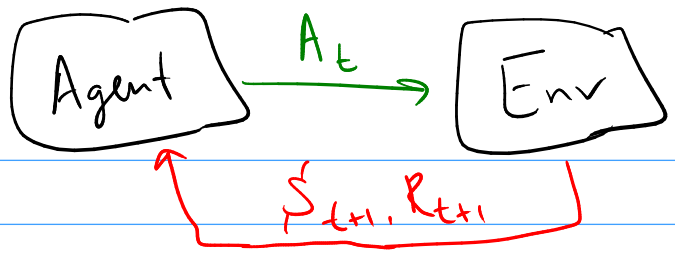


$$G_t = \sum_{k=t+1}^{\infty} \gamma^{k-t-1} R_k$$

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

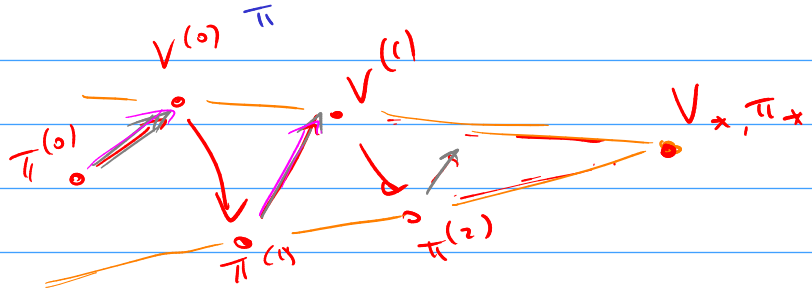
$\pi : S \to Prob[A(S)]$, $\forall s$ $\pi(a|s) = Pr[A_t = a \mid S_t = s]$

$$V_\pi(s) = E_\pi[G_t \mid S_t = s]$$

$$\overline{V_\pi(s)} = f(\overline{V_\pi(s)})$$

$$Q_\pi(s,a) = E_\pi[G_t \mid S_t = s, A_t = a]$$

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



MC estimation: $V_\pi(S) = Avg[Returns(S)]$

$Q_\pi(S,a) = Avg[Ret(S,a)]$

on-policy MC control:

$Q(s_t, a_t) := Avg(Ret(S,a))$

ε-soft $\forall a \ \pi(a|s_t) = \begin{cases} 1 - \varepsilon + \varepsilon/|A(s_t)|, & a = a^* \\ \varepsilon/|A(s_t)|, & \forall a \neq a^* \end{cases}$

Off-policy MC control:

- moving average no sp. $b(a|s)$ [behaviour]
- obtained $Q_\pi(s,a)$, a не $Q_b(s,a)$

Importance sampling $E_p[f] = \int p(\bar{x}) f(\bar{x}) d\bar{x} =$ (importance weights)

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

$$V_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_b\left[G_t \cdot \frac{Pr[Traj|\pi]}{Pr[Traj|b]} \mid S_t = s\right] =$$

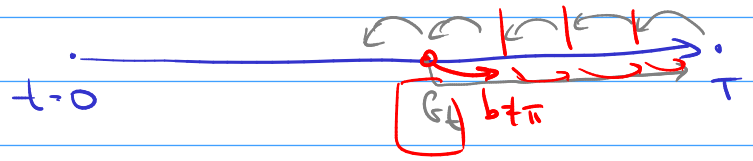
$S_t, A_t, S_{t+1}, R_{t+1}, A_{t+1}, S_{t+2}, \ldots$

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

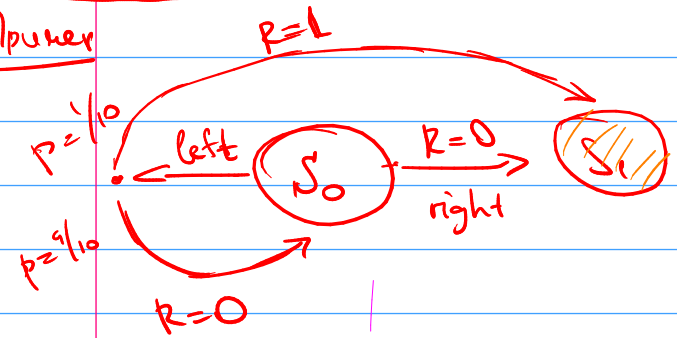
$$= \mathbb{E}_b\left[G_t \cdot \frac{\prod_{k=t}^{T} \pi(A_k|S_k)}{\prod_{k=t}^{T} b(A_k|S_k)} \mid S_t = s\right]$$

Условие: если $\pi(a|s) > 0$, то $b(a|s) > 0$

$V_\pi(s) = Avg(Ret(s))$ ← $G_t \cdot \gamma_{t..T}$



Пример



$\pi(left|S_0) = 1$

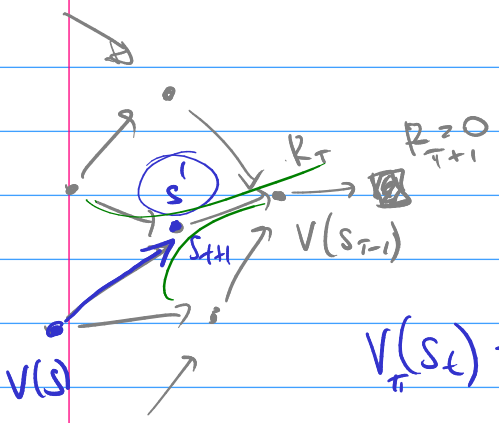
$b(left|S_0) = b(right|S_0) = \frac{1}{2}$

$$\mathbb{E}_b\left[\left(G \cdot \frac{\prod_k \pi(a_k|S_k)}{\prod_k b(a_k|S_k)}\right)^2\right] = \left(\frac{1}{2}\cdot\frac{1}{10}\right)\cdot\left(\frac{1}{1/2}\right)^2 +$$

$$+ \left(\frac{1}{2}\cdot\frac{9}{10}\cdot\frac{1}{2}\cdot\frac{1}{10}\right)\left(\frac{1}{1/2\cdot 1/2}\right)^2 + \ldots =$$

$$= \frac{1}{10}\sum_{k=0}^{\infty}\left(\frac{9}{10}\right)^k\left(\frac{1}{2}\right)^{k+1}\cdot 2^{2k+1} = \frac{1}{5}\cdot\sum_{k=0}^{\infty}\left(\frac{9}{5}\right)^k$$

TD-learning (temporal difference)



TD(0)

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

$$\text{New} = \text{Old} + \underbrace{\text{Learning Rate}}\left[\text{Target} - \text{Old}\right]$$

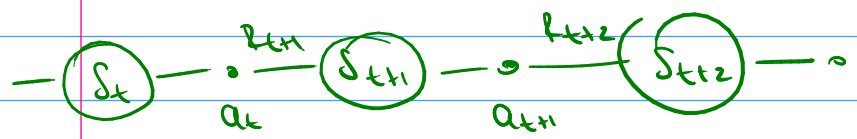
$$V_\pi(S_t) = \mathbb{E}_\pi[G_t|S_t] \approx \text{Avg}(G_t)$$

$$V_\pi(S_t) = V_\pi(S_t) + \frac{1}{n+1}\left(G_t - V_\pi(S_t)\right)$$

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots =$$

$$= R_{t+1} + \gamma \cdot G_{t+1} \approx R_{t+1} + \gamma \cdot V_\pi(S_{t+1})$$

On-policy TD-control — Sarsa



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

- init
- loop по каждому $t = 0, \dots, T$:
  - (s, a, r, s')
  - выберем a' в s' по $\pi(s') = \varepsilon$-жадн. стр. по $Q(s,a)$
  - $Q(s,a) := Q(s,a) + \alpha\left(\underbrace{r + \gamma \cdot Q(s',a')}_{\text{target}} - Q(s,a)\right)$

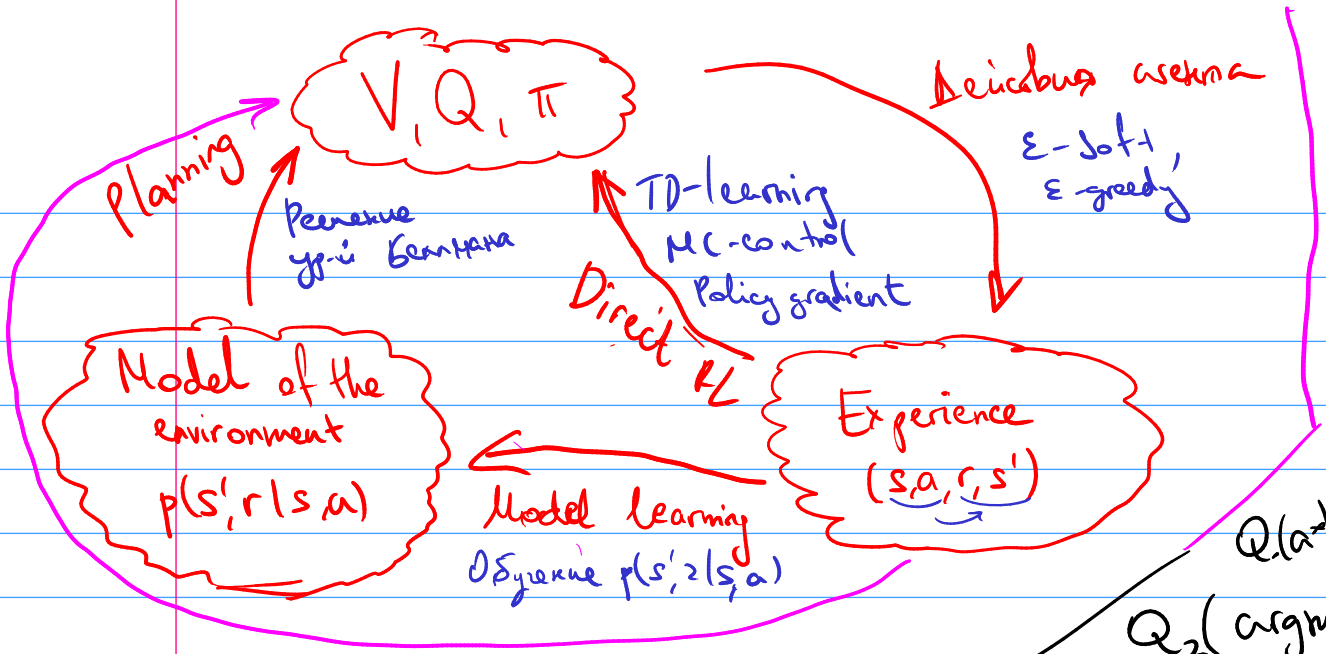
Off-policy TD-control — Q-learning — 1989

$$Q_2(s,a) := Q_2(s,a) + \alpha\left(r + \gamma \cdot \max_{a'} Q_1(s',a') - Q_2(s,a)\right)$$

Q_*

$\pi_*(s) = \text{argmax}_a\, Q_*(s,a)$

Проблема: winner's curse, maximization bias



Double Q-learning

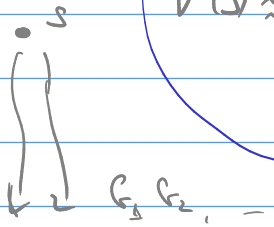
Q_1, Q_2

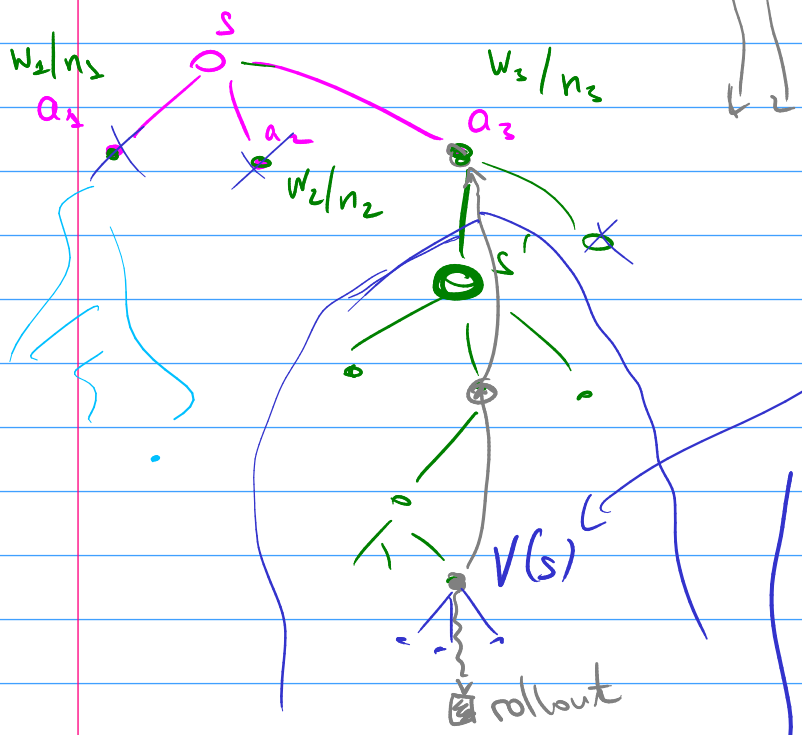
$a^* = \arg\max_a Q_1$

$Q(a^*) = Q_2$

$Q_2(\arg\max_a Q_1(a))$

$V'(S) \approx \alpha V(s) + (1-\alpha) \cdot \text{Avg}(\text{Rollouts})$

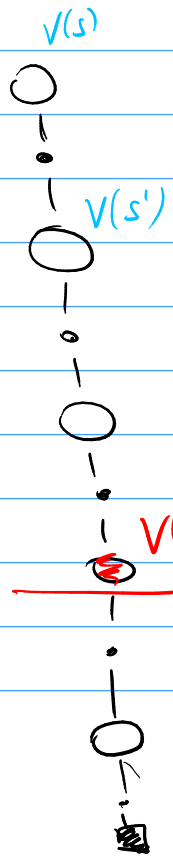




$G_t = R_{t+1} + \gamma G_{t+1} \approx V(S_{t+1})$

$= R_{t+1} + \gamma R_{t+2} + \gamma^2 G_{t+2} \approx V(S_{t+2})$

$\sum_{i=1}^{k-1} \gamma^{i-1} R_{t+i} + \gamma^k G_{t+k} \approx V(S_{t+k})$



Approximate RL

$$Q(s,a) \approx \hat{Q}(s,a,\bar{\theta})$$

$$V(s) \approx \hat{V}(s,\bar{\theta})$$

$$L(\bar{\theta}) = \sum_s \mu(s) \left(V(s) - \hat{V}(s,\bar{\theta}) \right)^2 \xrightarrow[\bar{\theta}]{} \min$$

SGD $\quad s_t$

$$= \nabla_{\bar{\theta}} L(\bar{\theta})$$

$$\bar{\theta} := \bar{\theta} + \alpha \left(V(s_t) - \hat{V}(s_t,\bar{\theta}) \right) \cdot \nabla_{\bar{\theta}} \hat{V}(s_t,\bar{\theta})$$

Gradient MC: $\quad V(s_t) \approx G_t$

$$\bar{\theta} := \bar{\theta} + \alpha \left(G_t - \hat{V}(s_t,\bar{\theta}) \right) \cdot \nabla_{\bar{\theta}} \hat{V}(s_t,\bar{\theta})$$

Semi-gradient TD(0):

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

— " — Q-Learning

— " — $\quad R_{t+1} + \gamma \max_a \hat{Q}(S_{t+1},a,\bar{\theta})$ — " —