30.11.18 Reinforcement Learning

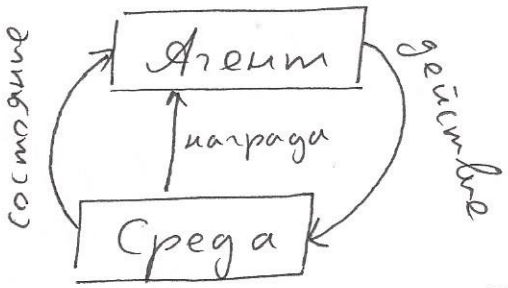
ML $\{x_i, y_i\}_{i=1}^N$, $f(x, \Theta)$

обучаем параметры Θ

$$F(\Theta) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i, \Theta)) + \lambda R(\Theta) \to \min_{\Theta}$$

отложенное вознаграждение?
взаимодействие со средой?

RL



$s \in S$ - сост. среды
$a \in A$ - действие агента

$\pi(a|s)$ - политика агента
$p(s'|s,a)$ - вер-ти переходов
$r(s,a)$ - награда

MDP, Markov Decision Process

$\{s_0, a_0, r_0, s_1, a_1, r_1, \dots\}$

суммарный дисконтированный реворд

$$R_{t_0} = r(s_0, a_0) + \gamma r(s_1, a_1) + \gamma^2 r(s_2, a_2) + \dots, \quad \gamma \in (0, 1)$$

$$\to \max_{\pi}$$

$$R_t = r(s_t, a_t) + \gamma r(s_{t+1}, a_{t+1}) + \dots$$

Value Iteration

$$V^{\pi}(s) = \mathbb{E}[R_t | s_t = s]$$

$$V^*(s) = \max_{\pi} V^{\pi}(s)$$

$$\pi^*(s) = \arg\max_{\pi} V^{\pi}(s)$$

$$R_0 = r(s_0, a_0) + \gamma r(s_1, a_1) + \gamma^2 r(s_2, a_2) + \dots = r(s_0, a_0) + \gamma_1$$

$$V^*(s) = \max_{\pi} \mathbb{E}[R_0 \mid s_0 = s] = \max_{a_0, a_1, a_2, \dots} \mathbb{E}[R_0 \mid s_0 = s] =$$

марковский процесс, отсюда равенство

$$= \max_{a_0, a_1, a_2, \dots} \left[r(s_0, a_0) + \gamma \mathbb{E}_{p(s' \mid s_0, a_0)} \mathbb{E}[R_1 \mid s_1 = s'] \right] =$$

$$= \max_{a_0} \left[r(s_0, a_0) + \gamma \mathbb{E}_{p(s' \mid s_0, a_0)} \underbrace{\max_{a_1, a_2, \dots} \mathbb{E}[R_1 \mid s_1 = s']}_{V^*(s')} \right]$$

$$V^*(s) = \max_a \left(r(s, a) + \gamma \mathbb{E}_{p(s' \mid s, a)} V^*(s') \right)$$

оптимальное ур-е Беллмана, $x = f(x)$

$x_{k+1} = f(x_k)$

fixed point iteration, сжимающее отображение

$$V_{new}(s) = \max_a \left(r(s, a) + \gamma \mathbb{E}_{p(s' \mid s, a)} V_{obs}(s') \right) \quad \forall s$$

$$\pi^*(s) = \arg\max_a \left(r(s, a) + \gamma \mathbb{E}_{p(s' \mid s, a)} V^*(s') \right)$$

Пример

терм. сост.

r:

0	0	0	1
0		0	-1
0	0	0	0

V_0:

0	0	0	1
0		0	-1
0	0	0	0

$\rightarrow V_1$:

0	0	γ	1
0		0	-1
0	0	0	0

$\rightarrow V_2$:

0	γ^2	γ	1
0		γ^2	-1
0	0	0	0

$\rightarrow \dots \rightarrow V_5$:

γ^3	γ^2	γ	1
γ^4		γ^2	-1
γ^5	γ^4	γ^3	γ^4

Q-learning

$$Q^{\pi}(s,a) = \mathbb{E}[R_t \mid s_t = s, a_t = a]$$

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a)$$

$$\pi^*(s) = \arg\max_a Q^*(s,a)$$

$$V^*(s) = \max_a Q^*(s,a)$$

$$Q^*(s,a) = \max_{a_1, a_2, a_3, \dots} \mathbb{E}[R_0 \mid s_0 = s, a_0 = a] =$$

$$= \max_{a_1, a_2, a_3, \dots} [r(s_0,a_0) + \gamma \mathbb{E}_{p(s'|s_0,a_0)} \mathbb{E}[R_1 \mid s_1 = s_1', a_1 = a_1]] =$$

$$= r(s_0,a_0) + \gamma \mathbb{E}_{p(s'|s_0,a_0)} \max_{a_1} \underbrace{\max_{a_2, a_3, \dots} \mathbb{E}[R_1 \mid s_1 = s_1', a_1 = a_1]}_{Q^*(s', a_1)}$$

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)} \max_{a'} Q^*(s',a')$$

$$\forall s, a$$

$$F(Q^*) = \frac{1}{|S||A|} \sum_{s,a} \left(Q^*(s,a) - r(s,a) - \gamma \mathbb{E}_{p(s'|s,a)} \max_{a'} Q^*(s',a') \right)^2 \to \min_{Q^*}$$

1) (Эмулируем (s, a, r, s')

2) $Q^*_{new}(s,a) = Q^*_{old}(s,a) - \alpha_{lr} \cdot 2 \left(Q^*_{old}(s,a) - r(s,a) - \gamma \max_{a'} Q^*_{old}(s',a') \right)$

Exploration - Exploitation dilemma

локально-оптимальные действия?

ε-жадная стратегия:

$$\pi(s) = \begin{cases} \arg\max_a Q(s,a), & \text{с вер-ю } 1-\varepsilon \\ \sim R, & \text{с вер-ю } \varepsilon \end{cases}$$

Softmax - стратегия

$$\pi(a|s) \sim Softmax(Q(s,a)/\tau)$$

сэмплинг, $\tau \to \infty$ и

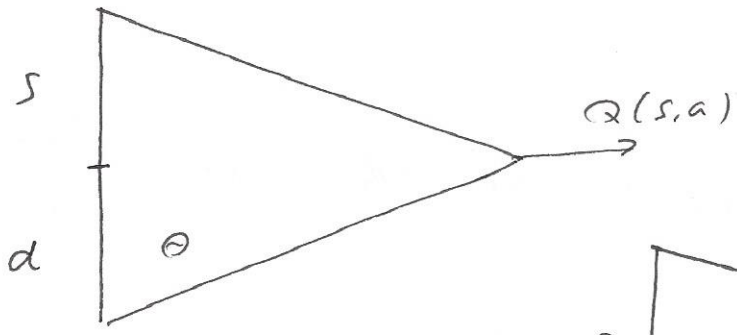
$\tau \to 0$ argmax

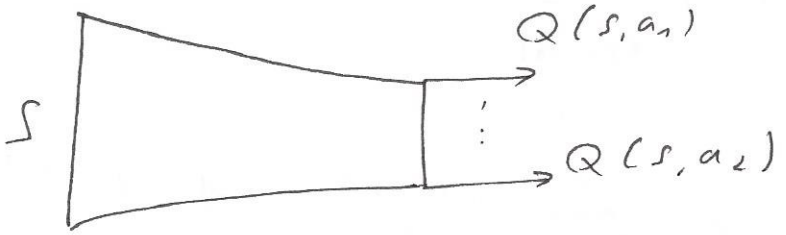
$$Q^*(s,a) \approx Q^*(s,a|\theta)$$

параметризация функции

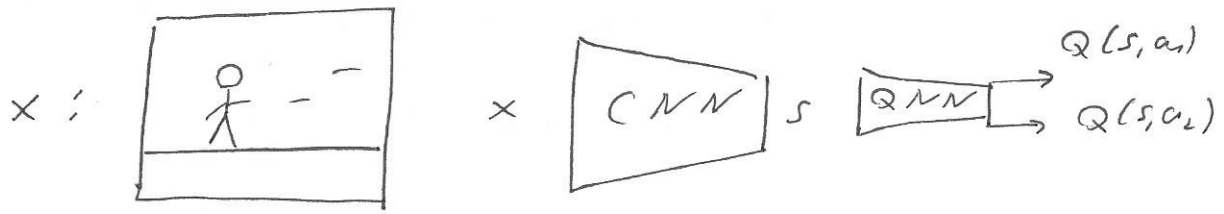
$$F(\theta) = \frac{1}{|S||A|} \sum_{s,a} (Q(s,a|\theta) - r(s,a) - \gamma \mathbb{E}_{p(s'|s,a)} \max_{a'} Q(s',a'|\theta))^2 \to \min_{\theta}$$

$$\theta_{new} = \theta_{old} - \alpha \cdot 2 (Q(s,a|\theta_{old}) - \dots) \nabla_\theta Q(s,a|\theta_{old})$$





DQN



Инициализация θ, M

/ / /

Инициализируем Θ, M

Для эпизодов $1 \dots P$

инициализация x_1

$s_1 = CNN(x_1)$

для $t = 1 \dots T$:

$$a_t = \begin{cases} \underset{a}{\operatorname{argmax}}\, Q(s_t, a; \Theta) & \text{с вер-ю } 1-\varepsilon \\ \sim R, & \varepsilon \end{cases}$$

$(s_t, a_t, \hat{r}_t, x_{t+1} \to s_{t+1}) \to M$ } experience replay, mini-batch

$(s_j, a_j, r_j, s_{j+1}) \leftarrow M$ }

$$y_j = \begin{cases} r_j + \gamma \max\limits_{a'} Q(s_j, a' | \Theta), & s_j - \text{нетерм.} \\ r_j, \text{ если } & s_j - \text{терм.} \end{cases}$$

$$\Theta_{new} = \Theta_{old} - \alpha \nabla_{\Theta} (Q(s_j, a_j | \Theta) - y_j)^2 \; Q(s_j, a_j | \Theta)$$