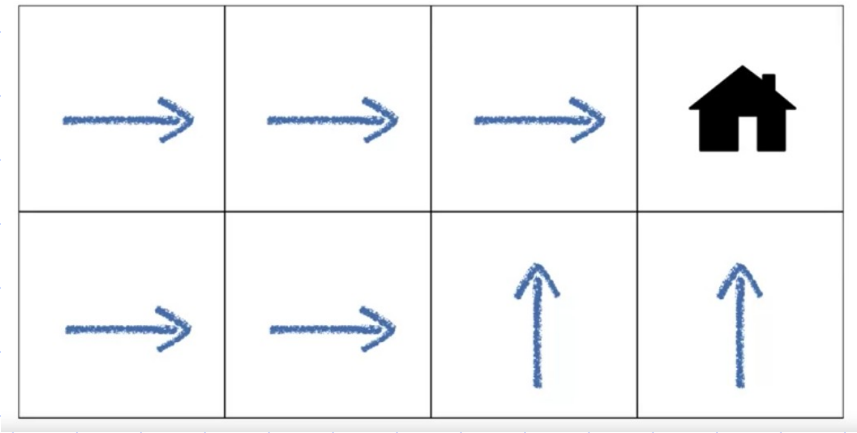
$$\pi(s) = a$$



State	Action
s_0	a_1
s_1	a_0
s_2	a_0

Deterministic Policy Example

- An agent moves towards its house on a grid.
- The states correspond to the locations on the grid.
- The actions move the agent up, down, left, and right.
- The arrows describe one possible policy, which moves the agent towards its house.
- Each arrow tells the agent which direction to move in each state.

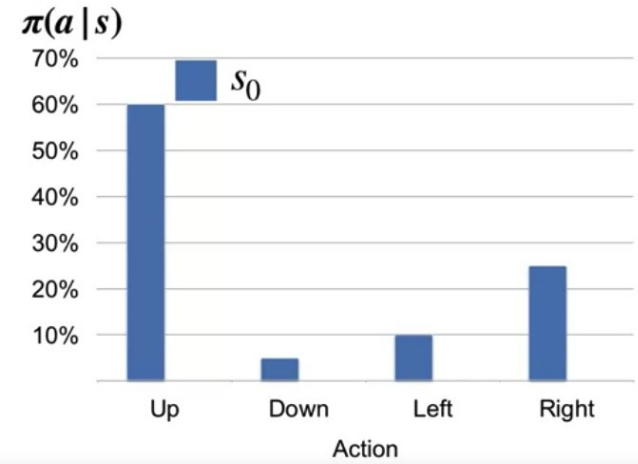


Policy types

- Stochastic Policy:
  - The agent selects actions based on a probability distribution over possible actions for each state.
  - This means that the agent's action selection is probabilistic rather than deterministic.
  - It is represented as $\pi(a|s) = P(A=a|S=s)$, indicating the probability of selecting action a given state

Policy type

- Stochastic Policy:
  - The distribution over actions for state S0 according to Pi
  - Pi specifies a separate distribution over actions for each state.
  - The sum over all action probabilities must be one for each state
  - Each



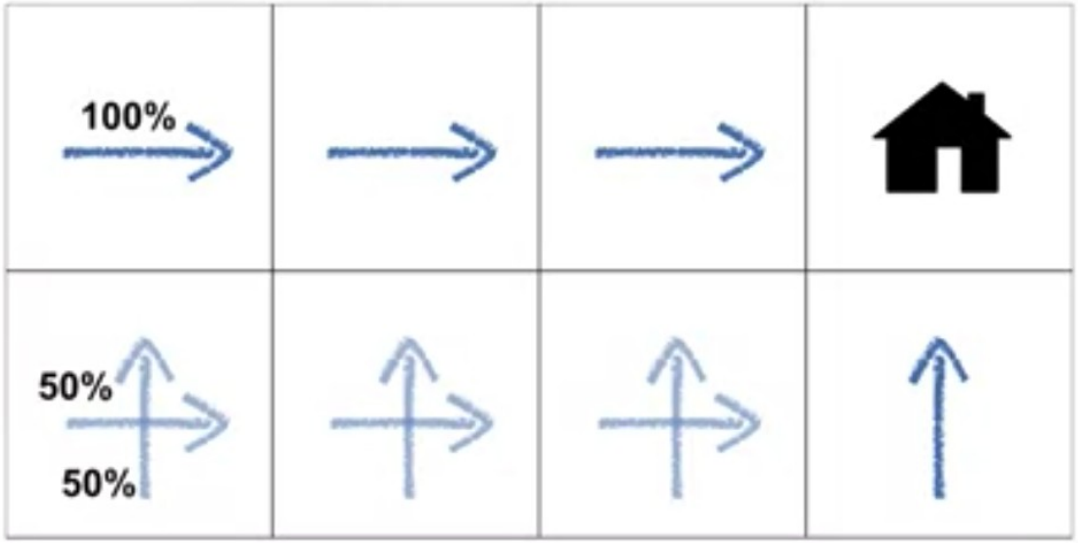
$$\pi(a \mid s)$$

$$\sum_{a \in \mathcal{A}(s)} \pi(a \mid s) = 1$$

$$\pi(a \mid s) \geq 0$$

Stochastic Policy Example

- A stochastic policy might choose up or right with equal probability in the bottom row.
- The stochastic policy will take the same number of steps to reach the house as the deterministic policy.



Value Function

- The value function is a critical concept used to estimate the long-term rewards or expected return that an agent can achieve from a given state or state-action pair.
- It helps the agent make decisions by providing a measure of how desirable it is to be in a particular state or to take a specific action from that state

Value Function Types

- State Value Function (V(s)):
  - The state value function predicts the expected return when starting from a particular state and following a specific policy thereafter.
  - Formally, it is defined as the expected sum of future rewards, discounted by a factor γ (gamma), when starting in state s and following policy π:

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t} R_{t+1} \mid s_0 = s\right]$$

  - It represents how good it is for the agent to be in a given state under a particular policy.

Value Function Types

- Action Value Function (Q(s, a)):
  - The action value function predicts the expected return when starting from a particular state, taking a specific action, and then following a particular policy thereafter.
  - Formally, it is defined as the expected sum of future rewards, discounted by a factor γ, when starting in state s, taking action a, and following policy π thereafter:

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t} R_{t+1} \mid s_0 = s, a_0 = a\right]$$

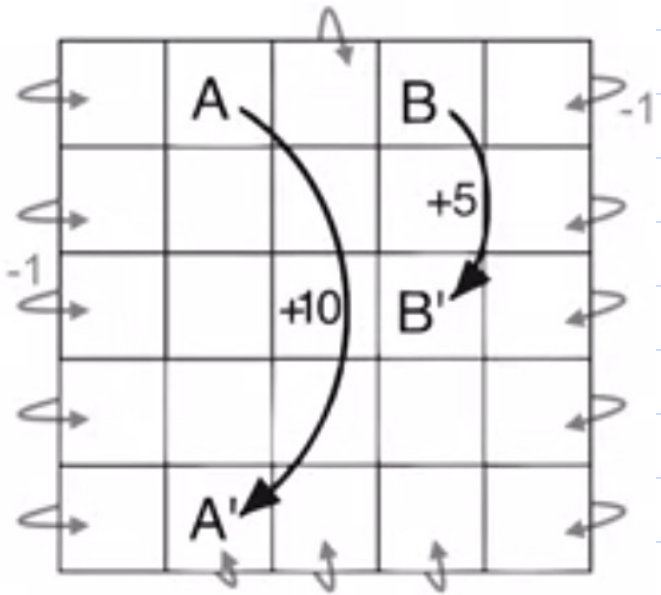
  - It represents how good it is for the agent to take a particular action in a given state under a particular policy.

Value Function

- Value function enable us to judge the quality of different policies.
- It guides the agent's decision-making process. By estimating the value of different states or state-action pairs, the agent can select actions that maximize its l

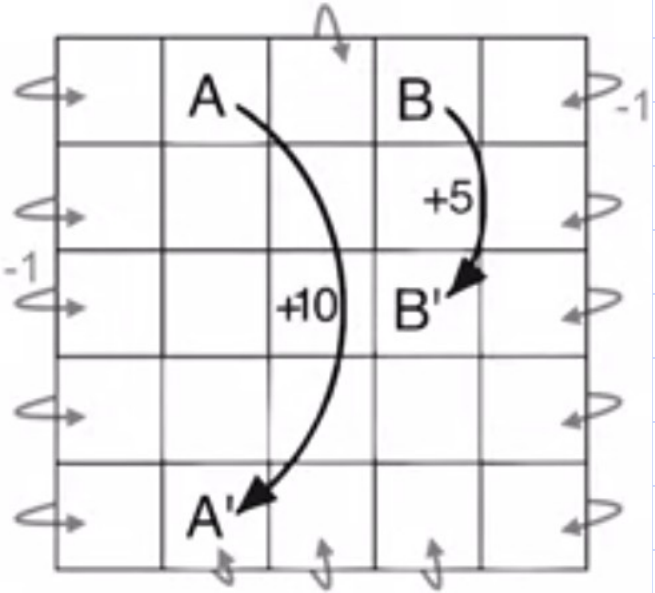
Value Function Example

- A simple continuing MDP.
  - The states are defined by the locations on the grid, the actions move the agent up, down, left, or right.
  - The agent cannot move off the grid and bumping generates a reward of minus on
  - Most other actions yield no rew



Value Function Example

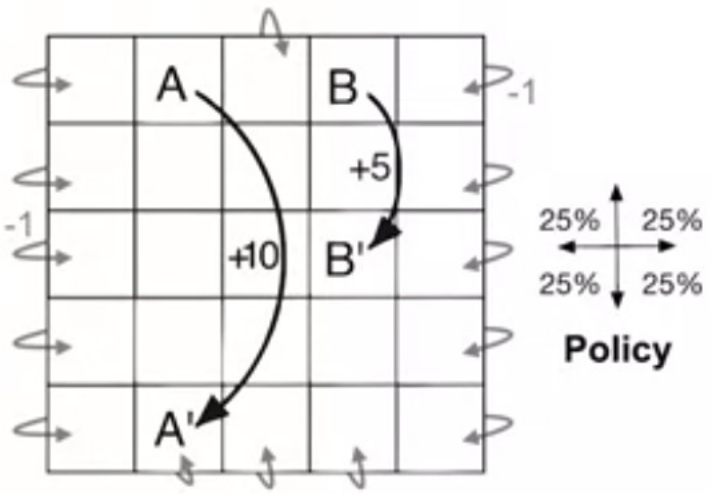
- A simple continuing MDP.
  - There are two special states however, these special states are labeled A and B.
  - Every action in state A yields plus 10 reward and plus five reward in state B.
  - Every action in state A and B transitions the agents to states A prime and B prime respectively.



Value Function Example

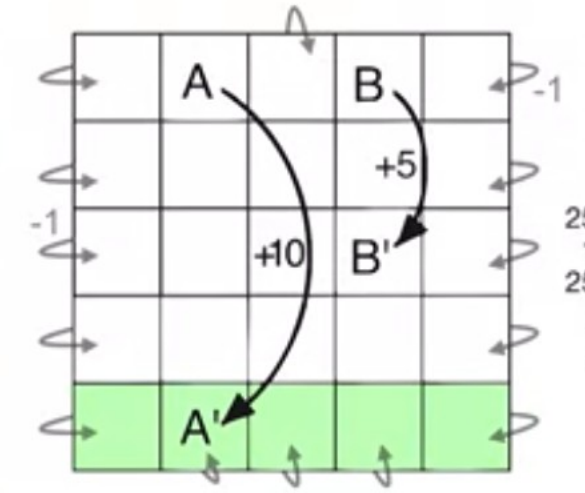
- A simple continuing MDP.
  - We must specify the policy before we can figure out what the value function is.
  - This is a continuing task, we need to specify Gamma=

$\gamma = 0.9$



Value Function Example

- A simple continuing MDP.
  - The value of each state is shown in the table.
  - The negative values near the bottom, these values are low because the agent is likely to bump into the wall before reaching the distance states A and B.

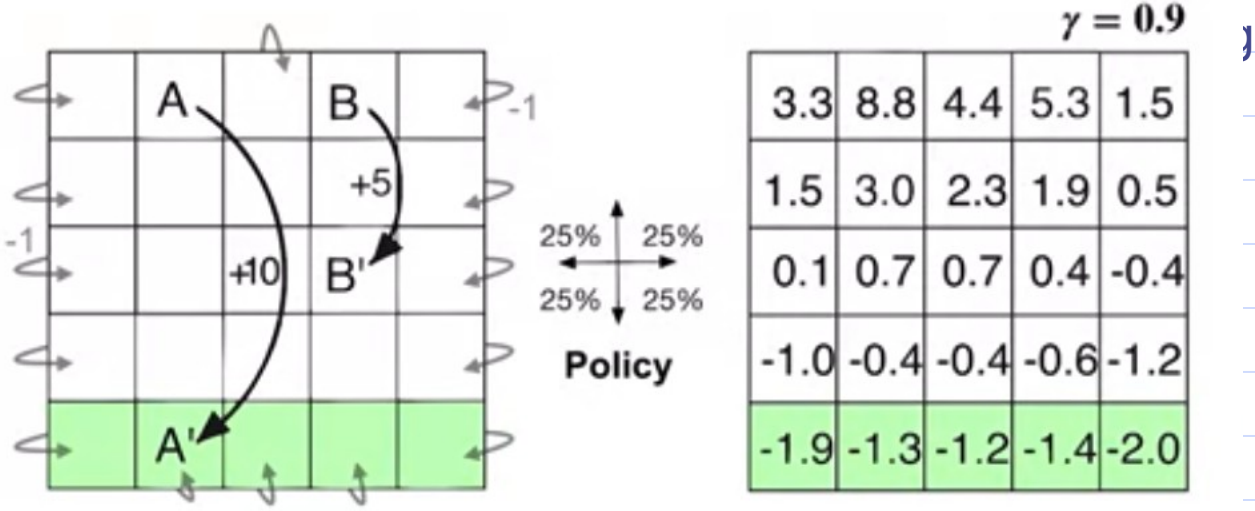


$\gamma = 0.9$

3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

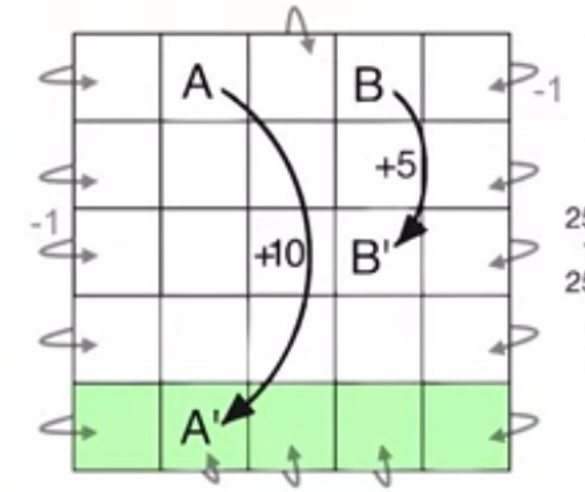
Value Function Example

- A simple continuing MDP.
  - A and B are both the only sources of positive reward in this MDP.
  - State A has the highest value: every transition from A moves the agent close to the lower wall and near the



Value Function Example

- A simple continuing MDP.
  - The value of state B is slightly greater than five.
  - The transition from B moves the agent to the middle.
  - In the middle, the agent is unlikely to bump and is close to the high-valued states A and B.



$\gamma = 0.9$

3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

Value Function Example

- The grid world as a 3x3 grid where the agent can move up, down, left, or right, and each cell has a reward associated with it.

Value Function Example

```python
# 1.7 Example for Value Function- HoaDNt@fe.edu.vn
import numpy as np

class GridWorld:
    def __init__(self):
        self.grid_size = (3, 3)
        self.num_actions = 4  # Up, Down, Left, Right
        self.rewards = np.array([
            [0, 0, 0],
            [0, 0, 0],
            [0, 1, 0]  # Reward of +1 in the bottom-right cell
        ])

    def get_reward(self, state):
        return self.rewards[state[0], state[1]]

class ValueFunction:
    def __init__(self, grid_size):
        self.values = np.zeros(grid_size)

    def update_value(self, state, new_value):
        self.values[state[0], state[1]] = new_value

    def get_value(self, state):
        return self.values[state[0], state[1]]

```

Value Function Example

```
26
27 # Create a grid world environment
28 grid_world = GridWorld()
29
30 # Create a value function for the grid world
31 value_function = ValueFunction(grid_world.grid_size)
32
33 # Initialize value function with rewards
34 for i in range(grid_world.grid_size[0]):
35     for j in range(grid_world.grid_size[1]):
36         state = (i, j)
37         value_function.update_value(state, grid_world.get_reward(state))
38
39 # Print the initial value function
40 print("Initial Value Function:")
41 print(value_function.values)
42
```

```
Initial Value Function:
[[0. 0. 0.]
 [0. 0. 0.]
 [0. 1. 0.]]
```

Summary

- Recognize that a policy is a distribution over actions for each state,
- Describe the roles of the state value and action value functions in reinforcement learning
- Examples of policies and value functions for a given MDP
-

FPT Education
FPT UNIVERSITY

Q & A