Policy improvement

Last lesson

We compared two policies

Last lesson

We compared two policies

- Random policy

Last lesson

We compared two policies

- Random policy
- Pole direction policy

Last lesson

We compared two policies

- Random policy
- Pole direction policy

$$\begin{aligned} \pi_{\text{pole-direction}} &\not\geq \pi_{\text{random}} \qquad (1) \\ \pi_{\text{random}} &\not\geq \pi_{\text{pole-direction}} \qquad (2) \end{aligned}$$

Last lesson

We compared two policies

- Random policy
- Pole direction policy

$$\pi_{\text{pole-direction}} \ngeq \pi_{\text{random}} \tag{1}$$

$$\pi_{\text{random}} \ngeq \pi_{\text{pole-direction}} \tag{2}$$

Policy improvement

Given a policy π, there is a procedural way to generate a better policy.

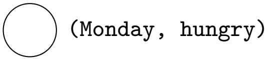
Two new italian restaurants in the neighborhood

Figure: Blue restaurant

Figure: Green restaurant

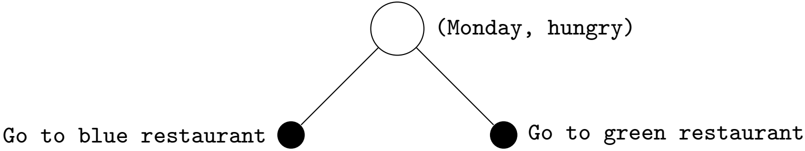
- You are the agent.
- Goal is to maximize pleasure from eating.

A simple, one step MDP

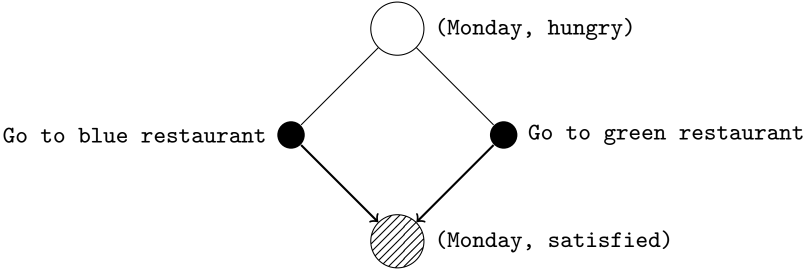


(Monday, hungry)

A simple, one step MDP



A simple, one step MDP



Policy improvement

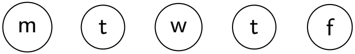


Figure: Week 1

Policy improvement

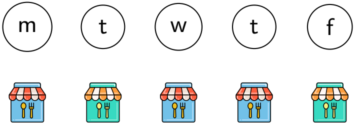


Figure: Week 1

- Follow random policy

Policy improvement

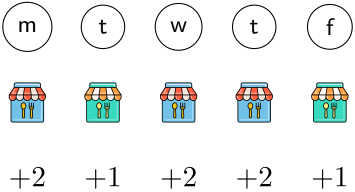


Figure: Week 1

- Follow random policy

Policy improvement

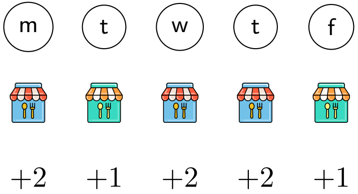


Figure: Week 1

- Follow random policy
- $Q(s, \text{blue restaurant}) = 2$

Policy improvement

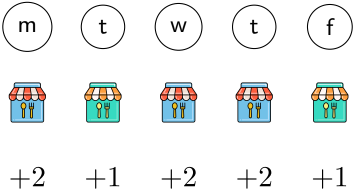


Figure: Week 1

- Follow random policy
- $Q(s, \text{blue restaurant}) = 2$
- $Q(s, \text{green restaurant}) = 1$

Policy improvement

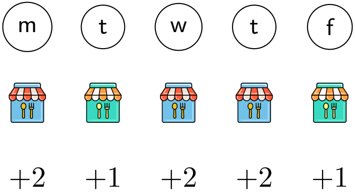


Figure: Week 1

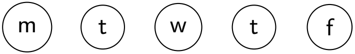


Figure: Week 2

- Follow random policy
- $Q(s, \text{blue restaurant}) = 2$
- $Q(s, \text{green restaurant}) = 1$

Policy improvement

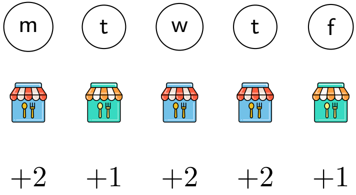


Figure: Week 1

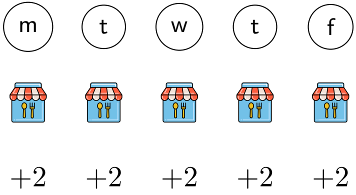


Figure: Week 2

- Follow random policy
- $Q(s, \text{blue restaurant}) = 2$
- $Q(s, \text{green restaurant}) = 1$

- Take action with max Q (**greedy** policy/**exploitation**)