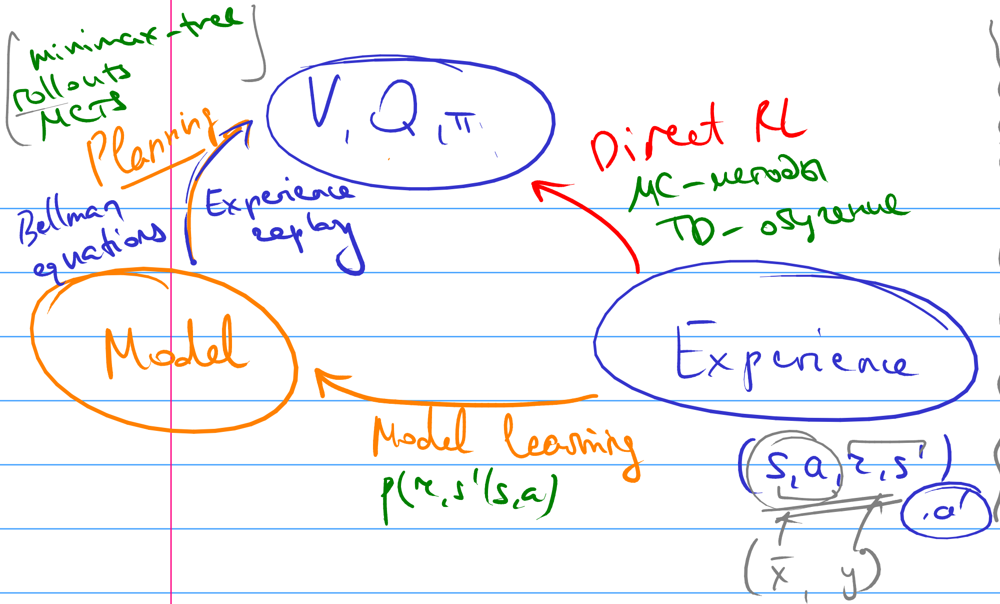
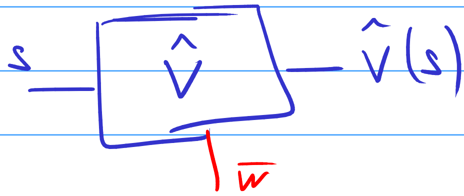


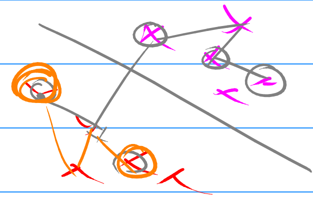
Bellman eq:



Approximate RL

веса, отраж. частоту посещ. состояний

$$L(\bar{w}) = \sum_{s \in S} \mu(s) \left(V(s) - \hat{V}(s, \bar{w})\right)^2$$



$$\nabla_{\bar{w}} L = -2 \sum_{s} \mu(s) \left(V(s) - \hat{V}(s, \bar{w})\right) \nabla_{\bar{w}} \hat{V}(s, \bar{w})$$

S_t: $\quad \bar{w} := \bar{w} + \alpha \left(V(S_t) - \hat{V}(S_t, \bar{w})\right) \nabla_{\bar{w}} \hat{V}(S_t, \bar{w})$ — Stochastic GD

$$\approx U_t$$

S_t: $\quad \bar{w} := \bar{w} + \alpha \left(U_t - \hat{V}(S_t, \bar{w})\right) \nabla_{\bar{w}} \hat{V}(S_t, \bar{w})$

$$\mathbb{E}\left[U_t \mid S_t = s\right] = V(s)$$

$U_t = G_t \Rightarrow$ Gradient MC

evaluation

$\pi, \quad V_\pi(s) = ?$

- generate $S_0, A_0, R_1, S_1, \ldots, S_T$
- $\forall t = T-1, \ldots, 0$
  - $\bar{w} := \bar{w} + \alpha \left(G_t - \hat{V}(S_t, \bar{w})\right) \nabla_{\bar{w}} \hat{V}(S_t, \bar{w})$

control

on-policy

off-policy

$$\rho_t = \frac{\pi(A_t | S_t)}{b(A_t | S_t)}$$

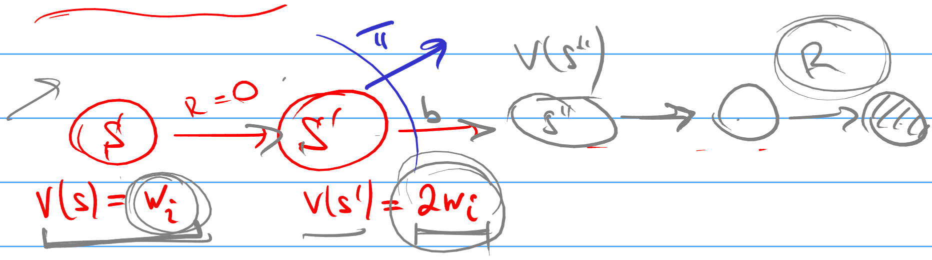
Approximate TD

$$\bar{w} := \bar{w} + \alpha \Big(\underbrace{R + \gamma \hat{V}(s', \bar{w})}_{U_t} - \hat{V}(s, \bar{w}) \Big) \cdot \nabla_{\bar{w}} \hat{V}(s, \bar{w})$$

Semi-gradient TD

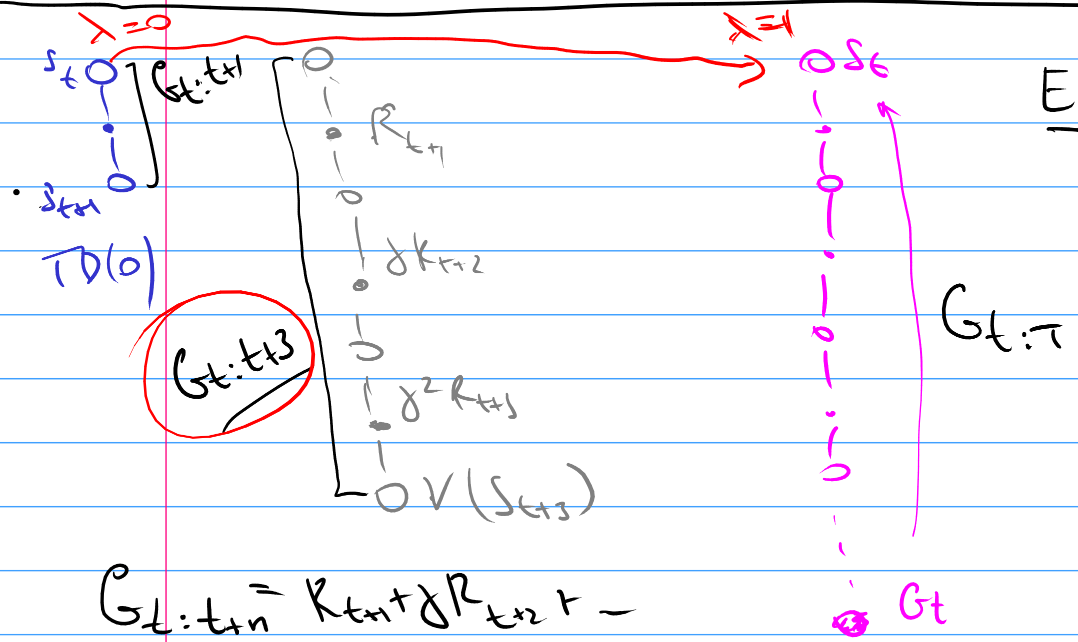
~~Gradient TD~~

off-policy TD



The deadly triad:

1) approximation
2) bootstrapping
3) off-policy



Eligibility traces

$$G_{t:t+n} = R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{n-1} R_{t+n} + \gamma^n V(S_{t+n})$$

$$G_t^{\lambda} = (1-\lambda) G_{t:t+1} + (1-\lambda)\lambda G_{t:t+2} + (1-\lambda)\lambda^2 G_{t:t+3} + \dots$$

$=$

$$\bar w := \bar w + \alpha\left(G_t^\lambda - \hat V(S_t, \bar w)\right) \nabla_{\bar w} \hat V(S_t, \bar w)$$

Semi-gradient TD(λ):

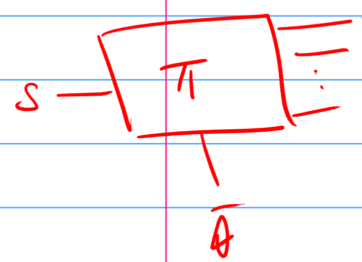
$\lambda = 0 \Rightarrow$ TD

$\lambda = 1 \Rightarrow$ MC

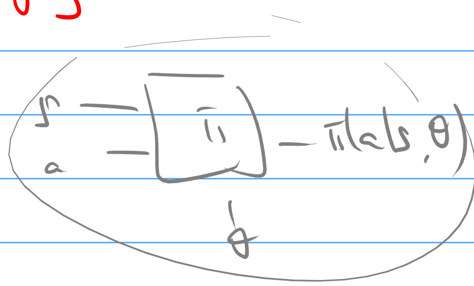
- init S_0, $\bar z = 0$
- $\forall t = 0, \dots T$:
  - избери A, получи R, S' в S
  - $\bar z := \underline{\gamma \cdot \lambda}\, \bar z + \nabla_{\bar w} \hat V(S_t, \bar w)$
  - $\bar w := \bar w + \alpha\left(R + \gamma \hat V(S', \bar w) - \hat V(S, \bar w)\right) \bar z$

Policy gradient

Q, V ~~(state-value / action-value) methods~~



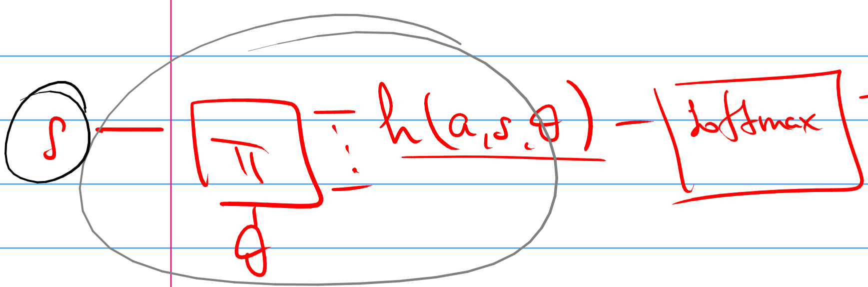
$\pi(a|s, \theta) = \Pr[A_t = a \mid S_t = s, \bar\theta]$



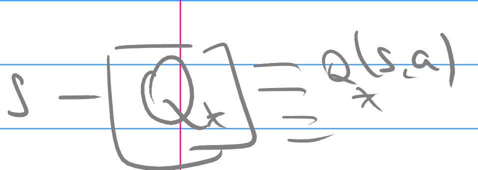
$J(\theta) = ?$

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

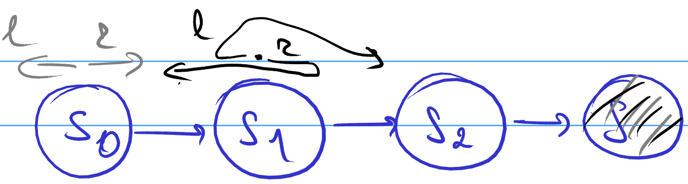
Actor-critic methods: $\pi(a|s, \bar\theta) + V(s, \bar w)$



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



$\Rightarrow \pi := \arg\max_a Q_*(s, a)$



$$J(\bar{\theta}) = V_{\pi_\theta}(s_0) \qquad \underline{\nabla_{\bar\theta} J(\bar\theta) = ?}$$

$$V_{\pi_\theta}(s) = \sum_a \pi(a|s,\theta)\, Q_{\pi_\theta}(s,a)$$

$$\nabla_\theta V_{\pi_\theta}(s) = \sum_a \left(\left(\nabla_\theta \pi(a|s,\theta)\right) \cdot Q_{\pi_\theta}(s,a) + \pi(a|s,\theta) \cdot \nabla_\theta Q_{\pi_\theta}(s,a) \right) =$$

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

$\gamma < 1$

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

$$= \sum_a \Big(Q_\pi(s,a) \cdot \nabla_\theta \pi(a|s) + \pi(a|s,\theta) \cdot \sum_{s',r} p(s',r|s,a) \Big(r +$$

$$+ \sum_{a'} \left(\nabla_\theta \pi(a'|s') \cdot Q_\pi(s',a') + \pi(a'|s',\theta) \cdot \nabla_\theta Q_\pi(s',a') \right) = \Big)$$

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

$$\nabla_\theta J(\theta) = \nabla_\theta V_\pi(s_0) = \sum_s \left(\sum_{t=0}^{\infty} \Pr\left[s_0 \to s' \text{ за } t \text{ шагов по } \pi\right] \right) \cdot \boxed{\sum_a \nabla\pi \cdot Q_\pi}$$

$$= \sum_s \eta(s) \cdot \sum_a \nabla_\theta \pi(a|s) \cdot Q_\pi(s,a) \quad \propto$$

$= \mathbb{E}[\#$ попад. в $s]$

$$\propto \sum_s \mu(s) \cdot \sum_a \nabla_\theta \pi(a|s)\, Q_\pi(s,a)$$

доля норм. сост. s

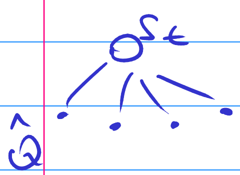
Policy gradient theorem

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

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

① All-actions

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



② REINFORCE

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

$$= \mathbb{E}_\pi \left[\sum_a \pi(a|s) \cdot \frac{1}{\pi(a|s)} Q_\pi(S_t,a) \nabla_\theta \pi(a|S_t,\theta) \right] =$$

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

$$Q_\pi(S_t, A_t) = \mathbb{E}_\pi[G_t | S_t, A_t]$$

$$J(\theta) \propto \mathbb{E}_\pi \left[G_t \cdot \frac{\nabla_\theta \pi(A_t|S_t,\theta)}{\pi(A_t|S_t,\theta)} \right]$$

$$\bar{\theta} := \bar{\theta} + \alpha G_t \cdot \frac{\nabla_\theta \pi(A_t|S_t,\theta)}{\pi(A_t|S_t,\theta)} = \bar{\theta} + \alpha G_t \nabla_\theta \ln \pi(A_t|S_t,\theta)$$

REINFORCE:

- loop
  - gen. ep. $S_0, A_0, R_1, S_1, \dots, S_T$
  - for $t = 0, \dots, T-1$:
    - $G := R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{T-t-1} R_T$

$$- \theta := \theta + \alpha \gamma^t G \nabla \ln \pi(A_t | S_t, \theta)$$

$$\nabla J(\bar{\theta}) \propto \sum_s \mu(s) \sum_a \underline{Q_\pi(s,a)} \cdot \nabla_\theta \pi(a|s,\theta)$$

$b(s)$ – baseline

$$\sum_s \mu(s) \sum_a \left(Q_\pi(s,a) - \underline{b(s)}\right) \nabla_\theta \pi(a|s,\theta) =$$

$$= \sum_s \mu(s) \left(\sum_a Q \nabla \pi - \cancel{\sum_a b(s) \underline{\nabla_\theta} \pi(a|s,\theta)} \right)$$

$$b(s) \cdot \nabla_\theta \Big(\underbrace{\sum_a \pi(a|s,\theta)}_{=1} \Big)$$

REINFORCE w/ baseline:

$$- \theta := \theta + \alpha \left(G_t - \underbrace{b(S_t)}_{= \hat{V}(S_t, \bar{w})}\right) \cdot \nabla_\theta \ln \pi(A_t | S_t, \theta)$$

$$- \bar{w} := \bar{w} + \alpha' \left(\underbrace{G_t}_{R_{t+1} + \gamma \hat{V}(S_{t+1}, \bar{w})} - \hat{V}(S_t, \bar{w}) \right) \nabla_{\bar{w}} \hat{V}(S_t, \bar{w})$$