

RL

Dimtirov Vladimir

13 февраля 2023 г.

Содержание

Markov decision process	3
Expected objective	5
State value function	5
Action value function	6
Bellman expectation equation	7
Bellman optimality equation	8
Generalized Policy Iteration	9
Policy evaluation	9
Policy improvement	10

Markov decision process

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

- \mathcal{S} - set of states of the world
- \mathcal{A} - set of actions
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ - state-transition function, giving us $p(s_{t+1} \mid s_t, a_t)$
- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ - reward function, given us $\mathbb{E}_R[R(s_t, a_t) \mid s_t, a_t]$

Reward hypothesis (R.Sutton)

That all of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward)

Cumulative rewards is called a return:

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

There are **2 problems** in the design of the system:

1. Infinite sum -> solve: discounting coefficient
2. Simple solve -> watch the pic 1

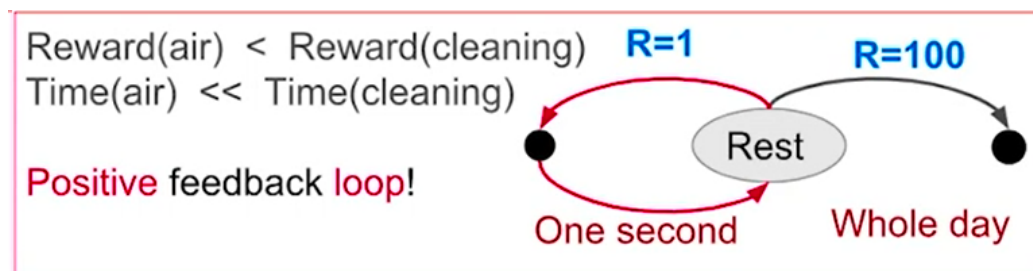


Рис. 1: Simple solution

Solving the infinite sum problem: get discounting coefficient ($0 \leq \gamma < 1$)

and cumulative rewards will take the form:

$$\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

This method has its pluses:

- Human likeness
- Mathematical convenience
- fast optimization

Notes: multiplying by γ changes the task and it's solution!

Take away 1: reward only for what, but never for how

Take away 2: do not subtract mean from rewards

Transformation politic for reward (ML advice):

- Rewards scaling - division by positive constant

- Reward shaping - we could add to all rewards in MDP values of potential-based shaping function

Expected objective

Optimal policy maximizes expected return:

$$\begin{aligned}
\mathbb{E}[G_0] &= \mathbb{E}[R_0 + \gamma R_1 + \dots + \gamma^T R_T] \\
&= \mathbb{E}_{E, \pi_\theta}[G_0] \\
&= \mathbb{E}_{\pi_\theta}[G_0] \\
&= \mathbb{E}[G_0 \mid \pi_\theta] \\
&= \mathbb{E}_{s_{0:T}}[G_0] \\
&= \mathbb{E}_{a_{0:T}}[\mathbb{E}_{a_0|s_0}[R_0 + \mathbb{E}_{s_1|s_0, a_0}[\mathbb{E}_{a_1|s_1}[\gamma R_1 + \dots]]]] \\
&= \sum_{t=0}^T \mathbb{E}_{(s_t, a_t) \sim p_\theta}[\gamma^t R_t]
\end{aligned}$$

State value function

” We want to know value function not only from 0 point in time ”

Is the expected return conditional on state:

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

Intuition: mean value of following policy π from state s

In Russian: 'внешнее ожидание по политики и внутреннее мат ожидание по среде'

Action value function

Is expected return conditional on state and action:

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

Intuition: value of following policy π after committing action a in state s

In Russian: 'Какую ценность имеет политика π , если я в состоянии s

сделаю действие a ,

Expression value function through state function:

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

Bellman expectation equation

$v(s)$ is calculated:

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

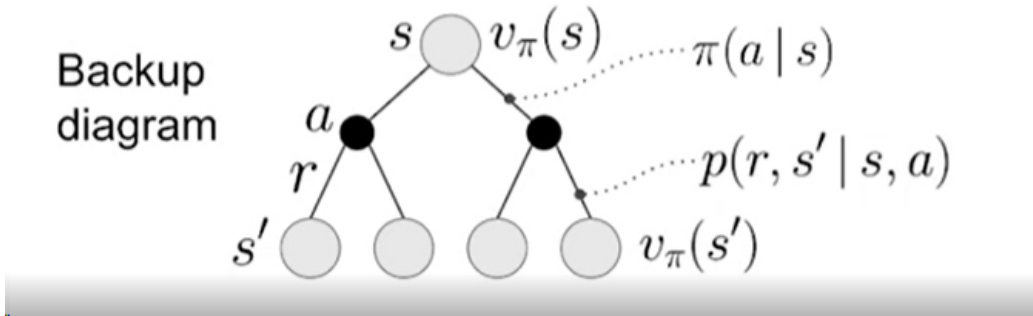


Рис. 2: Backup diagram for $v(s)$

$q(s)$ is calculated:

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

Backup
diagram
for $q(s, a)$

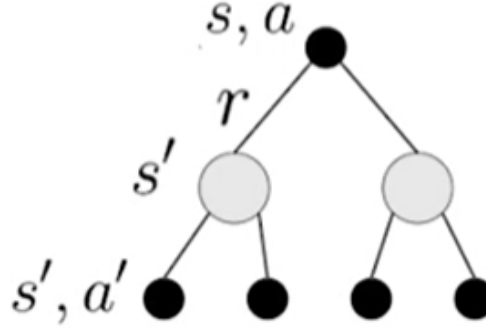


Рис. 3: Backup diagram for $q(s)$

Bellman optimality equation

We could compare policies on the basis of $v(s)$

$$\pi \geq \pi' \Leftrightarrow v_\pi(s) \geq v_{\pi'}(s) \quad \forall s$$

$$v_*(s) = \max_\pi v_\pi(s)$$

$$q_*(s, a) = \max_\pi q_\pi(s, a)$$

Notes: in any finite MDP there is always at least one deterministic

optimal policy

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

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

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

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

Generalized Policy Iteration

1. Policy Evaluation
2. Policy Improvement

Policy evaluation

Policy evaluation is also called **prediction problem**.

Predict value function for a particular policy

Bellman expectation equation:

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

is basically a system of linear equation

Algorithm:

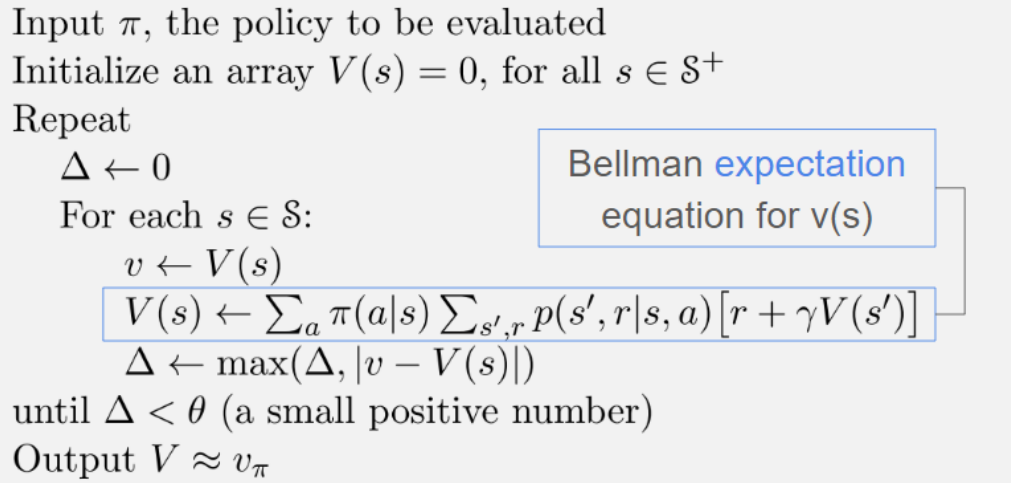


Рис. 4:

Policy improvement

Idea: once we know what $v(s)$ for a particular policy, we could improve it by acting greedily w.r.t $v(s)$!

$$\pi'(s) \leftarrow \arg \max_a \overbrace{\sum_{r,s'} p(r,s' | s, a) [r + \gamma v_\pi(s')]}^{q_\pi(s,a)}$$

This procedure is guaranteed to produce a better policy.