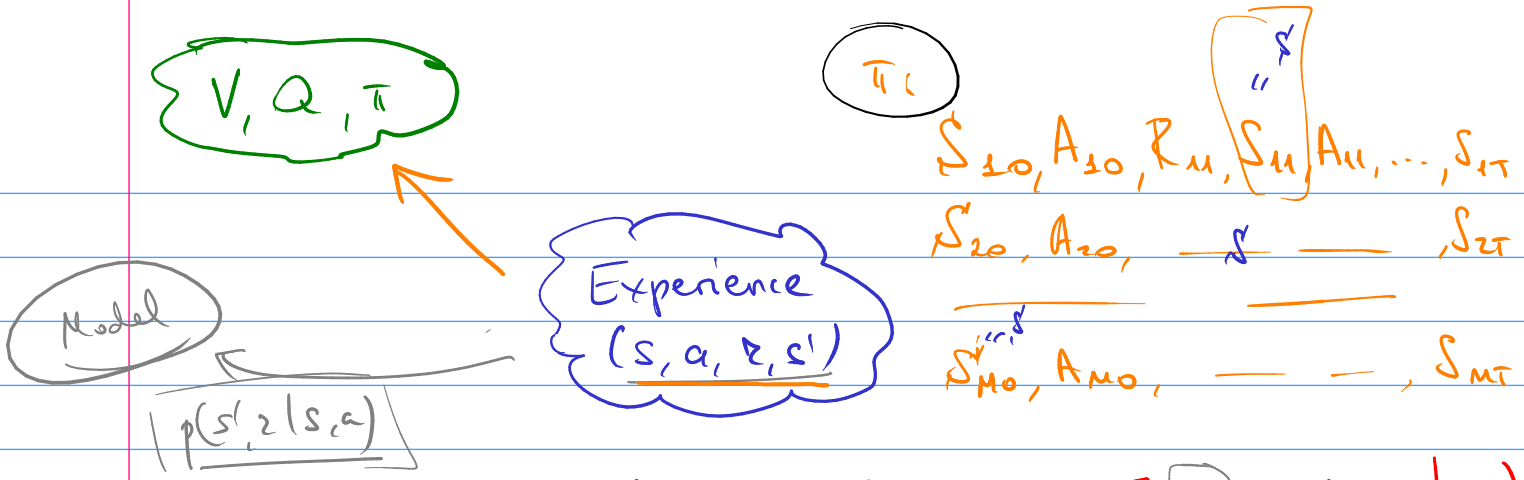



$$V_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] \approx \text{Avg}(G_t \mid S_t = s)$$

## Monte Carlo estimation:

- init $\pi$, $V(s)$
- repeat:
  - порождать $S_0, A_0, R_1, S_1, A_1, \ldots, S_t, \ldots, S_T$ по стратегии $\pi$
  - $G := 0$
  - for $t = T-1, T-2, \ldots, 0$:
    - $G := \gamma G + R_{t+1}$
    - если надо, то добавить $G$ в $\text{Returns}(S_t)$ и $V(S_t) := \text{Avg}(\text{Returns}(S_t))$

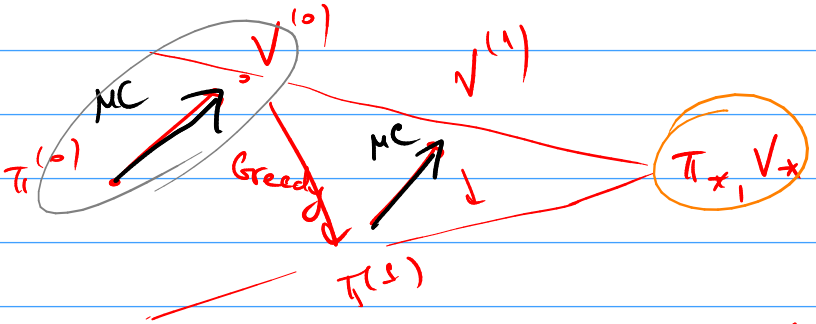
first-visit MC: $G_t$

every-visit MC: $G_t$

Exploring starts

- ——"—— $G$ в $\text{Returns}(S_t, A_t)$ и $Q(S_t, A_t) := \text{Avg}(\text{Returns}(S_t, A_t))$

$\pi \longrightarrow V_\pi, Q_\pi$




$$\pi'(s) = \underset{a}{\text{argmax}}\, Q_\pi(s, a)$$

**$\varepsilon$-soft** — $\forall a \; p(a|s) \geq \frac{\varepsilon}{|A|}$

Policy improv. thm for $\varepsilon$-soft strategies

# Monte Carlo control

— — — —

— repeat:

— — — — episode $S_0, A_0, \dots$

— $\forall t = T-1, \dots, 0$

— gos. G в Returns$(S_t, A_t)$

— $Q(S_t, A_t) := \text{Avg}(\text{ret}\_\_)$ expl. starts

→ оценка $q_\pi$ — $\pi(S_t) := \text{argmax}_a Q(S_t, a)$

→ мягкая стр-ра — $\pi'(S_t) := \begin{cases} \text{argmax}_a Q(S_t, a) & \text{с вер. } 1-\varepsilon \\ \text{uniform} & \text{с вер. } \varepsilon \end{cases}$

---

## Importance sampling

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

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

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

$$p(\text{Traj} \mid \pi) = p(A_t, \dots, S_T \mid \pi) = \pi(A_t \mid S_t)\, p(S_{t+1} \mid S_t, A_t) \cdot \pi(A_{t+1} \mid S_{t+1}) \cdot \dots \cdot p(S_T \mid S_{T-1}, A_{T-1})$$

$$\mathbb{E}_\pi[G_t] = \mathbb{E}_{\pi'}\left[G_t \cdot \frac{p(\text{Traj} \mid \pi)}{p(\text{Traj} \mid \pi')}\right] = \mathbb{E}_{\pi'}\left[G_t \cdot \prod_{k=t}^{T-1} \frac{\pi(A_k \mid S_k)}{\pi'(A_k \mid S_k)}\right]$$

$$\frac{p(\text{Traj} \mid \pi)}{p(\text{Traj} \mid \pi')} = \frac{\pi(A_t \mid S_t)\, p(S_{t+1} \mid S_t, A_t) \cdot \dots \cdot \pi(A_{T-1} \mid S_{T-1})\, p(S_T \mid S_{T-1}, A_{T-1})}{\pi'(A_t \mid S_t)\, p(S_{t+1} \mid S_t, A_t) \cdot \dots \cdot \pi'(A_{T-1} \mid S_{T-1})\, p(S_T \mid S_{T-1}, A_{T-1})}$$

$= \rho_{t:T-1}^{\pi, \pi'}$

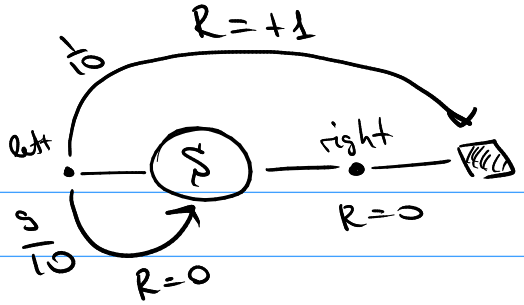
off-policy estimation

on-policy

$$V_\pi(s) = \mathbb{E}_{\pi'}\left[G_t \cdot \rho_{t:T-1}^{\pi, \pi'}\right]$$

$$V_\pi(s) = \frac{1}{N}\sum_{t=1}^{N} \rho_{t:T-1}^{\pi,\pi'} G_t$$

$$V_\pi(s) = \frac{\sum_{t=1}^{N} \rho_{t:T-1}^{\pi,\pi'} G_t}{\sum_{t=1}^{N} \rho_{t:T-1}^{\pi,\pi'}}$$




$$\pi(left|s) = 1$$

$$\pi'(left|s) = 1/2$$

$$V_\pi(s) = \mathbb{E}_{\pi'}\left[G \cdot \frac{\prod_k \pi(A_k|S_k)}{\prod_k \pi'(A_k|S_k)}\right]$$

$$Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

$$\mathbb{E}_{\pi'}\left[\left(G \cdot \prod_{k=0}^{T-1} \frac{\pi(A_k|S_k)}{\pi'(A_k|S_k)}\right)^2\right] = \sum_{traj} \pi'(traj) \cdot (\dots)^2 =$$

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

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

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

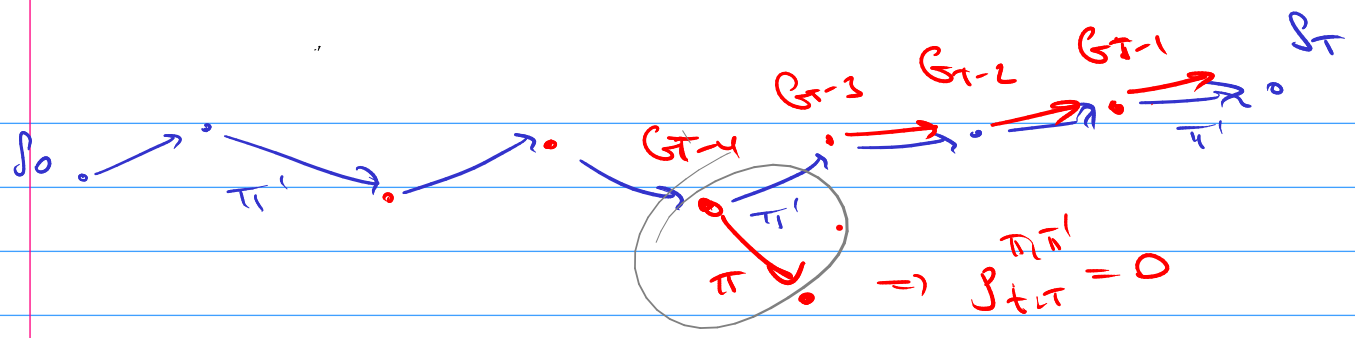
$$= \frac{1}{10}\sum_{k=0}^{\infty}\left(\frac{9}{10}\right)^k\left(\frac{1}{2}\right)^{k+1}\cdot 2^{2(k+1)} = \frac{1}{5}\sum_{k=0}^{\infty}\left(\frac{9}{10}\right)^k\cdot 2^k =$$

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

## Off-policy MC control

- init
- repeat:
  - порождаем $S_0, A_0, R_1, \dots, R_T, S_T$ по $\pi'$, $G=0$, $W=1$
  - for $t = T-1, \dots, 0$:
    - $G = \gamma G + R_{t+1}$, $W(S_t, A_t) = W(S_t, A_t) + W$
    - $Q(S_t, A_t) := Q(S_t, A_t) + \frac{W}{W(S_t, A_t)}\left[G - Q(S_t, A_t)\right]$
    - $\pi(S_t) := \operatorname{argmax}_a Q(S_t, a)$, $W := W \cdot \frac{1}{\pi'(A_t|S_t)}$

- if $A_t \neq \pi(S_t)$ then break




---

TD-learning

&

temporal difference

(bootstrapping)

MC

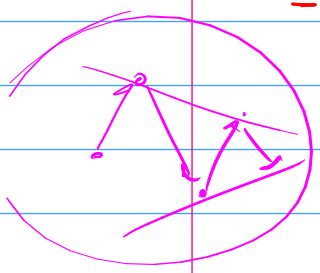
$$V_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] =$$
$$= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] =$$
$$= \mathbb{E}_\pi[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s]$$

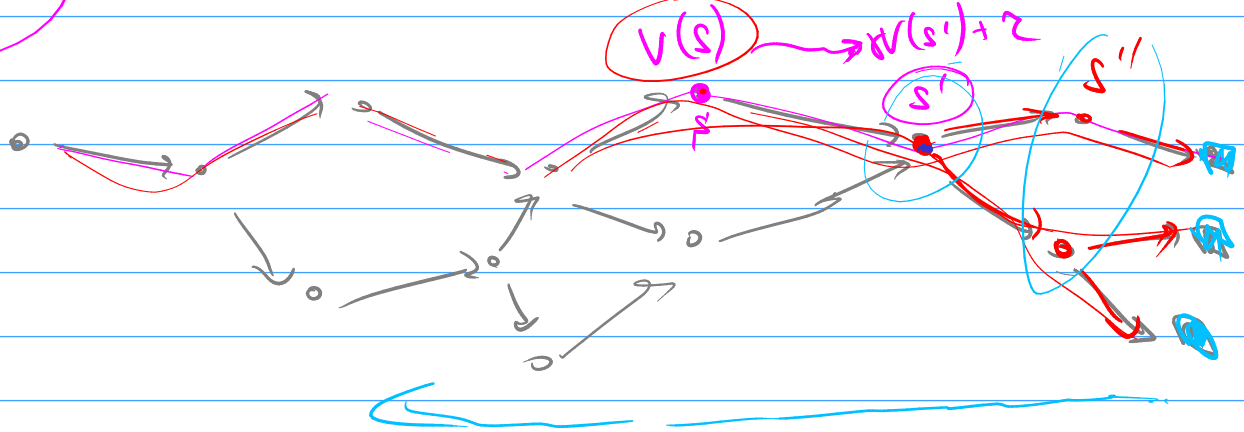
TD

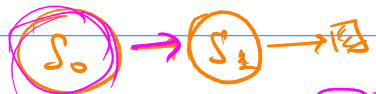
## TD(0) for estimation

- init $V_\pi(s)$, $s \in S$
- repeat:
  - сделаем ход $A_t$ в сост. $S_t$, получим $R_{t+1}, S_{t+1}$ (no $\pi$)
  - обновляем

$$V_\pi(S_t) := V_\pi(S_t) + \alpha\left(R_{t+1} + \gamma V_\pi(S_{t+1}) - V_\pi(S_t)\right)$$
$$Q_\pi(S_t, A_t) := Q_\pi(S_t, A_t) + \alpha\left(R_{t+1} + \gamma Q_\pi(S_{t+1}, A_{t+1}) - Q_\pi(S_t, A_t)\right)$$






$\gamma = 1$

$S_1 \to$ R=1
$S_1 \to$ R=0
$S_1 \to$ R=1
$S_1 \to$ R=1

$S_0 \to S_1 \to$ R=0

$S_1 \to$ R=1

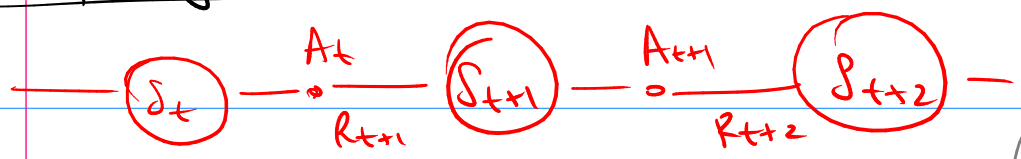
$V(S_1) = 2/3$

$V^{MC}(S_0) = 0$ $\quad G_1 = 0$

$V^{TD}(S_0) = 2/3$

# On-policy TD control

$(S, a, r, s')$




$(S, a, r, s', a')$

TD(0):

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

$\pi(S) = \arg\max_a Q(S, a)$

$\pi(S) = \begin{cases} \text{--''--}, & 1-\varepsilon \\ \text{unit}, & \varepsilon \end{cases}$

## SARSA

- init Q
- repeat
  - сделать A по $\pi^Q$, получив S', R
  - выбрать A' в S' по $\pi^Q$
  - обновить Q(S, A) по TD(0)
  - S := S', A := A'

---

# Off-policy TD control

$$\pi^Q = \begin{cases} \arg\max Q(S, a), & 1-\varepsilon \\ \text{unit}, & \varepsilon \end{cases}$$

TD(0):

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

## Q-learning

DQN