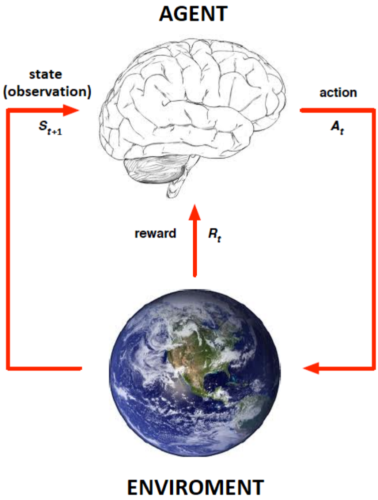
Лекция 3: Динамическое программирование

Антон Романович Плаксин

Reinforcement Learning



Цель агента - максимизировать $G = \sum_{t=0}^{T} \gamma^t R_t, \quad \gamma \in [0,1].$

Markov Decision Process

Markov Property

$$\mathbb{P}[S_{t+1}|S_t, A_t] = \mathbb{P}[S_{t+1}|S_1, A_1, S_2, A_2 \ldots, S_t, A_t]$$

$$\mathbb{P}[R_t|S_t, A_t] = \mathbb{P}[R_t|S_1, A_1, S_2, A_2 \ldots, S_t, A_t] = 1$$

Markov Decision Process

Markov Property

$$\mathbb{P}[S_{t+1}|S_t, A_t] = \mathbb{P}[S_{t+1}|S_1, A_1, S_2, A_2 \dots, S_t, A_t]$$

$$\mathbb{P}[R_t|S_t, A_t] = \mathbb{P}[R_t|S_1, A_1, S_2, A_2 \dots, S_t, A_t] = 1$$

Markov Decision Process $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$

- $\mathcal{S}$ — конечное ($|\mathcal{S}| = n$) пространство состояний
- $\mathcal{A}$ — конечное ($|\mathcal{A}| = m$) пространство действий
- $\mathcal{P}$ — известная функция (тензор) вероятностей переходов между состояниями

$$\mathcal{P}(s'|s, a) = \mathbb{P}[S_{t+1} = s'|S_t = s, A_t = a]$$

- $\mathcal{R}$ — известная функция (матрица) вознаграждений

$$\mathcal{R}(s, a) = R_t \quad \Leftrightarrow \quad \mathbb{P}[R_t|S_t = s, A_t = a] = 1$$

- $\gamma \in [0, 1]$ — коэффициент дисконтирования

Policy

$$\pi(a|s) \in [0,1], \quad a \in \mathcal{A}, \quad s \in \mathcal{S}$$

- Мы задаем π
- АГЕНТ находится в начальном состоянии $S_0 \in \mathcal{S}$
- совершает действие $A_0 \sim \pi(\cdot|S_0)$
- получает награду $R_0 = \mathcal{R}(S_0, A_0)$ и переходит в следующее состояние $S_1 \sim \mathcal{P}(\cdot|S_0, A_0)$
- совершает действие $A_1 \sim \pi(\cdot|S_1)$
- получает награду $R_1 = \mathcal{R}(S_1, A_1)$ и переходит в следующее состояние $S_2 \sim \mathcal{P}(\cdot|S_1, A_1)$
- $\ldots$
- $\tau = \{S_0, A_0, S_1, A_1, \ldots, S_T, A_T\}, \quad G(\tau) = \sum\limits_{t=0}^{T} \gamma^t \mathcal{R}(S_t, A_t)$

Наша задача

$$\mathbb{E}_\pi[G] \quad \longrightarrow \quad \max_\pi$$

Policy

$$\pi(a|s) \in [0,1], \quad a \in \mathcal{A}, \quad s \in \mathcal{S}$$

- Мы задаем π
- АГЕНТ находится в начальном состоянии $S_0 \in \mathcal{S}$
- совершает действие $A_0 \sim \pi(\cdot|S_0)$
- получает награду $R_0 = \mathcal{R}(S_0, A_0)$ и переходит в следующее состояние $S_1 \sim \mathcal{P}(\cdot|S_0, A_0)$
- совершает действие $A_1 \sim \pi(\cdot|S_1)$
- получает награду $R_1 = \mathcal{R}(S_1, A_1)$ и переходит в следующее состояние $S_2 \sim \mathcal{P}(\cdot|S_1, A_1)$
- $\ldots$
- $\tau = \{S_0, A_0, S_1, A_1, \ldots\}, \quad G(\tau) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_t, A_t)$

Наша задача

$$\mathbb{E}_\pi[G] \quad \longrightarrow \quad \max_\pi$$

State-Value Function

$$\mathbb{E}_\pi[G] = \sum_\tau G(\tau)\mathbb{P}(\tau|\pi),$$

где

$$G(\tau) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_t, A_t), \qquad \mathbb{P}(\tau|\pi) = \prod_{t=0}^{\infty} \pi(A_t|S_t)\mathcal{P}(S_{t+1}|S_t, A_t)$$

State-Value Function

$$\mathbb{E}_\pi[G] = \sum_\tau G(\tau)\mathbb{P}(\tau|\pi),$$

где

$$G(\tau) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_t, A_t), \qquad \mathbb{P}(\tau|\pi) = \prod_{t=0}^{\infty} \pi(A_t|S_t)\mathcal{P}(S_{t+1}|S_t, A_t)$$

State-Value Function

$$v_\pi(s) = \mathbb{E}_\pi[G \mid S_0 = s]$$

State-Value Function

$$\mathbb{E}_\pi[G] = \sum_\tau G(\tau)\mathbb{P}(\tau|\pi),$$

где

$$G(\tau) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_t, A_t), \qquad \mathbb{P}(\tau|\pi) = \prod_{t=0}^{\infty} \pi(A_t|S_t)\mathcal{P}(S_{t+1}|S_t, A_t)$$

State-Value Function

$$v_\pi(s) = \mathbb{E}_\pi[G \mid S_0 = s]$$

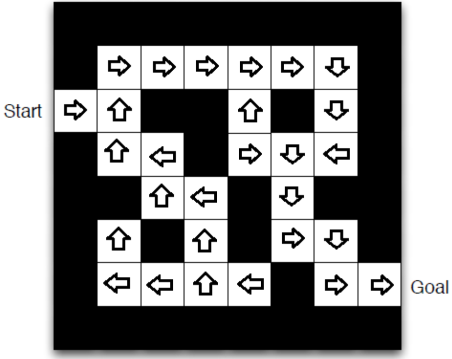
Замечание

Если Policy и Environment детерминированны (не стохастические), то

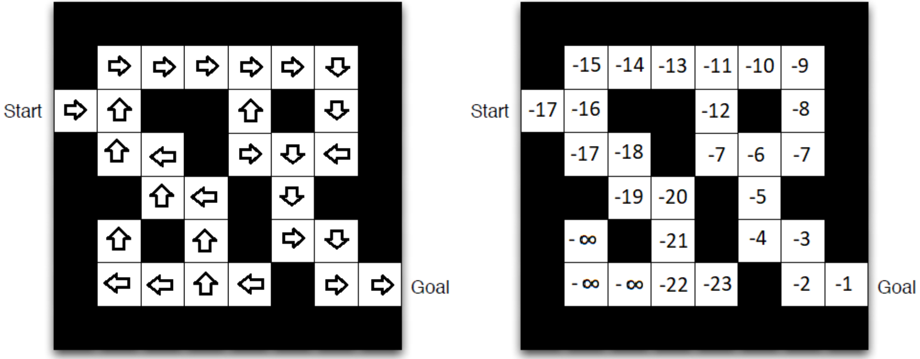
$$v_\pi(s) = G(\tau_\pi),$$

где $\tau_\pi : \mathbb{P}(\tau_\pi|\pi) = 1$.

Пример: Maze



Пример: Maze


Start
Goal
Start
-15 -14 -13 -11 -10 -9
-17 -16 -12 -8
-17 -18 -7 -6 -7
-19 -20 -5
- ∞ -21 -4 -3
- ∞ - ∞ -22 -23 -2 -1
Goal

Bellman Expectation Equation

$$\tau = (S_0, A_0, S_1, A_1, S_2, A_2, \ldots), \quad G(\tau) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_t, A_t)$$

Bellman Expectation Equation

$$\tau = (S_0, A_0, S_1, A_1, S_2, A_2, \ldots), \quad G(\tau) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_t, A_t)$$

$$\tilde{\tau} = (S_1, A_1, S_2, A_2, S_3, A_3 \ldots), \quad G(\tilde{\tau}) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_{t+1}, A_{t+1})$$

Bellman Expectation Equation

$$\tau = (S_0, A_0, S_1, A_1, S_2, A_2, \ldots), \quad G(\tau) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_t, A_t)$$

$$\tilde{\tau} = (S_1, A_1, S_2, A_2, S_3, A_3 \ldots), \quad G(\tilde{\tau}) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_{t+1}, A_{t+1})$$

$$G(\tau) = \mathcal{R}(S_0, A_0) + \gamma \sum_{t=1}^{\infty} \gamma^{t-1} \mathcal{R}(S_t, A_t) = \mathcal{R}(S_0, A_0) + \gamma G(\tilde{\tau})$$

Bellman Expectation Equation

$$\tau = (S_0, A_0, S_1, A_1, S_2, A_2, \dots), \quad G(\tau) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_t, A_t)$$

$$\tilde{\tau} = (S_1, A_1, S_2, A_2, S_3, A_3 \dots), \quad G(\tilde{\tau}) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_{t+1}, A_{t+1})$$

$$G(\tau) = \mathcal{R}(S_0, A_0) + \gamma \sum_{t=1}^{\infty} \gamma^{t-1} \mathcal{R}(S_t, A_t) = \mathcal{R}(S_0, A_0) + \gamma G(\tilde{\tau})$$

Bellman Expectation Equation для v_π

$$v_\pi(s) = \sum_a \pi(a|s) \Big(\mathcal{R}(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) v_\pi(s') \Big)$$

Решение Bellman Expectation Equation

Bellman Expectation Equation для v_π

$$v_\pi(s) = \sum_a \pi(a|s)\Big(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s')\Big)$$

Решение Bellman Expectation Equation

Bellman Expectation Equation для v_π

$$v_\pi(s) = \sum_a \pi(a|s)\Big(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s')\Big)$$

$$v_\pi(s) = \sum_a \pi(a|s)\mathcal{R}(s,a) + \gamma \sum_{s'} \sum_a \mathcal{P}(s'|s,a) v_\pi(s')$$

Решение Bellman Expectation Equation

Bellman Expectation Equation для v_π

$$v_\pi(s) = \sum_a \pi(a|s)\Big(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s')\Big)$$

$$v_\pi(s) = \sum_a \pi(a|s)\mathcal{R}(s,a) + \gamma \sum_{s'} \sum_a \mathcal{P}(s'|s,a) v_\pi(s')$$

$$\mathcal{R}_\pi(s) = \sum_a \pi(a|s)\mathcal{R}(s,a), \quad \mathcal{P}_\pi(s', s) = \sum_a \mathcal{P}(s'|s,a)$$

Решение Bellman Expectation Equation

Bellman Expectation Equation для v_π

$$v_\pi(s) = \sum_a \pi(a|s)\Big(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s')\Big)$$

$$v_\pi(s) = \sum_a \pi(a|s)\mathcal{R}(s,a) + \gamma \sum_{s'} \sum_a \mathcal{P}(s'|s,a) v_\pi(s')$$

$$\mathcal{R}_\pi(s) = \sum_a \pi(a|s)\mathcal{R}(s,a), \quad \mathcal{P}_\pi(s',s) = \sum_a \mathcal{P}(s'|s,a)$$

$$v_\pi(s) = \mathcal{R}_\pi(s) + \gamma \sum_{s'} \mathcal{P}_\pi(s',s) v_\pi(s')$$

Решение Bellman Expectation Equation

Bellman Expectation Equation для v_π

$$v_\pi(s) = \sum_a \pi(a|s)\Big(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s')\Big)$$

$$v_\pi(s) = \sum_a \pi(a|s)\mathcal{R}(s,a) + \gamma \sum_{s'} \sum_a \mathcal{P}(s'|s,a) v_\pi(s')$$

$$\mathcal{R}_\pi(s) = \sum_a \pi(a|s)\mathcal{R}(s,a), \quad \mathcal{P}_\pi(s',s) = \sum_a \mathcal{P}(s'|s,a)$$

$$v_\pi(s) = \mathcal{R}_\pi(s) + \gamma \sum_{s'} \mathcal{P}_\pi(s',s) v_\pi(s')$$

$$v_\pi = \begin{pmatrix} v_\pi(s_1) \\ \cdots \\ v_\pi(s_n) \end{pmatrix}, \mathcal{R}_\pi = \begin{pmatrix} \mathcal{R}_\pi(s_1) \\ \cdots \\ \mathcal{R}_\pi(s_n) \end{pmatrix}, \mathcal{P}_\pi = \begin{pmatrix} \mathcal{P}_\pi(s_1,s_1) & \dots & \mathcal{P}_\pi(s_1,s_n) \\ \vdots & \ddots & \vdots \\ \mathcal{P}_\pi(s_n,s_1) & \dots & \mathcal{P}_\pi(s_n,s_n) \end{pmatrix}$$

Решение Bellman Expectation Equation

Bellman Expectation Equation для v_π

$$v_\pi(s) = \sum_a \pi(a|s)\Big(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s')\Big)$$

$$v_\pi = \mathcal{R}_\pi + \gamma \mathcal{P}_\pi v_\pi$$

Решение Bellman Expectation Equation

Bellman Expectation Equation для v_π

$$v_\pi(s) = \sum_a \pi(a|s)\Big(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s')\Big)$$

$$v_\pi = \mathcal{R}_\pi + \gamma \mathcal{P}_\pi v_\pi$$

$$(E - \gamma \mathcal{P}_\pi) v_\pi = \mathcal{R}_\pi$$

Решение Bellman Expectation Equation

Bellman Expectation Equation для v_π

$$v_\pi(s) = \sum_a \pi(a|s)\Big(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s')\Big)$$

$$v_\pi = \mathcal{R}_\pi + \gamma \mathcal{P}_\pi v_\pi$$

$$(E - \gamma \mathcal{P}_\pi) v_\pi = \mathcal{R}_\pi$$

$$v_\pi = (E - \gamma \mathcal{P}_\pi)^{-1} \mathcal{R}_\pi$$

Решение Bellman Expectation Equation

Bellman Expectation Equation для v_π

$$v_\pi(s) = \sum_a \pi(a|s)\Big(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s')\Big)$$

$$v_\pi = \mathcal{R}_\pi + \gamma \mathcal{P}_\pi v_\pi$$

$$(E - \gamma \mathcal{P}_\pi) v_\pi = \mathcal{R}_\pi$$

$$v_\pi = (E - \gamma \mathcal{P}_\pi)^{-1} \mathcal{R}_\pi$$

Теорема

Если $\gamma < 1$, то существует единственное v_π решение Bellman Expectation Equation.

Iterative Policy Evaluation

Пусть задана Policy π; $v^0(s)$, $s \in \mathcal{S}$ — любая инициализация;
K — число итераций.
Для каждого $k \in \overline{0, K}$ делаем

- Для каждого $s \in \mathcal{S}$ определяем

$$v^{k+1}(s) = \sum_a \pi(a|s)\Big(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v^k(s')\Big)$$

  или сокращенно

$$v^{k+1} = \mathcal{R}_\pi + \gamma \mathcal{P}_\pi v^k$$

Теорема

Iterative Policy Evaluation сходится, то есть $v^k \to v_\pi$, $k \to \infty$.
Сходимость имеет порядок $o(mn^2)$

Action-Value Function

Action-Value Function

$$q_\pi(s, a) = \mathbb{E}_\pi[G \mid S_0 = s,\ A_0 = a]$$

Action-Value Function

Action-Value Function

$$q_\pi(s,a) = \mathbb{E}_\pi[G \mid S_0 = s,\ A_0 = a]$$

Связь с v_π

$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a), \quad q_\pi(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s')$$

Action-Value Function

Action-Value Function

$$q_\pi(s,a) = \mathbb{E}_\pi[G \mid S_0 = s,\ A_0 = a]$$

Связь с v_π

$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a), \quad q_\pi(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s')$$

Bellman Expectation Equation для q_π

$$q_\pi(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \sum_{a'} \pi(a'|s') q_\pi(s',a')$$

Policy Improvement

Частичный порядок для Policy

$$\pi' \geq \pi \quad \Leftrightarrow \quad v_{\pi'}(s) \geq v_{\pi}(s), \quad \forall s \in \mathcal{S}$$

Policy Improvement

Частичный порядок для Policy

$$\pi' \geq \pi \quad \Leftrightarrow \quad v_{\pi'}(s) \geq v_{\pi}(s), \quad \forall s \in \mathcal{S}$$

Greedy Policy Improvement

$$\pi'(a|s) = \begin{cases} 1, & \text{если } a \in \operatorname{argmax}_{a' \in \mathcal{A}} q_{\pi}(s, a') \\ 0, & \text{иначе} \end{cases}$$

Policy Improvement

Частичный порядок для Policy

$$\pi' \geq \pi \quad \Leftrightarrow \quad v_{\pi'}(s) \geq v_{\pi}(s), \quad \forall s \in \mathcal{S}$$

Greedy Policy Improvement

$$\pi'(a|s) = \begin{cases} 1, & \text{если } a \in \operatorname{argmax}_{a' \in \mathcal{A}} q_{\pi}(s, a') \\ 0, & \text{иначе} \end{cases}$$

Policy Improvement Theorem

Пусть задана Policy π. Если Policy π' определяется согласно Greedy Policy Improvement, то

$$\pi' \geq \pi$$

Optimal Policy

(Optimal) State-Value Function и Action-Value Function

$$v_*(s) = \max_{\pi} v_\pi(s), \quad q_*(s,a) = \max_{\pi} q_\pi(s,a)$$

Optimal Policy

(Optimal) State-Value Function и Action-Value Function

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

Optimal Policy Existence Theorem

Существует (оптимальная) Policy π_* такая, что

- $\pi_* \geq \pi, \ \forall \pi$
- $v_{\pi_*}(s) = v_*(s), \ \forall s \in \mathcal{S}$
- $q_{\pi_*}(s,a) = q_*(s,a), \ \forall s \in \mathcal{S}, \ \forall a \in \mathcal{A}$

Policy Iteration

Пусть инициализированы π^0 и заданы числа $L, K \in \mathbb{N}$.
Для каждого $k \in \overline{0, K}$ делаем

- (Policy evaluation) Iterative Policy Evaluation

$$v^{l+1} = \mathcal{R}_{\pi^k} + \mathcal{P}_{\pi^k} v^l, \quad l \in \overline{0, L-1}.$$

  По $v^L(s)$ построить $q^L(s, a)$.

- (Policy improvement) Greedy Policy Improvement

$$\pi^{k+1}(a|s) = \begin{cases} 1, & \text{если } a \in \operatorname{argmax}_{a' \in \mathcal{A}} q^L(s, a') \\ 0, & \text{иначе} \end{cases}$$

Policy Iteration

Пусть инициализированы π^0 и заданы числа $L, K \in \mathbb{N}$.
Для каждого $k \in \overline{0, K}$ делаем

- (Policy evaluation) Iterative Policy Evaluation

$$v^{l+1} = \mathcal{R}_{\pi^k} + \mathcal{P}_{\pi^k} v^l, \quad l \in \overline{0, L-1}.$$

  По $v^L(s)$ построить $q^L(s, a)$.

- (Policy improvement) Greedy Policy Improvement

$$\pi^{k+1}(a|s) = \begin{cases} 1, & \text{если } a \in \operatorname{argmax}_{a' \in \mathcal{A}} q^L(s, a') \\ 0, & \text{иначе} \end{cases}$$

Теорема

Policy Iteration сходится, то есть $\pi^k \to \pi_*$, $k \to \infty$. Сходимость имеет порядок $o(mn^2)$

Bellman Optimality Equations

Bellman Optimality Equations для v_*

$$v_*(s) = \max_{a \in \mathcal{A}} \left(\mathcal{R}(s, a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s, a) v_*(s') \right)$$

Bellman Optimality Equations

Bellman Optimality Equations для v_*

$$v_*(s) = \max_{a \in \mathcal{A}} \left(\mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) v_*(s') \right)$$

Bellman Optimality Equations для q_*

$$q_*(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) \max_{a' \in \mathcal{A}} q_*(s',a')$$

Bellman Optimality Equations

Bellman Optimality Equations для v_*

$$v_*(s) = \max_{a\in\mathcal{A}} \left(\mathcal{R}(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s,a) v_*(s')\right)$$

Bellman Optimality Equations для q_*

$$q_*(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s,a) \max_{a'\in\mathcal{A}} q_*(s',a')$$

Связь v_* и q_*

$$v_*(s) = \max_{a\in A} q_*(s,a), \quad q_*(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s,a) v_*(s')$$

Bellman Optimality Equations

Bellman Optimality Equations для v_*

$$v_*(s) = \max_{a \in \mathcal{A}} \left(\mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) v_*(s') \right)$$

Bellman Optimality Equations для q_*

$$q_*(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) \max_{a' \in \mathcal{A}} q_*(s',a')$$

Связь v_* и q_*

$$v_*(s) = \max_{a \in A} q_*(s,a), \quad q_*(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) v_*(s')$$

Связь π_* и q_*

$$\pi_*(a|s) = \begin{cases} 1, & \text{если } a \in \operatorname{argmax}_{a' \in \mathcal{A}} q_*(s,a') \\ 0, & \text{иначе} \end{cases}$$

Value Iteration

$v^0(s)$, $s \in \mathcal{S}$ — любая инициализация; K — число итераций.
Для каждого $k \in \overline{0, K}$ делаем

- Для каждого $s \in \mathcal{S}$ определяем

$$v^{k+1}(s) = \max_{a \in \mathcal{A}} \left(\mathcal{R}(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) v^k(s') \right)$$

Теорема

Value Iteration сходится, то есть $v^k \to v_*$, $k \to \infty$. Сходимость имеет порядок $o(mn^2)$

- Определения v_π, q_π, v_*, q_*, π_* будут использоваться в самом общем случае MDP (когда $\mathcal{S}$ и $\mathcal{A}$ не обязательно конечные, и $\mathcal{P}$ и $\mathcal{R}$ не обязательно известны)

- Bellman Expectation Equation для v_π и q_π, и Bellman Optimality Equation для v_* и q_* (в том виде, в котором они представлены), а также Policy Improvement Theorem и Optimal Policy Existence Theorem справедливы в случае, когда в MDP $\mathcal{S}$ и $\mathcal{A}$ конечны, но $\mathcal{P}$ и $\mathcal{R}$ не обязательно известны

- Алгоритмы Policy Iteration и Value Iteration работают в случае, когда в MDP $\mathcal{S}$ и $\mathcal{A}$ конечны, и $\mathcal{P}$ и $\mathcal{R}$ известны

Организационные вопросы

- Пятница, 17:50, аудитория 622
- Отчетность: домашние работы
- Странчика курса: https://github.com/imm-rl-lab/UrFU_course
- E-mail для связи:

ВОПРОСЫ?