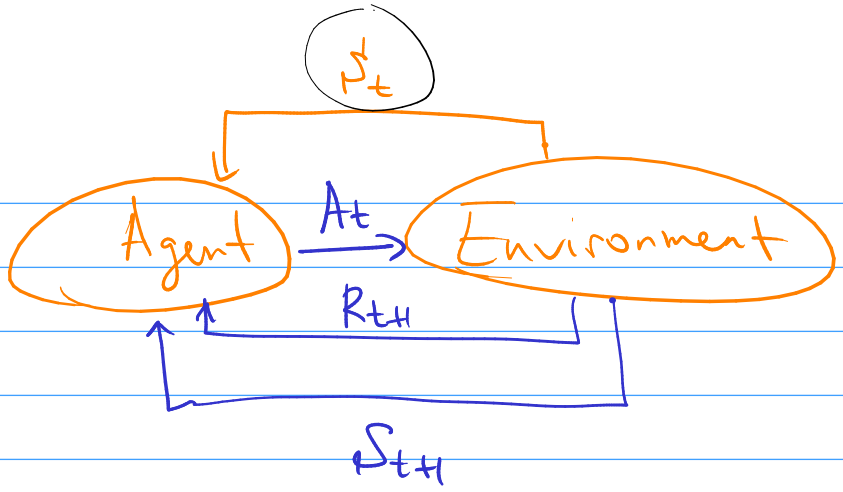


MDP – Markov Decision Process:

S, A Strategy: $\pi: S \to Prob(A)$

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

Dynamics: $p(r,s'|s,a) = Pr[R_{t+1} = r, S_{t+1} = s' | S_t = s, A_t = a]$

$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, S_3, \ldots$

π π π

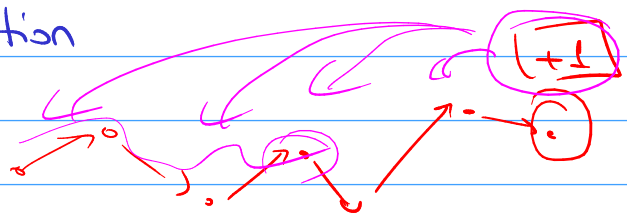
P P P

$$\sum_{t=1}^{\infty} R_t \cdot \gamma^{t-1} \to \max$$

$|\gamma = 1|$

1) Exploration vs. exploitation

2) Credit assignment

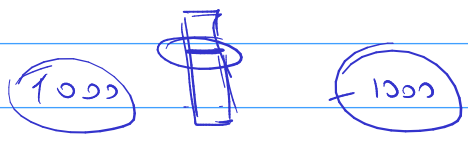


Multiarmed bandits $|S| = 1$

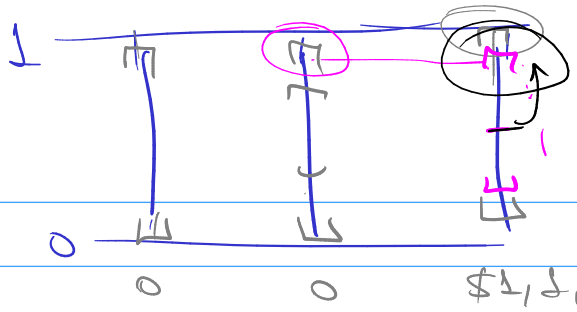
A

a_1 a_n

exploitation



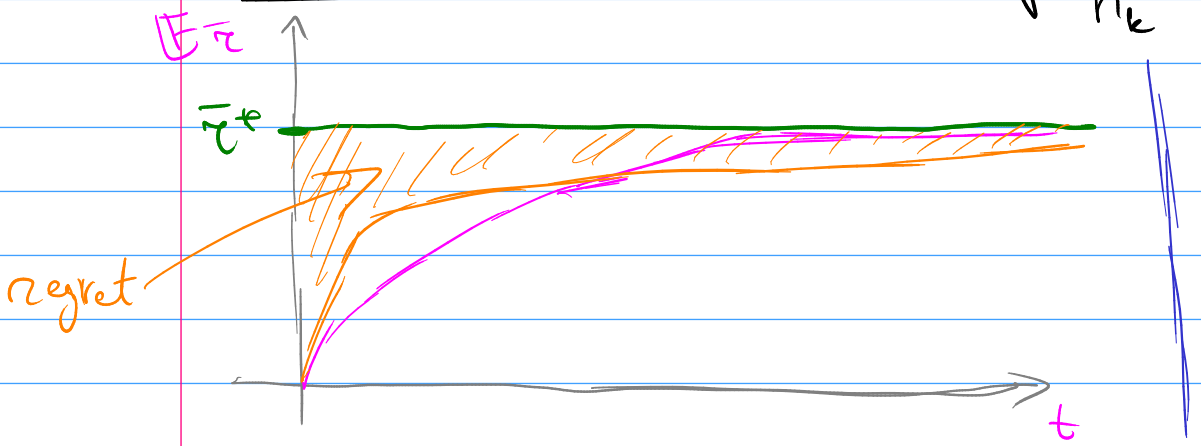
exploration



$Pri(a_k) = \bar{r}_k + (\dots)$

N, n_k, $\sum n_k = N$,

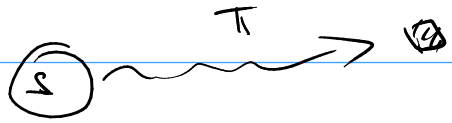
UCB1: $Pri(a_k) = \bar{r}_k + c\sqrt{\frac{\log N}{n_k}}$



Thm $c = \sqrt{2}$

$O(\log T)$

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



Value function

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

$$Q_\pi(s,a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=t+1}^{\infty} \gamma^{k-t-1} R_k \mid S_t = s, A_t = a\right]$$

$$V_\pi(s) = \sum_a \pi(a|s)\, Q(s,a)$$

G_{t+1}

$$Q_\pi(s,a) = \mathbb{E}_\pi[G_t \mid s,a] = \mathbb{E}_\pi\left[R_{t+1} + \gamma \left(\sum_{k=t+2}^{\infty} \gamma^{k-t-1} R_k\right) \mid s,a\right]$$

$$= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid s,a] = \underbrace{\mathbb{E}_\pi[R_{t+1} \mid s,a]}_{r(s,a)} + \gamma \underbrace{\mathbb{E}_\pi[G_{t+1} \mid s,a]}_{p(s'|s,a)} =$$

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

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

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

Bellman equations

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

$$V_*(s) = \max_\pi \mathbb{E}_\pi[G_t \mid S_t = s]$$

$$Q_*(s,a) = \max_\pi \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$$

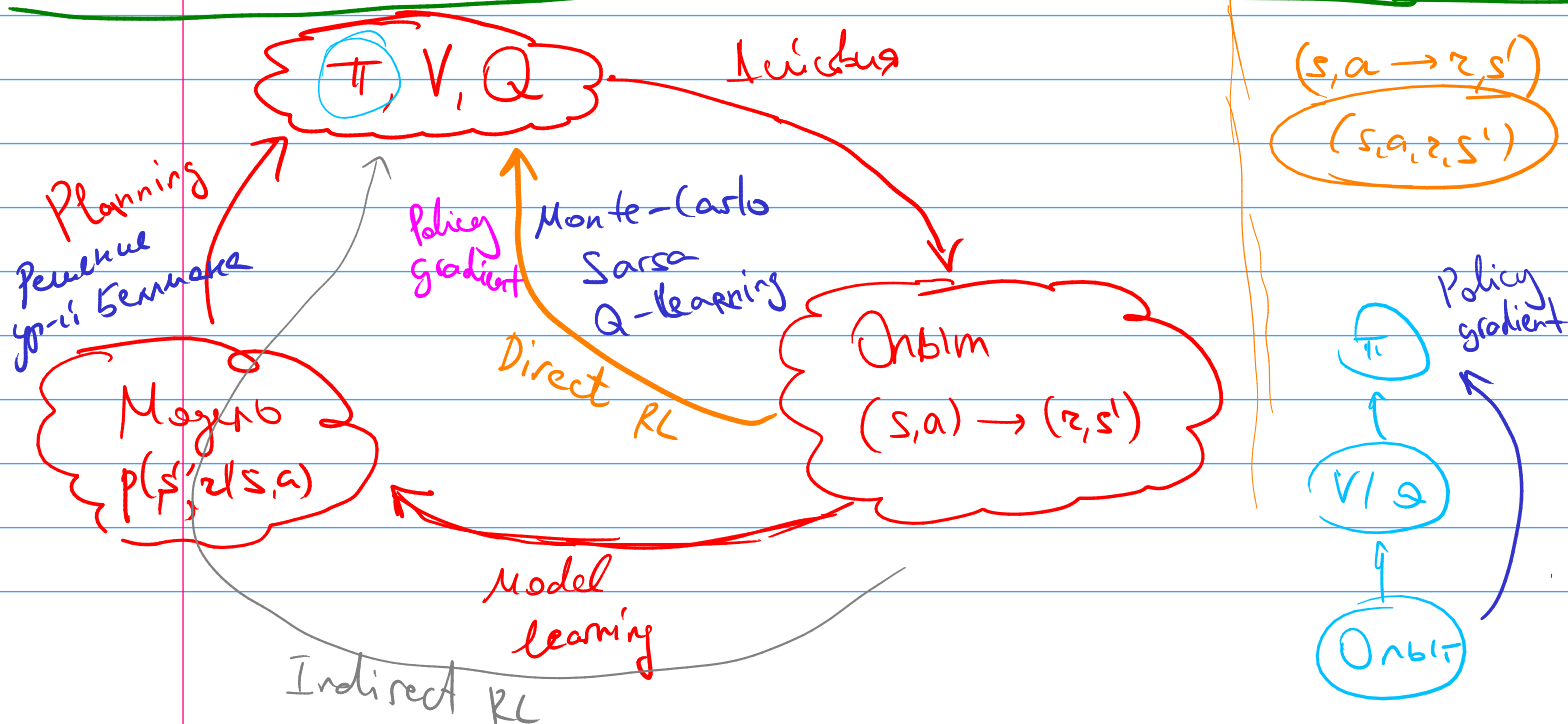
$$\pi_*(s) = \arg\max_a Q_*(s,a)$$

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

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

Bellman equation

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



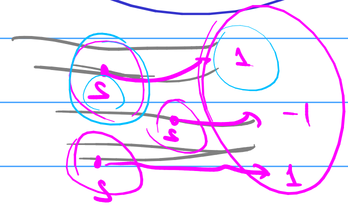
1) Monte-Carlo

Estimation $\quad V_\pi = \mathbb{E}_\pi [G_t | S_t = s] \approx \frac{1}{R} \sum_{r=1}^{R} G_t^{(r)}$

- нагенер. эпизод по π
- $\forall t \quad G_T = 0$

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

- $\forall t \quad Returns[S_t].append(G_t)$
- $V_\pi(S_t) := Avg(Returns(S_t))$



$Q_\pi(S_t, A_t)$

Control $\quad$ —"— $\quad \pi$ —

—"—

on-policy MC control

- $S_t, A_t: \quad Q(S_t, A_t) = Avg(Returns(S_t, A_t))$

$$\pi(a|S_t) = \begin{cases} 1-\varepsilon, & \arg\max_a Q(S_t, a) \\ \varepsilon, & Unif(a) \end{cases}$$

ε-soft, ε-greedy

Off-policy
- нагенер. по b (behaviour)
- обуч. $V_\pi(s)$, $Q_\pi(S, a)$

Importance sampling $\quad V_\pi(s) = \mathbb{E}_\pi [G_t | S_t = s] \overset{?}{\approx} \frac{1}{R} \sum_1^R G_t^{(r)}$

сэмплы не по π, а по b

$q(\bar{x})$

$$\mathbb{E}_{p(\bar{x})}[f(\bar{x})] = \int f(\bar{x}) p(\bar{x}) d\bar{x} =$$

$$= \int f(\bar{x}) \frac{p(\bar{x})}{q(\bar{x})} q(\bar{x}) d\bar{x} = \mathbb{E}_q \left[f(\bar{x}) \cdot \frac{p(\bar{x})}{q(\bar{x})} \right]$$

importance weights

---

2) TD-learning

temporal difference

$S_0, A_0, R_1, S_1, A_1, \ldots, S_t, A_t, R_{t+1}, S_{t+1}, \ldots$

$V_\pi(S_{t+1})$

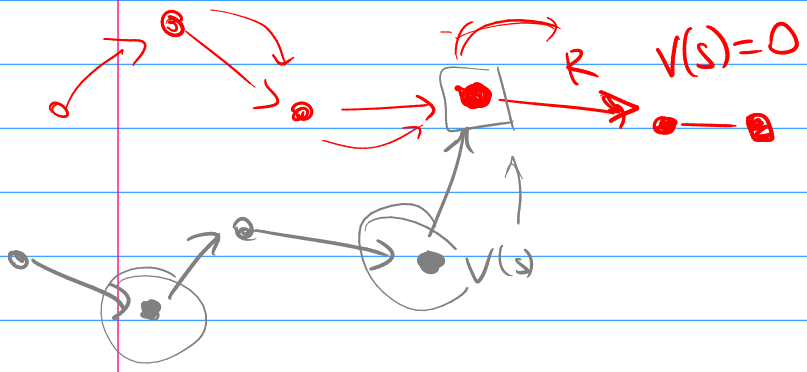
$$V_\pi(s) = \mathbb{E}_\pi [G_t | S_t = s] = \mathbb{E}_\pi [R_{t+1} + \gamma G_{t+1} | S_t = s]$$

$$V_\pi(S_t) \approx R_{t+1} + \gamma V_\pi(S_{t+1})$$

TD(λ)

TD(0): $V_\pi(S_t) := V_\pi(S_t) + \alpha\left[R_{t+1} + \gamma V_\pi(S_{t+1}) - V_\pi(S_t)\right]$

target



On-policy TD control — Sarsa

$(S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1})$

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

$\pi(S_t) = \varepsilon$-zach. ep. no Q

Off-policy TD-control Q-Learning

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

$\rightarrow Q_*(S_t, A_t)$

S — Model — $\hat{V}(s; \bar{w}) \approx V(s)$

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

Semi-gradient TD

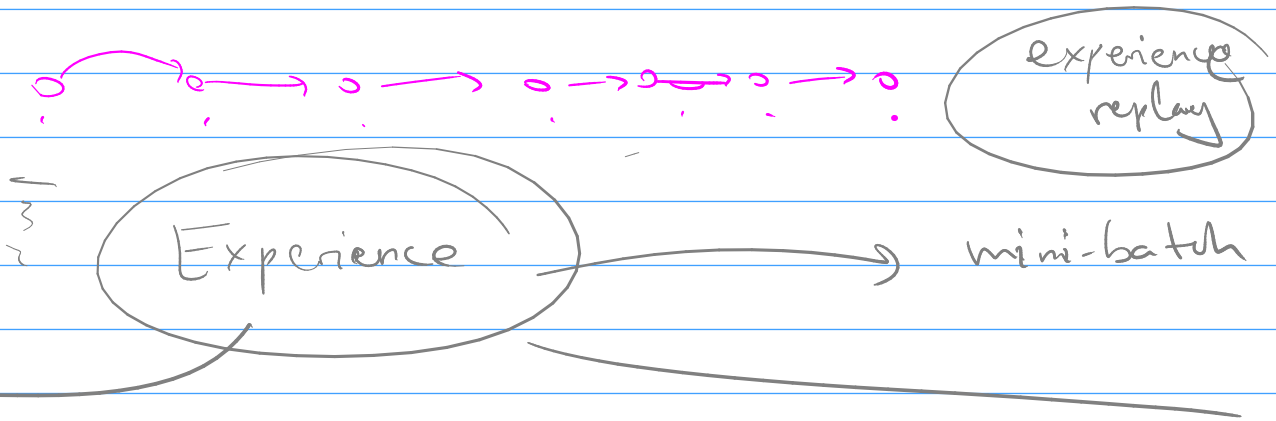
$$\bar{w} \mathrel{+}= \alpha\left(R_{t+1} + \gamma \hat{V}(S_{t+1}, \bar{w}) - \hat{V}(S_t, \bar{w})\right) \nabla_{\bar{w}} \hat{V}(S_t, \bar{w})$$

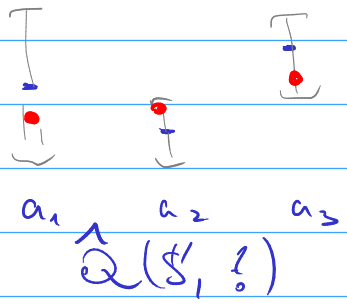
Policy gradient

$$\pi(a|s, \bar{\theta}) \quad s \rightarrow [\pi] \equiv \pi(a|s)$$

$$J(\bar{\theta}) = V_{\pi_{\bar{\theta}}}(S_0)$$

$$\bar{\theta}_{t+1} = \bar{\theta} + \alpha \cdot \nabla_{\bar{\theta}} J(\bar{\theta})$$





$$s \rightarrow [V] \rightarrow \hat{V}(s)$$

$$s, a \rightarrow [Q] \rightarrow \hat{Q}(s,a)$$

Dueling networks

$$s \rightarrow [Q] \equiv \hat{Q}(s,a)$$

advantage

$$s \rightarrow [\quad] \equiv \hat{A}(s,a), \hat{V}(s) \Rightarrow (+) \equiv \hat{Q}(s,a) = \hat{V}(s) + \hat{A}(s,a)$$