① TD-learning

(s,a,r,s')

оценка для Q

$$Q_\pi(S_t,A_t) := Q_\pi(S_t,A_t) + \alpha\left(R_{t+1} + \gamma\, Q_\pi(S_{t+1},A_{t+1}) - Q_\pi(S_t,A_t)\right)$$

TD:

} backup diagrams

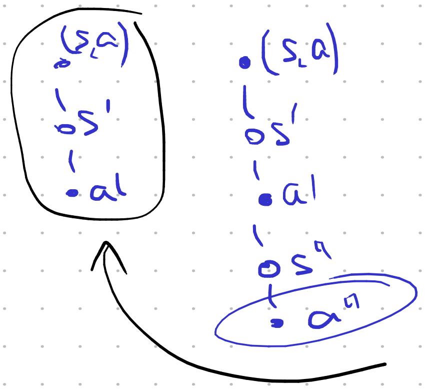
MC:

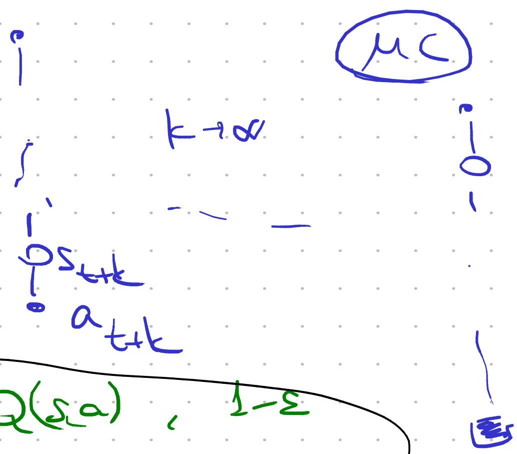
Bellman:

V_π:

n-step TD:

$$Q_\pi(S_t,A_t) := \dots\ \alpha\left(R_{t+1} + \gamma R_{t+2} + \gamma^2 Q_\pi(S_{t+2},A_{t+2}) - Q_\pi(S_t,A_t)\right)$$





SARSA

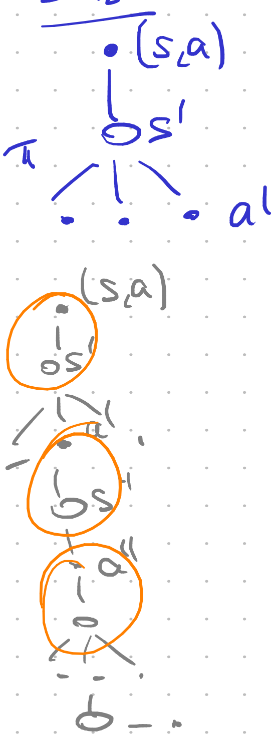
On-policy TD control:

$$\pi_Q(s) = \begin{cases} \arg\max_a Q(s,a), & 1-\varepsilon \\ \text{unif}, & \varepsilon \end{cases}$$

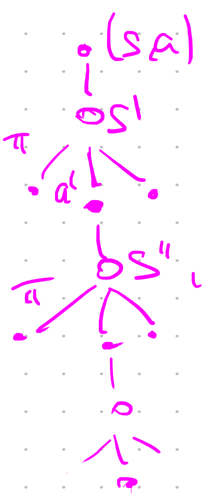
- выбираю a по страт.
- $Q(s,a) := Q(s,a) + \alpha\left(r + \gamma \cdot Q(s',a') - Q(s,a)\right)$

Expected Sarsa

(s,a,r,s',a')



$$Q(s,a) := Q(s,a) + \alpha\left(r + \gamma \cdot \sum_{a'} \pi(a'|s')\, Q(s',a') - Q(s,a)\right)$$

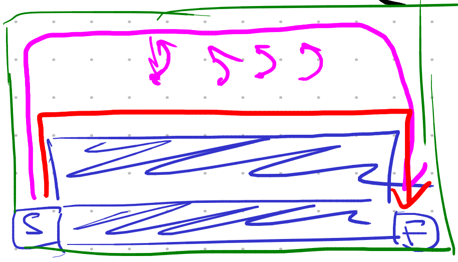


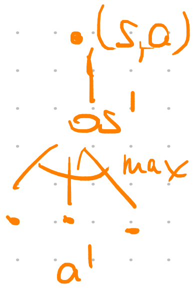
$$r + \gamma \cdot \sum_{a_1 \neq a'} \pi(a_1|s') \cdot Q(s',a_1) +$$

$$+ \gamma \cdot \pi(a'|s') \cdot \left(r' + \gamma \cdot \sum_{a_2} \pi(a_2|s'')\, Q(s'',a_2)\right)$$

② Q-learning (1989, Watkins) — off policy TD control

$$Q(s,a) := Q(s,a) + \alpha\Big(r + \gamma \cdot \max_{a'} Q(s',a') - Q(s,a)\Big)$$





$$\mathbb{E}\big[Q_1(s, \arg\max Q_2)\big]$$

$$Q(s,a) \rightsquigarrow r_t + \gamma \cdot \max_{a'} Q(s',a')$$

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

$$\mathbb{E}_\pi[G_t \mid s,a]$$

$$\mathbb{E}[\max] \neq \max \mathbb{E}[\cdot]$$

$$r \sim \mathcal{N}\left(-\tfrac{1}{10}, 1\right)$$

$$Q(s,a) = \max Q(s',a')$$

Winner's curse

Mechanism design

$$\hat{X}_1, \hat{X}_2, \dots, \hat{X}_k$$

$$\hat{X}_i \sim \mathcal{N}(\hat{X}_i \mid X, \sigma^2)$$

Double Q-learning (double DQN)

$$Q_1(s,a) \qquad Q_2(s,a)$$

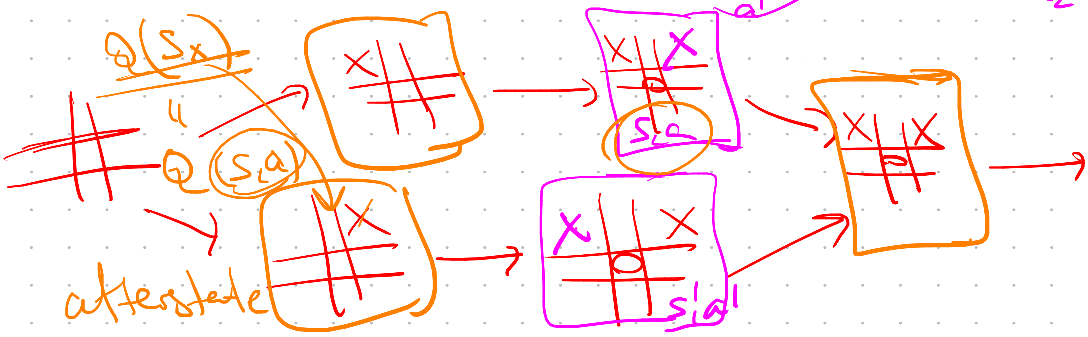
(s,a,r,s'):

$$- p = \tfrac{1}{2}: \; Q_1(s,a) := Q_1(s,a) + \alpha\big[r + \gamma \cdot Q_2(s', \arg\max_{a'} Q_1(s',a')) - Q_1(s,a)\big]$$

$$- p = \tfrac{1}{2}: \; Q_2(s,a) := Q_2(s,a) + \alpha\big[r + \gamma \cdot Q_1(s', \arg\max_{a'} Q_2(s',a')) - Q_2(s,a)\big]$$

③ Afterstates

$$\frac{Q(s,x)}{4}$$



Tabular RL

s	a	Q
.	.	.

$s, a \to \boxed{\hat{Q}} \leftarrow \bar{w}$

(4) Approximate RL

$s \to \boxed{\hat{V}} \to \hat{V}(s, \bar{w}) \approx V(s)$, $\bar{w}$

$s \to \boxed{\hat{Q}} \to \hat{Q}(s, a_k, \bar{w})$, $a_1, \dots, a_k$, $\bar{w}$ $\quad \approx Q(s, a_k)$

$$\cancel{L(\bar{w}) = \sum_{s \in S} \left(V_\pi(s) - \hat{V}_\pi(s, \bar{w})\right)^2} \xrightarrow[\bar{w}]{} \min$$

GD

$F(\bar{x}) \to \min$

$\bar{x} := \bar{x} - \alpha \cdot \nabla_{\bar{x}} F(\bar{x})$

$$F(\bar{x}) = \frac{1}{N} \sum_{n=1}^{N} \ell(d_n, \bar{x}) = \mathbb{E}_{\text{unif}}\left[\ell(d, \bar{x})\right] \approx \frac{1}{M} \sum_{m=1}^{M} \ell(d_m, \bar{x})$$

SGD: mini-batch

stochastic gradient descent

$$L(\bar{w}) = \sum_{s \in S} \mu(s) \cdot \overbrace{\left(V_\pi(s) - \hat{V}_\pi(s, \bar{w})\right)^2}^{\ell(s, \bar{w})} \xrightarrow[\bar{w}]{} \min$$

$\mu(s) =$ „гопо вер., пробаб. в $s \in S$"

$\mu(s) = Pr_\pi[S_t = s]$

$\mathbb{E}_\pi[R + \gamma \cdot V_\pi(s', \bar{w})]$

- новай опит (s, a, r, s'):

$$\bar{w} := \bar{w} - \alpha \nabla_{\bar{w}} \ell(s, \bar{w}) =$$

$$= \bar{w} + 2\alpha \left(V_\pi(s) - \hat{V}_\pi(s, \bar{w})\right) \cdot \nabla_{\bar{w}} \hat{V}_\pi(s, \bar{w})$$

G_t (MC) $\qquad r + \gamma \cdot \hat{V}_\pi(s', \bar{w})$ (TD)

(3) Off-policy MC

Gradient MC estimation:

$$\bar{w} := \bar{w} + \alpha \left(G_t - \hat{V}_\pi(s, \bar{w})\right) \cdot \nabla_{\bar{w}} V_\pi(s, \bar{w})$$

Semi-gradient TD estimation

$$\bar{w} := \bar{w} + \alpha\Big(R_{t+1} + \gamma \cdot \hat{V}_\pi(S_{t+1}, \bar{w}) - \hat{V}_\pi(S_t, \bar{w})\Big) \cdot \nabla_{\bar{w}} \hat{V}(S_t, \bar{w})$$

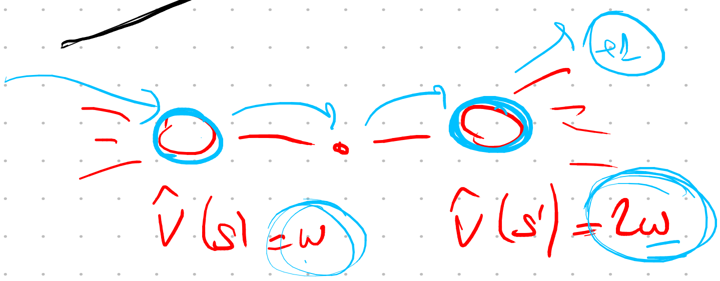
$\approx V(S_t)$

Semi-gradient Sarsa

n-step, expected Sarsa

$$\bar{w} := \bar{w} + \alpha\Big(R_{t+1} + \gamma \, \hat{Q}(S_{t+1}, A_{t+1}, \bar{w}) - \hat{Q}(S_t, A_t, \bar{w})\Big) \cdot \nabla_{\bar{w}} \hat{Q}(S_t, A_t, \bar{w})$$

~~Off-policy semi-gradient TD control~~



semi-gr. Sarsa

Deadly triad

- approximation
- bootstrapping
- off-policy

Q-learning

DQN

experience replay

(5) Policy gradient

s — $\boxed{Q}$ ≡ $Q(s,a)$ } argmax

s — $\boxed{\pi}$ ≡ $\pi(a|s, \bar{\theta}) = \Pr[A_t = a \mid S_t = s]$

$\bar{\theta}$

s — $\boxed{\pi}$ ≡ $h(a, s, \bar{\theta})$ — $\boxed{\text{softmax}}$ — $\pi(a|s)$

$\bar{\theta}$

$$\pi(a|s, \theta) = \frac{e^{h(a,s,\theta)}}{\sum_{a'} e^{h(a', s, \theta)}}$$

s_0 – fixed state (?)

$$J(\bar{\theta}) = V_{\pi_{\bar{\theta}}}(s_0) = \mathbb{E}_{\pi_{\bar{\theta}}}[G_t \mid s_0] \xrightarrow{\bar{\theta}} \max$$

$$\bar{\theta} := \bar{\theta} + \alpha \cdot \nabla_{\bar{\theta}} J(\bar{\theta})$$

$= \mathbb{E}_{\text{Traj}(\pi_{\bar{\theta}})}$

Policy gradient theorem

$(\gamma = 1)$, $r = 0$

$$\nabla_\theta V_{\pi_\theta}(s) = \nabla_\theta \Big[\sum_a \pi_{\bar{\theta}}(a|s) \cdot Q_{\pi_\theta}(s,a)\Big] =$$

$$= \sum_a \Big[\big(\nabla_\theta \pi_\theta(a|s)\big) \cdot Q_{\pi_\theta}(s,a) + \pi_\theta(a|s) \cdot \nabla_\theta Q_{\pi_\theta}(s,a) \Big] =$$

$$= \sum_a \left[Q \cdot \nabla\pi(a|s) + \pi_\theta(a|s) \cdot \nabla_\theta \left[\sum_{r,s'} p(r,s'|s,a)\left(r + V_{\pi_\theta}(s')\right) \right] \right] =$$

$$= \sum_a \left[Q \cdot \nabla\pi(a|s) + \pi_\theta(a|s) \cdot \sum_{r,s'} p(r,s'|s,a)\left(\cancel{r+} \nabla_\theta V_{\pi_\theta}(s')\right) \right] =$$

$$= \sum_a \left[(Q \cdot \nabla_\pi)(a,s) + \pi_\theta(a|s) \sum_{r,s'} p(r,s'|s,a) \times \right.$$

$$\left. \times \sum_{a'} \left[(Q \cdot \nabla_\pi)(a',s') + \pi_\theta(a'|s') \cdot \nabla_\theta Q_{\pi_\theta}(s',a') \right] \right]$$

$$= \sum_{s'} \left[\left(\sum_a Q_{\pi_\theta}(s',a) \cdot \nabla_{\pi_\theta}(a|s') \right) \cdot \sum_{k=0}^{\infty} \Pr\left[\begin{matrix} s \to s' \\ \text{за } k \text{ шагов} \\ \text{по } \pi \end{matrix} \right] \right]$$

2-й шаг: $(s') \times \left(\sum_a \pi_\theta(a|s) \cdot \sum_r p(r,s'|s,a) \right)$

$$\nabla_{\bar\theta} J(\bar\theta) = \nabla_{\bar\theta} V_\pi(s_0) = \sum_s \left(\underbrace{\sum_{k=0}^{\infty} \Pr\left[\begin{matrix} s_0 \to s \\ \text{за } k \text{ шагов} \\ \text{по } \pi \end{matrix} \right]}_{\mathbb{E}\,\#[t:\ S_t = s]\ \propto\ \mu(s)} \right) \left(\sum_a Q_{\pi_\theta}(s,a) \nabla_{\bar\theta} \pi_\theta(a|s) \right)$$

$$\boxed{\nabla_{\bar\theta} J(\bar\theta) \propto \sum_s \mu(s) \sum_a Q_{\pi_\theta}(s,a) \nabla_\theta \pi_\theta(a|s)}$$

1) All-actions method (actor-critic)

$$\bar\theta := \bar\theta + \alpha \cdot \sum_a \hat{Q}(s,a|\bar w) \cdot \nabla_\theta \pi(a|s,\bar\theta)$$

2) REINFORCE (Williams, 1992)

$$\sum_a \pi_\theta(a|S_t) \cdot \ldots$$

$$\nabla_{\bar\theta} J(\bar\theta) \propto \mathbb{E}_{\pi_\theta}\left[\sum_a Q_{\pi_\theta}(S_t, a)\, \nabla_\theta \pi_\theta(a|S_t)\right] =$$

$$= \mathbb{E}_{\pi}\left[\sum_a \pi_\theta(a|S_t) \cdot Q_{\pi_\theta}(S_t,a) \cdot \frac{\nabla_\theta \pi_\theta(a|S_t)}{\pi_\theta(a|S_t)}\right] =$$

$$= \mathbb{E}_{\pi_\theta}\left[\underbrace{Q_{\pi_\theta}(S_t,A_t)}_{= \mathbb{E}_{\pi_\theta}[G_t | S_t, A_t]} \cdot \frac{\nabla_\theta \pi_\theta(A_t|S_t)}{\pi_\theta(A_t|S_t)}\right] =$$

$$= \mathbb{E}_{\pi_\theta}\left[G_t \cdot \frac{\nabla_\theta \pi_\theta(A_t|S_t)}{\pi_\theta(A_t|S_t)}\right]$$

REINFORCE

$$\bar\theta := \bar\theta + \alpha \cdot G_t \cdot \nabla_\theta\left[\log \pi_\theta(A_t | S_t, \bar\theta)\right]$$

3) REINFORCE w/ baselines

$Q - b(s)$

$$\nabla J(\bar\theta) = \sum_s \mu(s) \sum_a Q_\pi(s,a)\, \nabla_\theta \pi(a|s)$$

$$\sum_s \mu(s) \sum_a \left(Q - b(s)\right) \cdot \nabla_\theta \pi(a|s) =$$

$$= \sum_s \ldots - \sum_s \mu(s) \sum_a b(s) \cdot \nabla_\theta \pi(a|s)$$

$$= \sum_a \nabla_\theta \pi(a|s) =$$

$$= \nabla_\theta \sum_a \pi(a|s) = \nabla_\theta 1 = 0$$

Actor-critic

$$\nabla J(\bar\theta) \propto \sum_s \mu(s) \sum_a \left(Q_\pi(s,a) - \hat V(s, \bar w)\right) \cdot \nabla_\theta \pi(a|s)$$

- gen. $S_0, A_0, R_1 \dots$ ———, $G = 0$
- for $t = T-1, \dots, 0$
  - $G = R_t + \gamma G$
  - $\bar{w} := \bar{w} + \alpha_w \left(G - \hat{V}(S_t, \bar{w})\right) \nabla_{\bar{w}} \hat{V}(S_t, \bar{w})$
  - $\bar{\theta} := \bar{\theta} + \alpha_\theta \left(G - \hat{V}(S_t, \bar{w})\right) \cdot \nabla_{\bar{\theta}} \log \pi(A_t | S_t, \bar{\theta})$