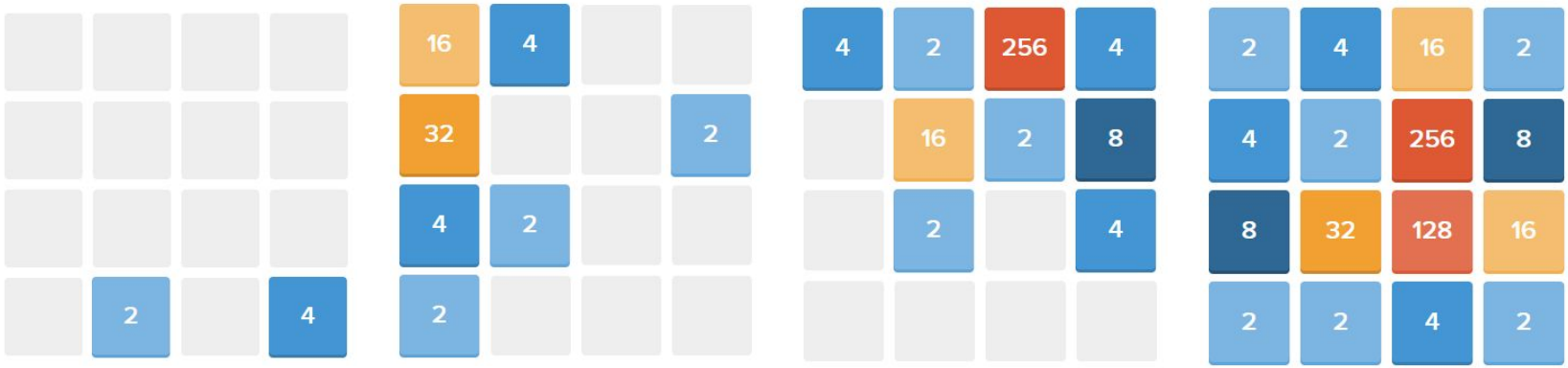
Марковский процесс
и уравнения Беллмана

Рахманов Сергей, БПМИ192

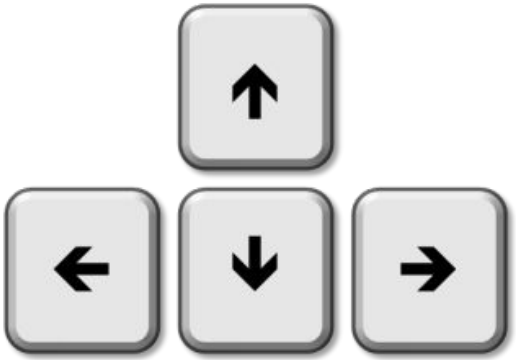
Markov Decision Process

Markov Decision Process



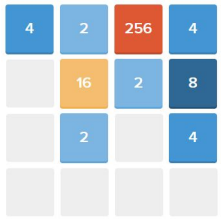
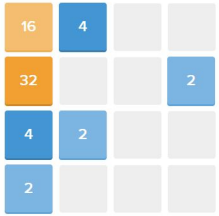
States Actions Probabilities Rewards

Markov Decision Process



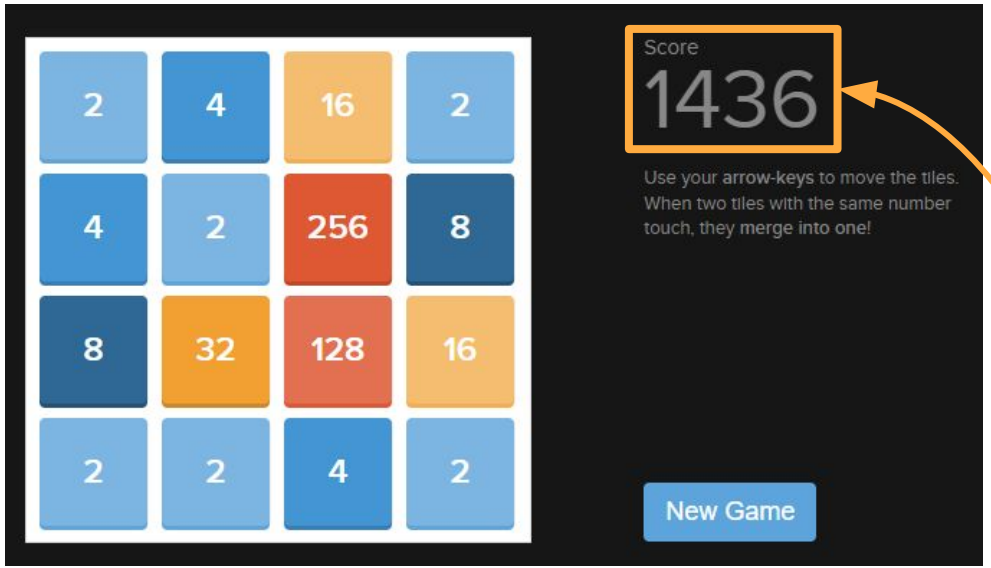
States **Actions** Probabilities Rewards

Markov Decision Process

P( |  , →)

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. $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a state-transiting function
4. $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function

Свойство

Свойство Марковского процесса:

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

следующее состояние и награда зависят от предыдущего состояния и действия

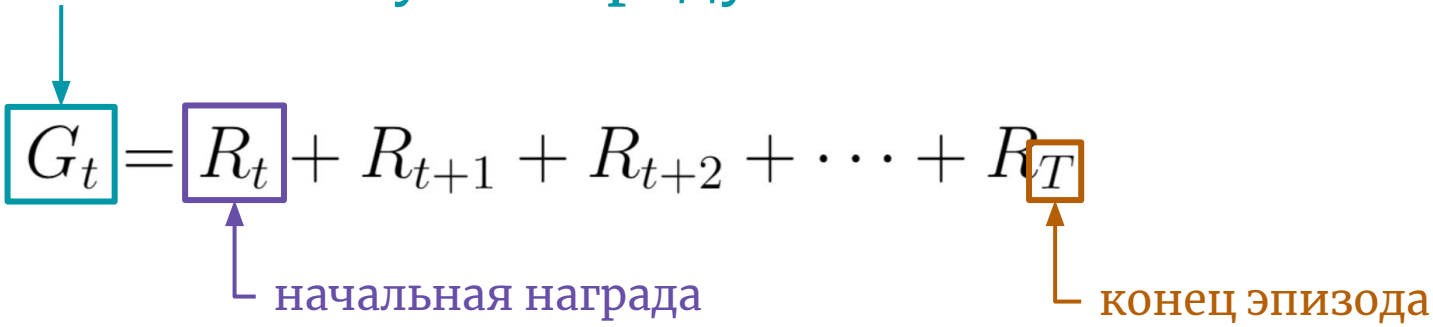
Return

Накопленную награду G назовем return:

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

Return

Накопленную награду G назовем return:



Return

В return добавляют коэффициент дисконтирования γ:

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

$$0 \leqslant \gamma < 1$$

State-value function *v(s)*

Ожидаемый return при фиксированном policy π и состоянии s.

$$v_\pi(s) = \mathbb{E}_\pi \left[G_t \,|\, S_t = s \right]$$

State-value function $v(s)$

Ожидаемый return при фиксированном policy π и состоянии s.

$$v_\pi(s) = \mathbb{E}_\pi \left[G_t \,|\, S_t = s \right]$$

$$= \mathbb{E}_\pi \left[R_t + \gamma G_{t+1} \,|\, S_t = s \right]$$

State-value function $v(s)$

Ожидаемый return при фиксированном policy π и состоянии s.

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

Action-value function *q(s, a)*

Ожидаемый return при фиксированном policy π, состоянии s и **действии a**.

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

Action-value function *q(s, a)*

Ожидаемый return при фиксированном policy π, состоянии s и **действии a**.

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

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

Ожидаемый return при фиксированном policy π, состоянии s и **действии a**.

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

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

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

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

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

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

$$= \sum_a \pi(a \mid s) q_\pi(s,a)$$

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

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

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

$$= \sum_a \pi(a \mid s) q_\pi(s,a)$$

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

Bellman’s equations: evaluation and optimality

Bellman’s expectation equations

Bellman's expectation equations

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

$$= \mathbb{E}_\pi \left[R_t + \gamma v_\pi(S_{t+1}) \mid S_t = s \right]$$

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

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

Bellman’s optimality equations

Bellman's optimality equations

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

$$= \max_a \; \mathbb{E}\left[R_t + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a\right]$$

$$q_*(s,a) = \mathbb{E}\left[R_t + \gamma \max_{a'} q_*(S_{t+1}, a') \;\middle|\; S_t = s, A_t = a\right]$$

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

Bellman's expectation equations

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

$$= \mathbb{E}_\pi \left[R_t + \gamma v_\pi(S_{t+1}) \mid S_t = s \right]$$

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

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

Как использовать?

Как использовать?

Как использовать?

- policy iteration
- value iteration

Источники

- **Markov Decision Processes: Discrete Stochastic Dynamic Programming**
  Martin L. Puterman, ISBN 978-0-471-72782-8
- **Algorithms of Reinforcement Learning**
  Csaba Szepesvari, ISBN 978-1608454921
  https://sites.ualberta.ca/~szepesva/rlbook.html
- **Курс ШАД по RL**
  https://github.com/yandexdataschool/Practical_RL