# Optimal Policy

# Policy improvement

$$\text{greedy}(\pi) \geq \pi$$

# Iterative greedy policy improvement

$\pi_1$

# Iterative greedy policy improvement

$$\pi_1 \leq \text{greedy}(\pi_1)$$

# Iterative greedy policy improvement

$$\pi_1 \leq \text{greedy}(\pi_1) = \pi_2$$

# Iterative greedy policy improvement

$$\pi_1 \leq \text{greedy}(\pi_1) = \pi_2 \leq \text{greedy}(\pi_2)$$

# Iterative greedy policy improvement

$$\pi_1 \leq \text{greedy}(\pi_1) = \pi_2 \leq \text{greedy}(\pi_2) = \pi_3$$

# Iterative greedy policy improvement

$$\pi_1 \leq \text{greedy}(\pi_1) = \pi_2 \leq \text{greedy}(\pi_2) = \pi_3 \leq \text{greedy}(\pi_3)$$

# Iterative greedy policy improvement

$$\pi_1 \leq \text{greedy}(\pi_1) = \pi_2 \leq \text{greedy}(\pi_2) = \pi_3 \leq \text{greedy}(\pi_3) = \pi_4$$

# Iterative greedy policy improvement

$$\pi_1 \leq \text{greedy}(\pi_1) = \pi_2 \leq \text{greedy}(\pi_2) = \pi_3 \leq \text{greedy}(\pi_3) = \pi_4 \leq \cdots$$

# Iterative greedy policy improvement

$$\pi_1 \leq \text{greedy}(\pi_1) = \pi_2 \leq \text{greedy}(\pi_2) = \pi_3 \leq \text{greedy}(\pi_3) = \pi_4 \leq \cdots$$

Does this improvement process converge?

# Iterative greedy policy improvement

$$\pi_1 \leq \text{greedy}(\pi_1) = \pi_2 \leq \text{greedy}(\pi_2) = \pi_3 \leq \text{greedy}(\pi_3) = \pi_4 \leq \cdots$$

Does this improvement process converge?

Yes!

# Optimal policy

# Optimal policy

## Definition

For finite MDPs (has terminal states) with bounded rewards, there exists an optimal policy $\pi_*$, such that

$$\boxed{\pi_* \geq \pi, \quad \forall \pi} \tag{1}$$

# Optimal policy

## Definition

For finite MDPs (has terminal states) with bounded rewards, there exists an optimal policy $\pi_*$, such that

$$\boxed{\pi_* \geq \pi, \quad \forall \pi} \tag{1}$$

- Iterative greedy policy improvement converges to the optimal policy.

$$\pi_1 \leq \text{greedy}(\pi_1) = \pi_2 \leq \text{greedy}(\pi_2) = \pi_3 \leq \cdots$$

# Optimal policy

## Definition

For finite MDPs (has terminal states) with bounded rewards, there exists an optimal policy $\pi_*$, such that

$$\boxed{\pi_* \geq \pi, \quad \forall \pi} \tag{1}$$

- Iterative greedy policy improvement converges to the optimal policy.

$$\pi_1 \leq \text{greedy}(\pi_1) = \pi_2 \leq \text{greedy}(\pi_2) = \pi_3 \leq \cdots \pi_*$$

# Optimal policy

## Definition

For finite MDPs (has terminal states) with bounded rewards, there exists an optimal policy $\pi_*$, such that

$$\boxed{\pi_* \geq \pi, \quad \forall \pi} \tag{1}$$

- Iterative greedy policy improvement converges to the optimal policy.

$$\pi_1 \leq \mathrm{greedy}(\pi_1) = \pi_2 \leq \mathrm{greedy}(\pi_2) = \pi_3 \leq \cdots \pi_*$$

- Greedy policy improvement wrt $\pi_*$ leads to $\pi_*$ itself

$$\pi_* \leq \mathrm{greedy}(\pi_*) \tag{2}$$

# Optimal policy

## Definition

For finite MDPs (has terminal states) with bounded rewards, there exists an optimal policy $\pi_*$, such that

$$\boxed{\pi_* \geq \pi, \quad \forall \pi} \tag{1}$$

- Iterative greedy policy improvement converges to the optimal policy.

$$\pi_1 \leq \text{greedy}(\pi_1) = \pi_2 \leq \text{greedy}(\pi_2) = \pi_3 \leq \cdots \pi_*$$

- Greedy policy improvement wrt $\pi_*$ leads to $\pi_*$ itself

$$\pi_* \leq \text{greedy}(\pi_*) \tag{2}$$

$$\implies \boxed{\text{greedy}(\pi_*) = \pi_*}$$

# Optimal policy

### Definition

For finite MDPs (has terminal states) with bounded rewards, there exists an optimal policy $\pi_*$, such that

$$\boxed{\pi_* \geq \pi, \quad \forall\pi} \tag{1}$$

- Optimal policy **maximizes the value of all states** in the MDP

# Optimal policy

### Definition

For finite MDPs (has terminal states) with bounded rewards, there exists an optimal policy $\pi_*$, such that

$$\boxed{\pi_* \geq \pi, \quad \forall \pi} \tag{1}$$

- Optimal policy **maximizes the value of all states** in the MDP

$$\pi^{'} \geq \pi \quad \text{if} \quad V_{\pi'}(s) \geq V_\pi(s), \quad \forall s \tag{2}$$

# Optimal policy

### Definition

For finite MDPs (has terminal states) with bounded rewards, there exists an optimal policy $\pi_*$, such that

$$\boxed{\pi_* \geq \pi, \quad \forall \pi} \tag{1}$$

- Optimal policy **maximizes the value of all states** in the MDP

$$\pi^{'} \geq \pi \quad \text{if} \quad V_{\pi^{'}}(s) \geq V_{\pi}(s), \quad \forall s \tag{2}$$

$$\implies \boxed{V_{\pi_*}(s) \geq V_{\pi}(s), \quad \forall s, \pi}$$

# We found a way to attain our goal in RL!

# We found a way to attain our goal in RL!

### Goal

Irrespective of the dynamics of the MDP, find the policy that maximize the discounted reward sum (value) for all states in the MDP.

# We found a way to attain our goal in RL!

### Goal

Irrespective of the dynamics of the MDP, find the *optimal policy* $\pi_*$.

# We found a way to attain our goal in RL!

## Goal

Irrespective of the dynamics of the MDP, find the *optimal policy* $\pi_*$.

## Method: iterative greedy policy improvement

$$\pi_1 \leq \text{greedy}(\pi_1) = \pi_2 \leq \text{greedy}(\pi_2) = \pi_3 \leq \text{greedy}(\pi_3) = \pi_4 \leq \cdots \pi_*$$

# We found a way to attain our goal in RL!