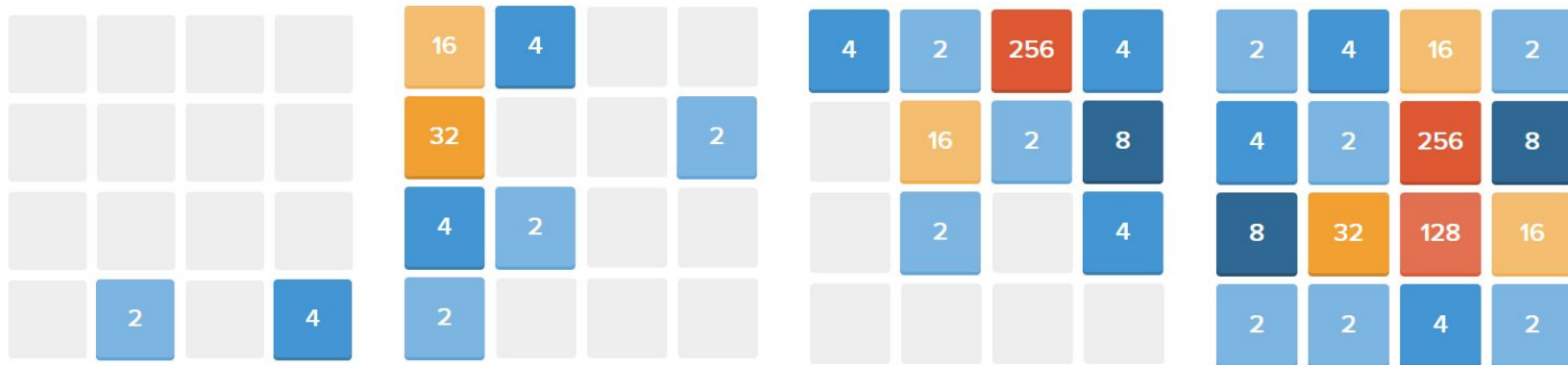


# Марковский процесс и уравнения Беллмана

# Markov Decision Process

# Markov Decision Process



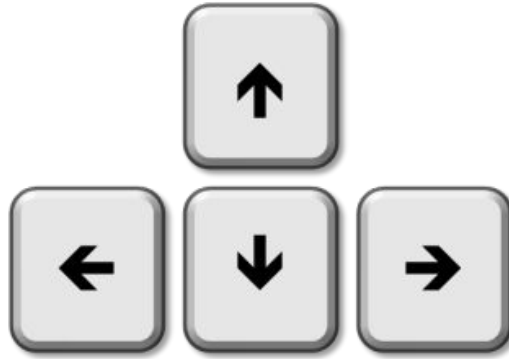
States

Actions

Probabilities

Rewards

# Markov Decision Process



States

**A**ctions

Probabilities

Rewards

# Markov Decision Process

$$P\left( \begin{array}{|c|c|c|c|} \hline 4 & 2 & 256 & 4 \\ \hline \text{ } & 16 & 2 & 8 \\ \hline \text{ } & 2 & \text{ } & 4 \\ \hline \text{ } & \text{ } & \text{ } & \text{ } \\ \hline \end{array} \mid \begin{array}{|c|c|c|c|} \hline 16 & 4 & \text{ } & \text{ } \\ \hline 32 & \text{ } & \text{ } & 2 \\ \hline 4 & 2 & \text{ } & \text{ } \\ \hline 2 & \text{ } & \text{ } & \text{ } \\ \hline \end{array} , \rightarrow \right)$$

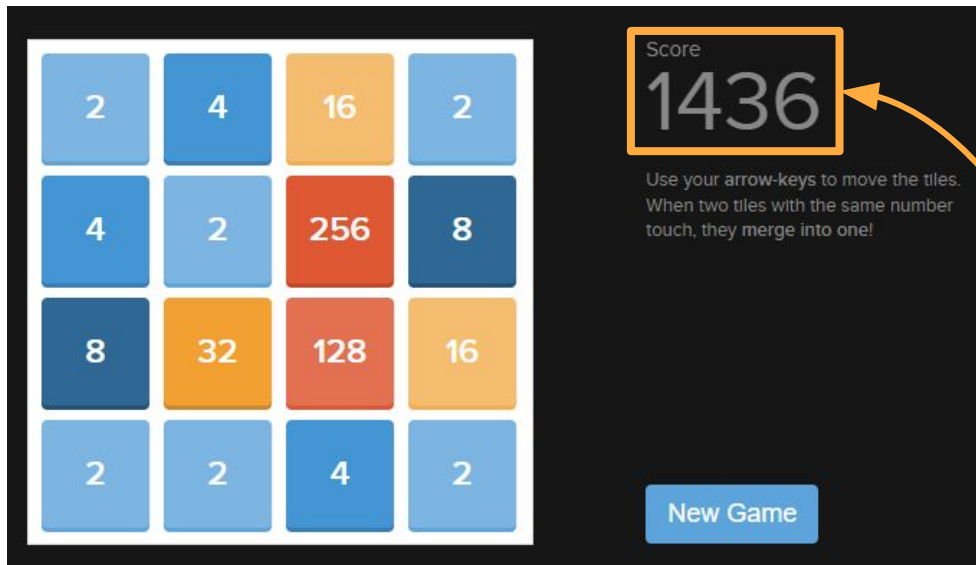
States

Actions

Probabilities

Rewards

# Markov Decision Process



States

Actions

Probabilities

Rewards

## Более формально

MDP is a tuple  $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$ , where:

1.  $\mathcal{S}$  is a set of states of the environment
2.  $\mathcal{A}$  is a set of actions
3.  $\mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$  is a state-transiting function
4.  $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  is a reward function

# СВОЙСТВО

Свойство Марковского процесса:

$$p(r_t, s_{t+1} | s_0, a_0, r_0, \dots, s_t, a_t) = p(r_t, s_{t+1} | s_t, a_t)$$

следующее состояние и награда зависят от предыдущего состояния и действия



# Return

Накопленную награду  $G$  назовем return:

$$G_t = R_t + R_{t+1} + R_{t+2} + \cdots + R_T$$

# Return

Накопленную награду  $G$  назовем return:

$$G_t = R_t + R_{t+1} + R_{t+2} + \dots + R_T$$

Diagram illustrating the components of the return calculation:

- $G_t$  is the total return (highlighted with a blue box).
- $R_t$  is the initial reward (highlighted with a purple box, labeled "начальная награда").
- $R_T$  is the reward at the end of the episode (highlighted with an orange box, labeled "конец эпизода").

# Return

В return добавляют коэффициент дисконтирования  $\gamma$ :

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k}$$

$$0 \leq \gamma < 1$$

# State-value function $v(s)$

Ожидаемый return при фиксированном policy  $\pi$  и состоянии  $s$ .

$$v_{\pi}(s) = \mathbb{E}_{\pi} [G_t \mid S_t = s]$$

# State-value function $v(s)$

Ожидаемый return при фиксированном policy  $\pi$  и состоянии  $s$ .

$$\begin{aligned} v_{\pi}(s) &= \mathbb{E}_{\pi} [G_t \mid S_t = s] \\ &= \mathbb{E}_{\pi} [R_t + \gamma G_{t+1} \mid S_t = s] \end{aligned}$$

# State-value function $v(s)$

Ожидаемый return при фиксированном policy  $\pi$  и состоянии  $s$ .

$$\begin{aligned} v_{\pi}(s) &= \mathbb{E}_{\pi} [G_t \mid S_t = s] \\ &= \mathbb{E}_{\pi} [R_t + \gamma G_{t+1} \mid S_t = s] \\ &= \sum_a \pi(a \mid s) \sum_{r, s'} p(r, s' \mid s, a) \left[ r + \gamma \mathbb{E}_{\pi} [G_{t+1} \mid S_{t+1} = s'] \right] \\ &= \sum_a \pi(a \mid s) \sum_{r, s'} p(r, s' \mid s, a) [r + \gamma v_{\pi}(s')] \end{aligned}$$

# Action-value function $q(s, a)$

Ожидаемый return при фиксированном policy  $\pi$ , состоянии  $s$  и действии  $a$ .

$$q_{\pi}(s, a) = \mathbb{E}_{\pi} [G_t \mid S_t = s, A_t = a]$$

# Action-value function $q(s, a)$

Ожидаемый return при фиксированном policy  $\pi$ , состоянии  $s$  и действии  $a$ .

$$\begin{aligned} q_{\pi}(s, a) &= \mathbb{E}_{\pi} [G_t \mid S_t = s, A_t = a] \\ &= \mathbb{E}_{\pi} [R_t + \gamma G_{t+1} \mid S_t = s, A_t = a] \end{aligned}$$



# Action-value function $q(s, a)$

Ожидаемый return при фиксированном policy  $\pi$ , состоянии  $s$  и действии  $a$ .

$$\begin{aligned} q_{\pi}(s, a) &= \mathbb{E}_{\pi} [G_t \mid S_t = s, A_t = a] \\ &= \mathbb{E}_{\pi} [R_t + \gamma G_{t+1} \mid S_t = s, A_t = a] \\ &= \sum_{r, s'} p(r, s' \mid s, a) \left[ r + \gamma \mathbb{E}_{\pi} [G_{t+1} \mid S_{t+1} = s'] \right] \\ &= \sum_{r, s'} p(r, s' \mid s, a) [r + \gamma v_{\pi}(s')] \end{aligned}$$

$v(s)$  vs  $q(s, a)$

$$q_{\pi}(s, a) = \sum_{r, s'} p(r, s' \mid s, a) [r + \gamma v_{\pi}(s')]$$

# $v(s)$ vs $q(s, a)$

$$q_{\pi}(s, a) = \sum_{r, s'} p(r, s' \mid s, a) [r + \gamma v_{\pi}(s')]$$

$$\begin{aligned} v_{\pi}(s) &= \sum_a \pi(a \mid s) \sum_{r, s'} p(r, s' \mid s, a) [r + \gamma v_{\pi}(s')] \\ &= \sum_a \pi(a \mid s) q_{\pi}(s, a) \end{aligned}$$

# $v(s)$ vs $q(s, a)$

$$q_{\pi}(s, a) = \sum_{r, s'} p(r, s' | s, a) [r + \gamma v_{\pi}(s')]$$

$$v_{\pi}(s) = \sum_a \pi(a | s) \sum_{r, s'} p(r, s' | s, a) [r + \gamma v_{\pi}(s')]$$

$$= \sum_a \pi(a | s) q_{\pi}(s, a)$$

$$q_{\pi}(s, a) = \sum_{r, s'} p(r, s' | s, a) \left[ r + \gamma \sum_{a'} \pi(a' | s') q_{\pi}(s', a') \right]$$

Bellman's equations:  
evaluation and optimality

# Bellman's expectation equations

# Bellman's **expectation** equations

$$\begin{aligned}v_{\pi}(s) &= \sum_a \pi(a \mid s) \sum_{r, s'} p(r, s' \mid s, a) [r + \gamma v_{\pi}(s')] \\&= \mathbb{E}_{\pi} [R_t + \gamma v_{\pi}(S_{t+1}) \mid S_t = s]\end{aligned}$$

$$\begin{aligned}q_{\pi}(s, a) &= \sum_{r, s'} p(r, s' \mid s, a) [r + \gamma v_{\pi}(s')] \\&= \sum_{r, s'} p(r, s' \mid s, a) \left[ r + \gamma \sum_{a'} \pi(a' \mid s') q_{\pi}(s', a') \right]\end{aligned}$$

# Bellman's optimality equations



# Bellman's optimality equations

$$\begin{aligned} v_*(s) &= \max_a \sum_{r, s'} p(r, s' \mid s, a) [r + \gamma v_*(s')] \\ &= \max_a \mathbb{E} [R_t + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a] \end{aligned}$$

$$\begin{aligned} q_*(s, a) &= \mathbb{E} \left[ R_t + \gamma \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a \right] \\ &= \sum_{r, s'} p(r, s' \mid s, a) \left[ r + \gamma \max_{a'} q_*(s', a') \right] \end{aligned}$$

# Bellman's **expectation** equations

$$\begin{aligned}v_{\pi}(s) &= \sum_a \pi(a \mid s) \sum_{r, s'} p(r, s' \mid s, a) [r + \gamma v_{\pi}(s')] \\&= \mathbb{E}_{\pi} [R_t + \gamma v_{\pi}(S_{t+1}) \mid S_t = s]\end{aligned}$$

$$\begin{aligned}q_{\pi}(s, a) &= \sum_{r, s'} p(r, s' \mid s, a) [r + \gamma v_{\pi}(s')] \\&= \sum_{r, s'} p(r, s' \mid s, a) \left[ r + \gamma \sum_{a'} \pi(a' \mid s') q_{\pi}(s', a') \right]\end{aligned}$$

Как использовать?

# Как использовать?



# Как использовать?



- policy iteration
- value iteration