$\pi(a|s)$

$S_0, A_0, R_1, S_1, A_1, \ldots$

$p(r, s'|s, a)$

$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=t+1}^{\infty} \gamma^{k-t-1} R_k$

trend following

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

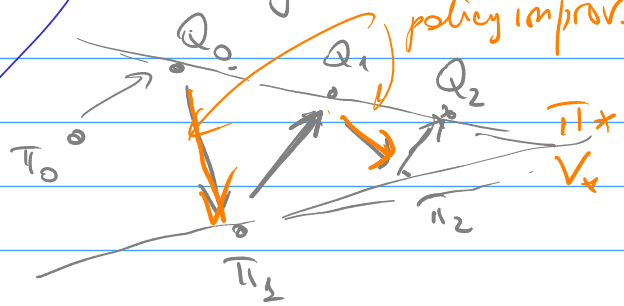
$Q_\pi(s,a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a]$

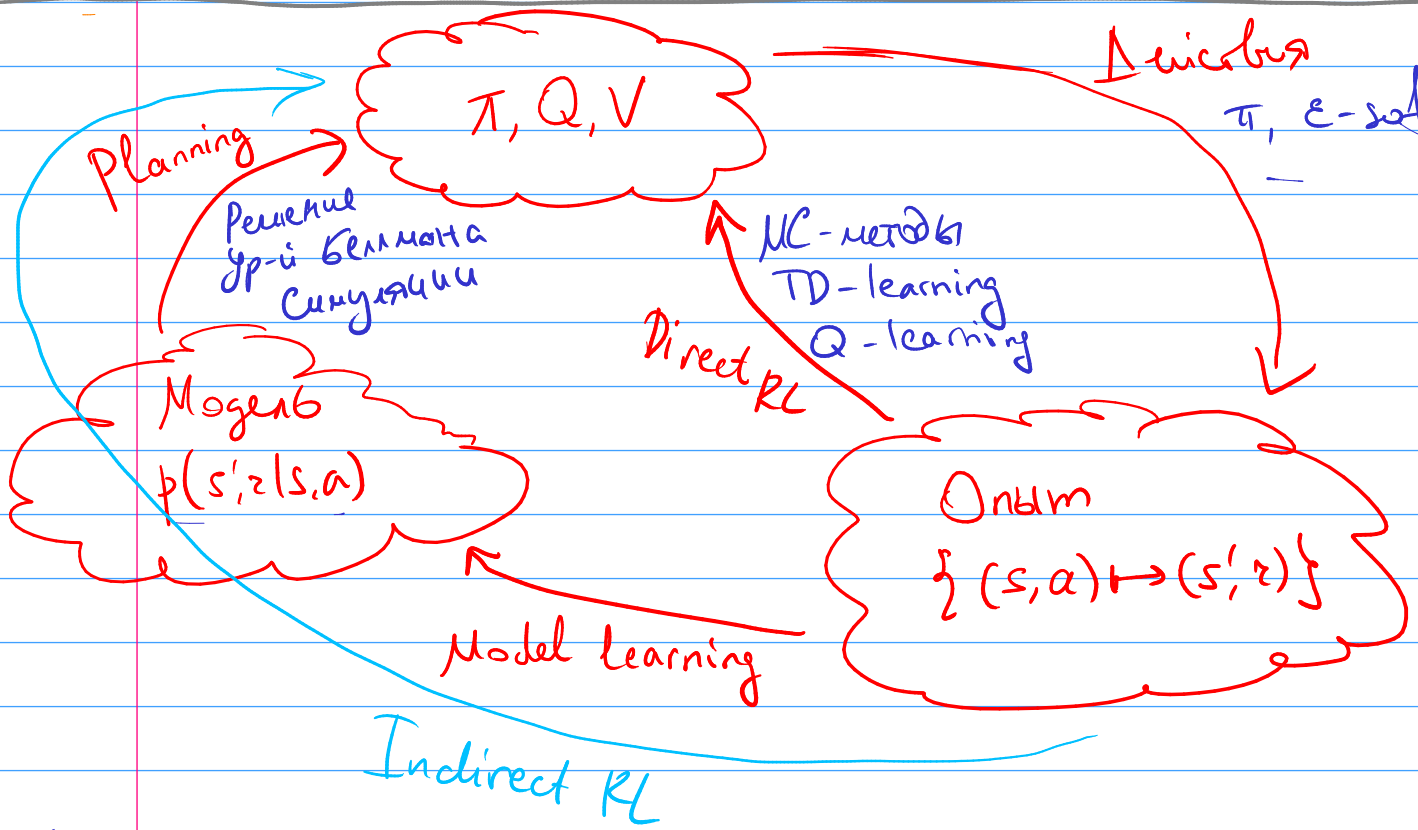
Bellman equations

$V_*(s) = \max_\pi V_\pi(s)$

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

Policy iteration





Monte Carlo

① MC estimation

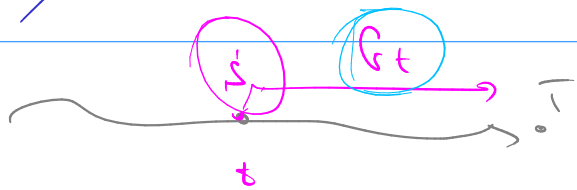
$V_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s] \approx \frac{1}{N} \sum_{n=1}^{N} G_t^{(n)}$

π: $S_0, A_0, R_1, S_1, A_1, \ldots$

MC estimation

- init π, $V_\pi(s)$
- loop:
  - порожд. эпизод по стр. π:

$S_0, A_0, R_1, S_1, A_1 \ldots = \ldots, S_T$



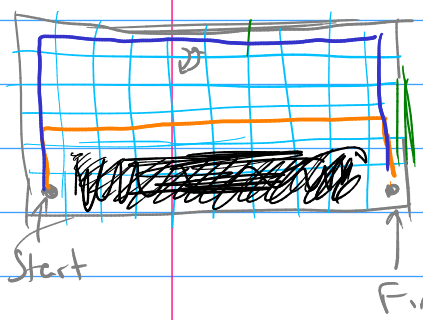
- считаю $G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ $\forall t$: $G_T = 0$, $G_{T-1} = R_T$, $G_{T-2} = \gamma R_T + R_{T-1}$

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

- $\forall t$ добавляем G_t в Returns$[S_t]$ / Returns$[S_t, A_t]$
- $V_\pi(S_t) = Avg($Returns$[S_t])$

② On-policy MC control

- init π - —
- loop
  - — — // — — // — — G_t, Returns$[S_t, A_t]$
  - $Q_\pi(S_t, A_t) = Avg($Returns$[S_t, A_t])$
  - $\forall t$ $\pi(a|S_t) = \begin{cases} 1-\varepsilon, & a = \arg\max_a Q(S_t, a) \\ \varepsilon, & a \sim Unif \end{cases}$



$\pi(a|S_t) = \arg\max_a Q(S_t, a)$

$b(a|S_t) = \begin{cases} 1-\varepsilon \\ \varepsilon \end{cases}$ — behaviour

③ Off-policy MC control :

- порождать по b
- обучать не Q_b, а Q_π

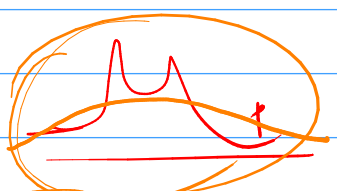
Importance sampling

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

$Q_b(S, a) = \mathbb{E}_b[G_t \mid - \; -]$

$\mathbb{E}_{p(\bar{x})}[f(\bar{x})]$, $\bar{x}_n \sim q(\bar{x})$ $\int q(\bar{x}) \ldots d\bar{x}$

$= \int p(\bar{x}) f(\bar{x}) d\bar{x} = \int \frac{p(\bar{x})}{q(\bar{x})} f(\bar{x}) q(\bar{x}) d\bar{x} =$



$= \mathbb{E}_{q(\bar{x})}\left[f(\bar{x}) \frac{p(\bar{x})}{q(\bar{x})}\right] \approx \frac{1}{N} \sum_n \frac{p(\bar{x}_n)}{q(\bar{x}_n)} f(\bar{x}_n)$

Если $q(\bar{x}) = 0$, то и $p(\bar{x}) = 0$

$S_t, A_t : R_{t+1}, S_{t+1}, A_{t+1}, R_{t+2}, S_{t+2}, \dots, S_T$ ← Traj

$$Pr[Traj|\pi, S_t, A_t] = p(S_{t+1}, R_{t+1}|S_t, A_t)\pi(A_{t+1}|S_{t+1})\, p(S_{t+2}, R_{t+2}|S_{t+1}, A_{t+1})\pi(A_{t+2}|S_{t+2})\dots$$

$$= \prod_{k=t+1}^{T} p(S_k, R_k|S_{k-1}, A_{k-1})\,\pi(A_k|S_k)$$

$$Pr[\dots|b\dots] = \prod p(S_k, R_k|S_{k-1}, A_{k-1})\, b(A_k|S_k)$$

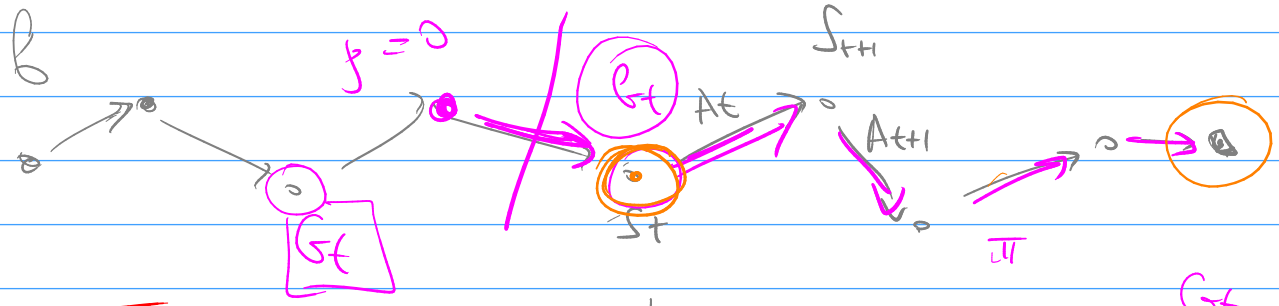
$$\rho_{t:T-1} = \frac{Pr[Traj|\pi]}{Pr[Traj|b]} = \prod_k \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$$

G_t : $G_T = 0$, $G_t = \gamma G_{t+1} + R_{t+1}$

$\rho_T = 1$, $\rho_t = \rho_{t+1} \cdot \frac{\pi(A_{t+1}|S_{t+1})}{b(A_{t+1}|S_{t+1})}$

b: $S_0, \dots, R_T, S_T$

$$Q_\pi = E_\pi[\dots] = E_b[\rho\dots] \approx \frac{1}{N}\sum_1^N G_t \cdot \rho_t$$



V, Q

$E_\pi[G_t|\dots]$

t

G_t

$V(s) \approx$ samples

$V = E_\pi[\dots] \approx$ samples

TD Learning

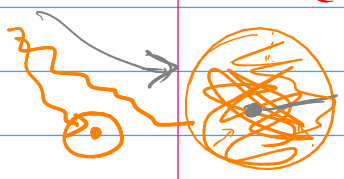
$$V(S_t) = \mathbb{E}_\pi [G_t | \ldots] =$$

$$= \mathbb{E}_\pi [\gamma G_{t+1} + R_{t+1} | \ldots]$$

$$= \mathbb{E}_\pi [R_{t+1} | \ldots] + \gamma \cdot \mathbb{E}_\pi [G_{t+1} | S_t]$$

Sample step $(t+1)$ $\approx R_{t+1} + \gamma \cdot \mathbb{E}_\pi [G_{t+1} | S_{t+1}] =$

$= R_{t+1} + \gamma V(S_{t+1})$

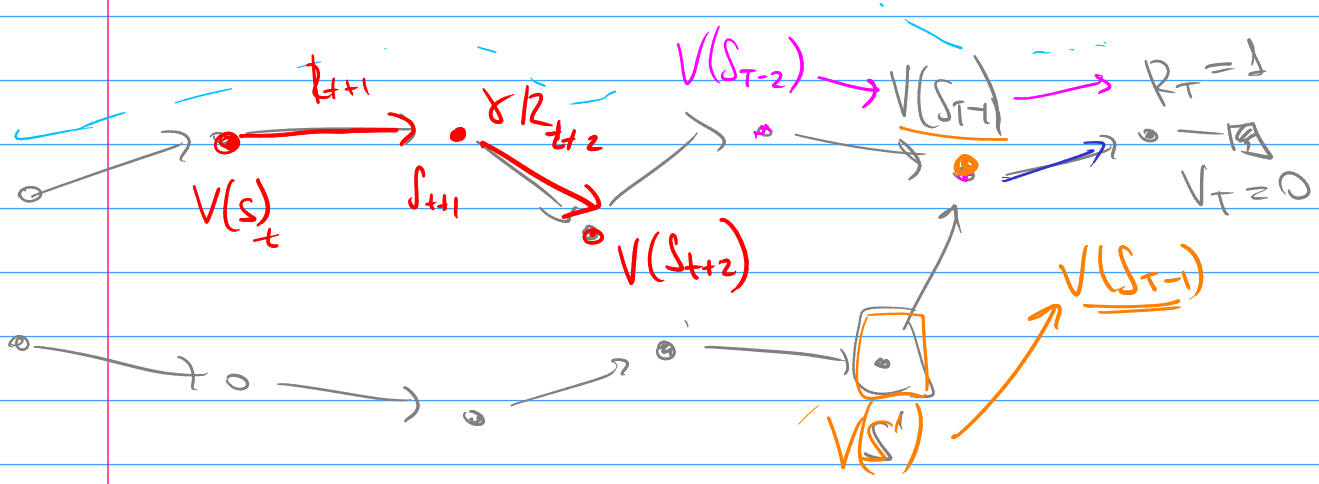


TD estimation π

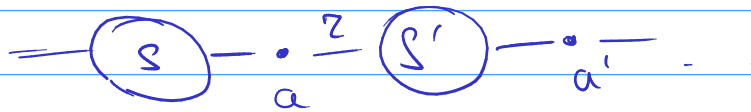
$$V_\pi(S_t) := V_\pi(S_t) + \alpha \left[\overbrace{R_{t+1} + \gamma V_\pi(S_{t+1})}^{\text{target}} - V_\pi(S_t) \right]$$

$$V(S_t) \approx G_t = R_{t+1} + \gamma (R_{t+2} + \ldots)$$

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



On-policy TD control — (S, A, R, S', A')



- init
- loop

Sarsa

Q (S, A, R, S', A') — $Q(S,A) := Q(S,A) + \alpha \left[R + \gamma Q(S', A') - Q(S,A) \right]$

TD-target

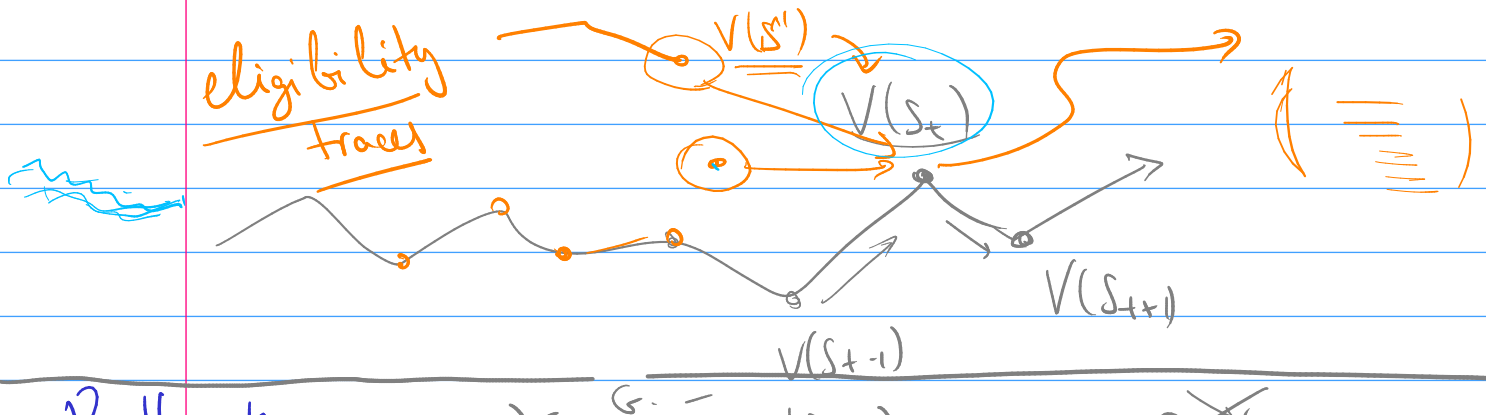
— π — ε-жадн. сл. по Q

Off-policy TD-control — Q-learning (1989)

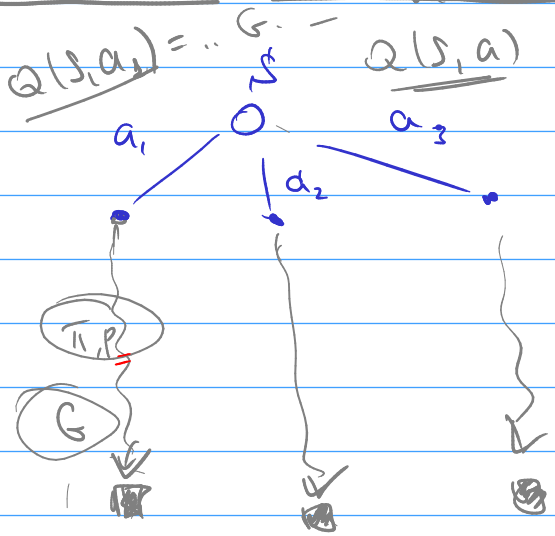
Q*

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

eligibility traces

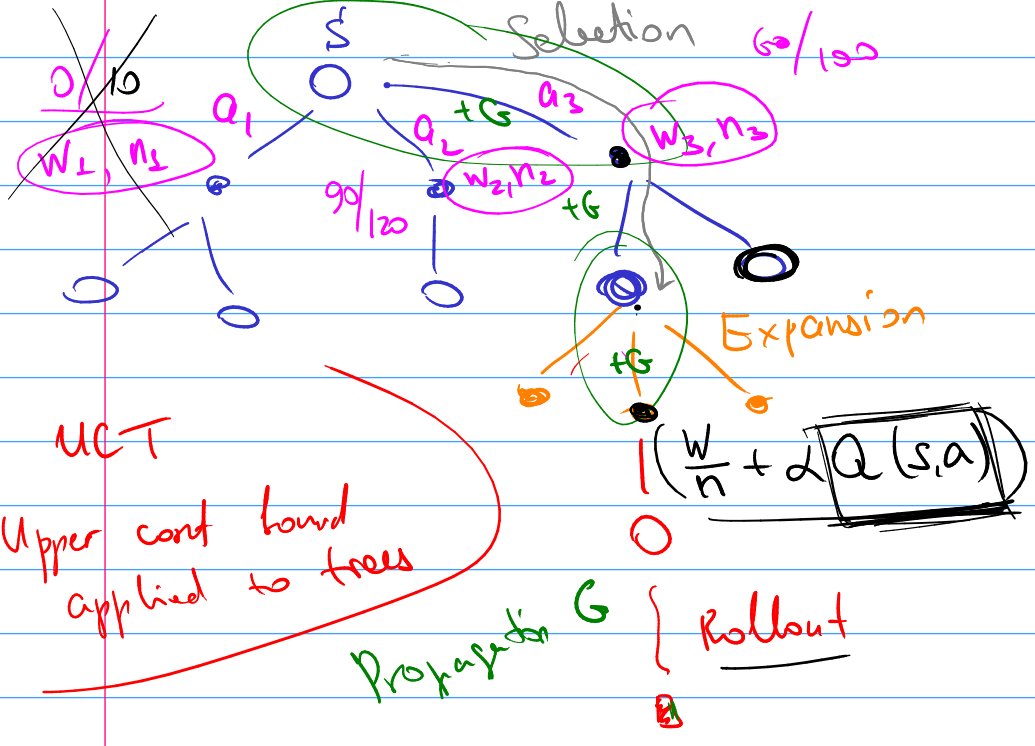


Rollouts



Q*

Decision-time planning



MCTS

Monte-Carlo Tree Search

800 rollouts

MCT

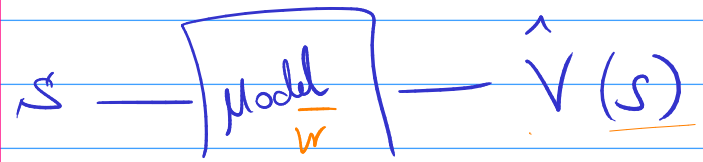
Upper conf bound applied to trees

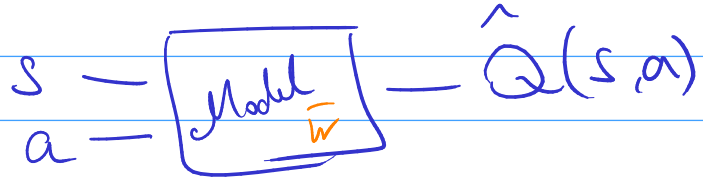
Propagation G } Rollout

D

$s = \bar{x}$

$\hat{V}(s) = \sigma(\bar{w}^T \bar{x})$





---

s: $V(s)$ $\qquad E = \frac{1}{2}\left(V(s) - \hat{V}(s)\right)^2 \xrightarrow[\bar{w}]{} \min$

$$\nabla_{\bar{w}} E = -\left(V(s) - \hat{V}(s)\right) \cdot \nabla_{\bar{w}} \hat{V}$$

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

Gradient MC:

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

Semi-gradient TD(0):

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

Semi-gradient Q-learning

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