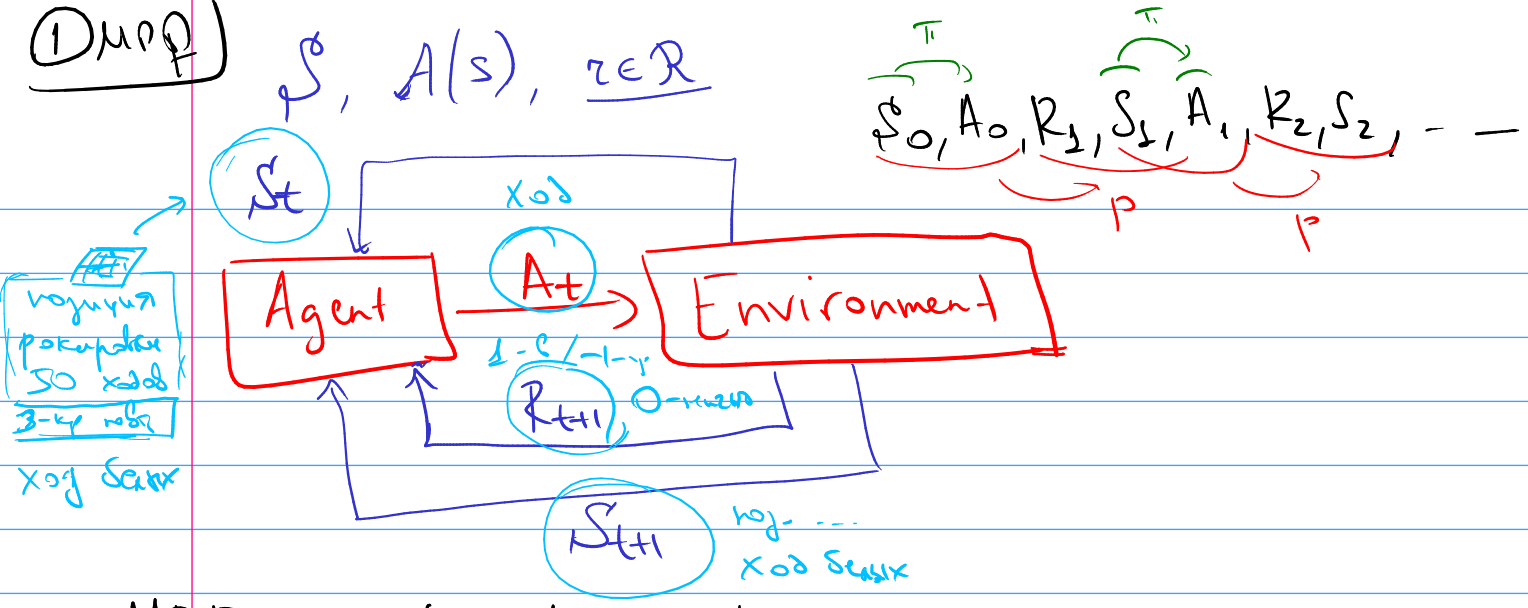
① MDP

S, $A(s)$, $r \in R$



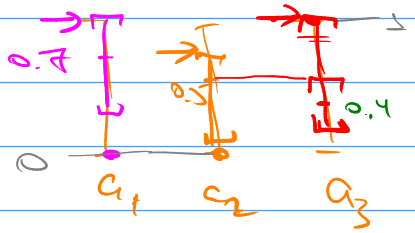
MDP Markov decision process

dynamics: $p(s', r \mid s, a) = P_r[S_{t+1}=s', R_{t+1}=r \mid S_t=s, A_t=a]$

$\forall s,a \quad \sum_{s',r} p(s', r \mid s, a) = 1$

strategy: $\pi(a|s) = P_r[A_t=a \mid S_t=s]$, $\forall s \sum \pi(a|s) = 1$

② Multiarmed bandits $|S| = 1$



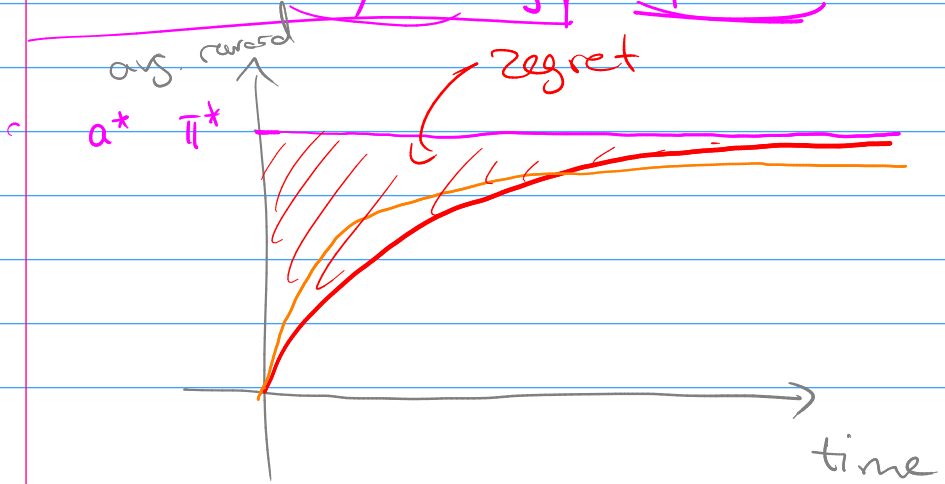
$\pi(a) = P_r[A_t=a]$

$p(r \mid a) = P_r[R_{t+1}=r \mid A_t=a]$

$/0/$ $/1$

$A = a_1, \ldots, a_k$

$D = \{(A_t, R_{t+1})\}$ exploration vs. exploitation



(3) Returns & Value functions

t: $R_{t+1}, R_{t+2}, R_{t+3}, \dots$

Episodic: $G_t = \sum_{\tau=t+1}^{T} R_\tau$ — return

discount factor

Continuous: (?) ($T = \infty$)

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

$\gamma \leq 1$

return →

$$G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$$

$T = \infty \Rightarrow \gamma < 1$

$T < \infty \Rightarrow \gamma = 1$

$$G_t = R_{t+1} + \gamma \cdot G_{t+1}$$

Value functions

$$V_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{k+t+1} \mid S_t = s\right]$$

$$Q_\pi(s,a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum \gamma^k R_{t+k+1} \mid \begin{matrix} S_t = s \\ A_t = a \end{matrix}\right]$$

$$V_\pi(s) = \mathbb{E}_{a \sim \pi}[Q_\pi(s,a)] = \sum_a \pi(a|s) Q_\pi(s,a)$$

$$Q_\pi(s,a) = \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a] =$$

$$= \mathbb{E}_\pi[R_{t+1} \mid S_t = s, A_t = a] + \gamma \mathbb{E}_\pi[G_{t+1} \mid S_t = s, A_t = a]$$

$r(s,a) = \mathbb{E}_p[R \mid s,a]$

$$= \mathbb{E}_{s' \sim p(s'|s,a)}[V_\pi(s') \mid s,a]$$

$$Q_\pi(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) V_\pi(s')$$

(4) Bellman equations

$$V_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s]$$

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

Ожид. мат. усл.

$$Q_\pi(s,a) = \sum_{s'} \sum_r p(s', r|s, a) \left[r + \gamma \sum_{a'} \pi(a'|s') Q_\pi(s', a') \right]$$

Ожид. мат. усл.

Bellman equations

π_* – опт. ур. $\forall \pi \; \forall s \quad V_*(s) \geq V_\pi(s)$

$$V_*(s) = \max_\pi V_\pi(s), \quad Q_*(s,a) = \max_\pi Q_\pi(s,a)$$

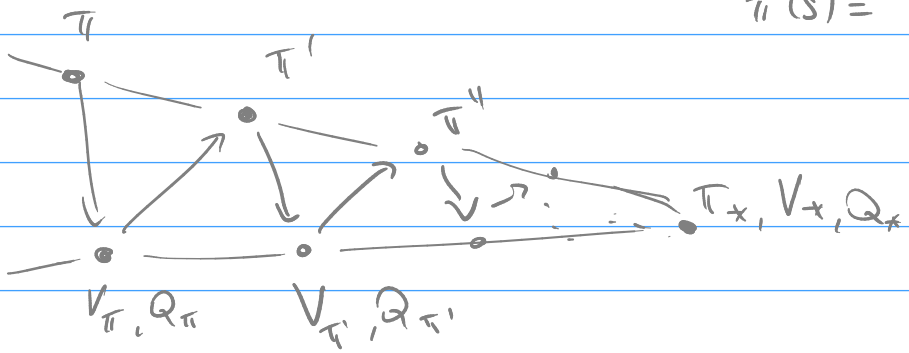
$$V_*(s) = \max_\pi V_\pi(s) = \max_\pi \left[\sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a) \left[r + \gamma V_\pi(s') \right] \right] =$$

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

Bellman eq

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

policy iteration

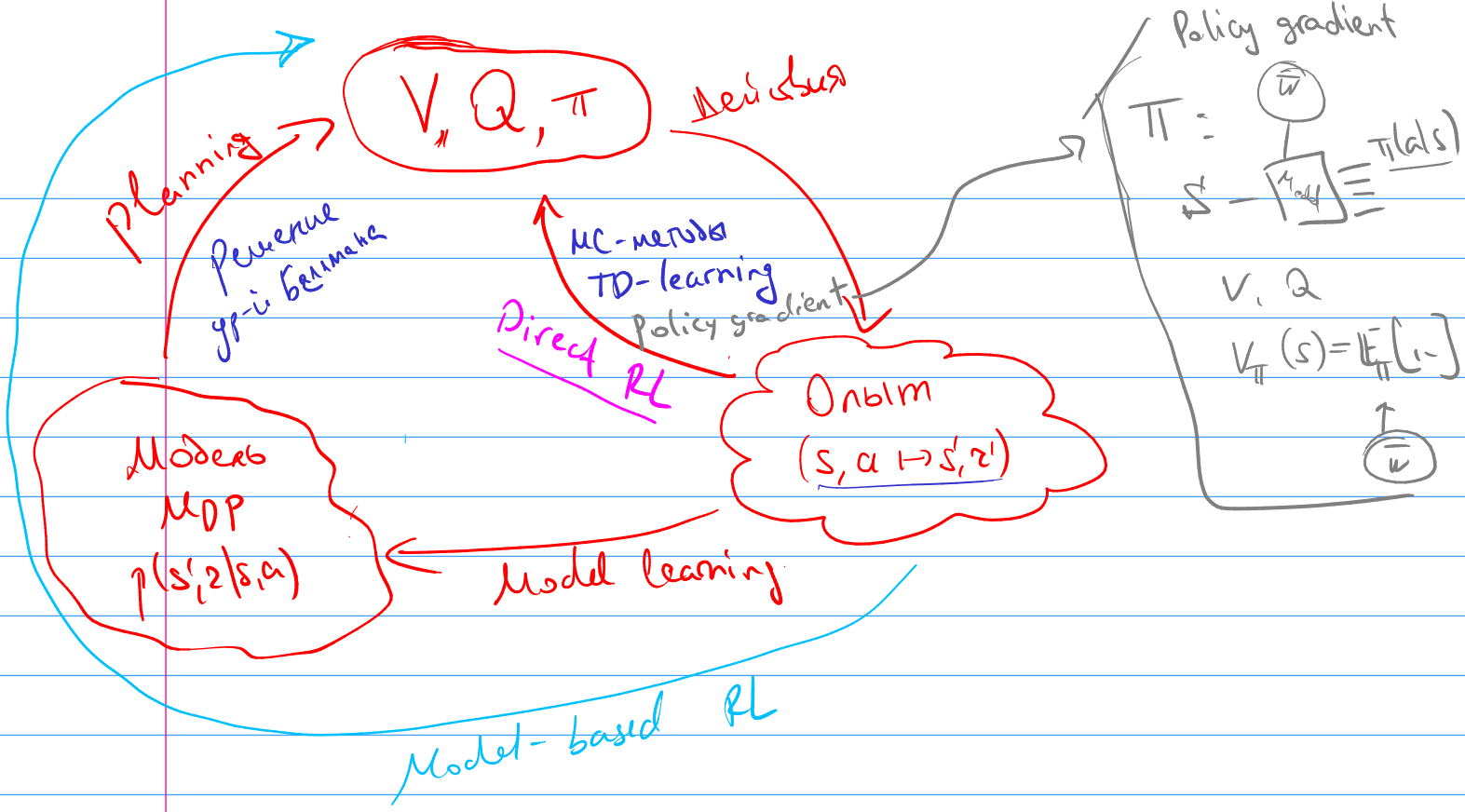


$$\pi'(s) = \arg\max_a Q_\pi(s,a)$$

policy improvement theorem

(5) Monte-Carlo

$p(r, s' \mid s, a) = ?$



MC: π, $V_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s]$

– loop:

MC evaluation — играем по π $S_0, A_0, R_1, S_1, A_1, \dots$

– считаем $G_t = \gamma G_{t+1} + R_{t+1}$

– для G_t в списках Returns(S_t)

– $V_\pi(S_t) := \text{Avg}(\text{Returns}(S_t))$

$\pi \to V_\pi$

MC control

$\pi \to \pi' \to \dots$

$Q(S_t, A_t) := \text{Avg}(\text{Returns}(S_t, A_t))$

$\pi(a|S_t) = \varepsilon\text{-Greedy}(Q)$

$= \begin{cases} 1-\varepsilon, & \text{argmax}_a Q \\ \varepsilon, & a \sim \text{Unif} \end{cases}$

on-policy control

играем по π

и оцениваем Q_π

importance sampling

off-policy control

играем по π

оцениваем Q_* $Q_{\pi'}$

⑥ TD-learning (temporal difference)

- эпизод S_0, A_0, R_1 - ——

$$V(S_t) \approx G_t = R_{t+1} + \gamma\left(R_{t+2} + \gamma^2 R_{t+3} + \dots\right)$$

$$G_{t+1} \approx V(S_{t+1})$$

TD(λ)

TD(0)

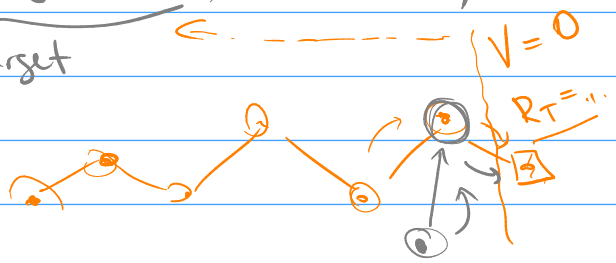
$$V(S_t) \longrightarrow R_{t+1} + \gamma \cdot V(S_{t+1})$$

1959

$$V(S_t) := V(S_t) + \alpha\left(R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right)$$

target

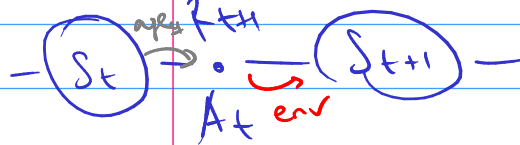
TD evaluation



TD control

ON-policy - SARSA

π: S_0, A_0, R_1, S_1, A_1 —— ~ ~ G



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

$\pi := \varepsilon$-Greedy(Q)

off-policy Q-learning

$(s, a) \mapsto (s', r)$

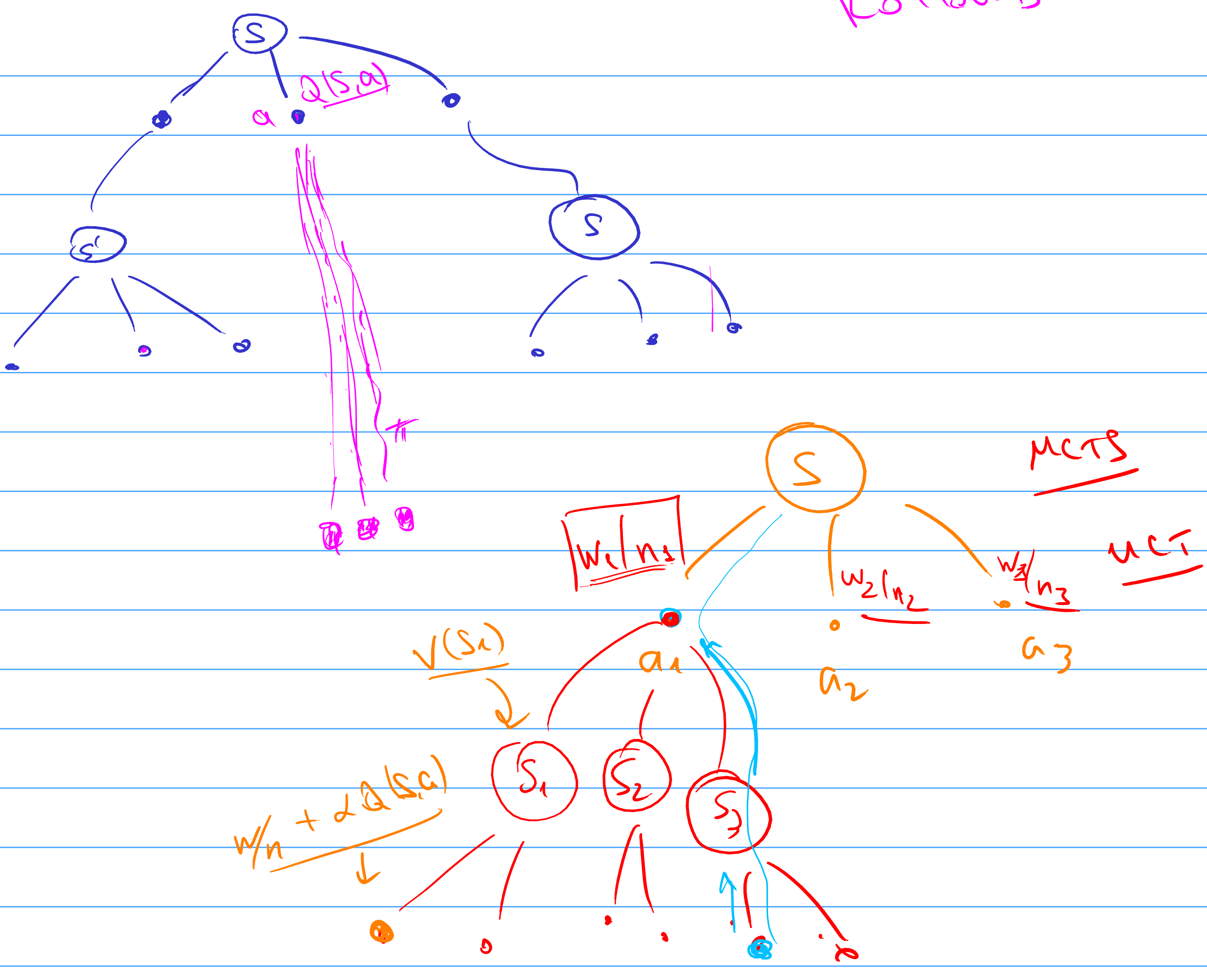
$$Q(s,a) = Q(s,a) + \alpha\left(r + \gamma \cdot \max_{a'} Q(s', a') - Q(s,a)\right)$$

$\rightarrow Q_*(s,a)$

$\pi := \varepsilon$-Greedy(Q)

MCTS – Monte-Carlo Tree Search

Rollouts

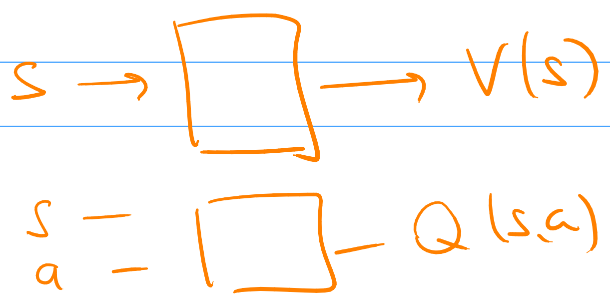


$$V(s) \approx \hat V(s, \bar w)$$

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

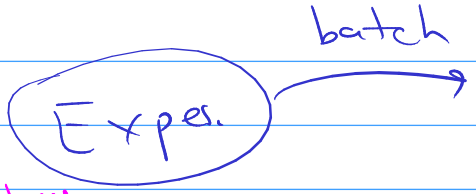
$$\bar w := \bar w + \alpha \left(R_{t+1} + \gamma \max_{a'} \hat Q(S_{t+1}, a', \bar w) - \hat Q(S_t, a_t, \bar w) \right) \nabla_{\bar w} \hat Q$$

S	V
s_1	–
s_2	–
⋮	–
$s_{10^{130}}$	–

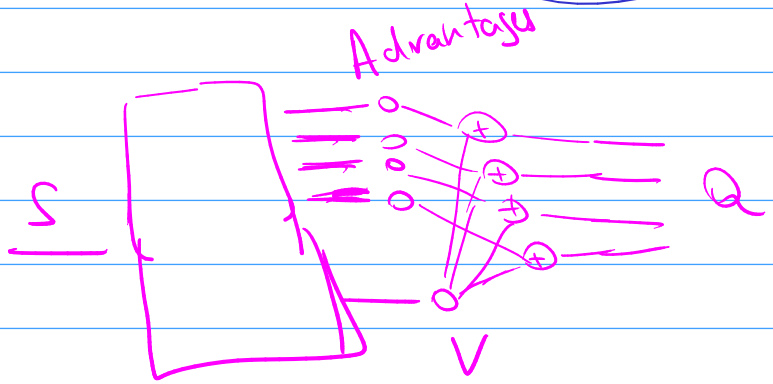


DQN deep Q-networks

1) Experience replay



2)



3) Double DQN