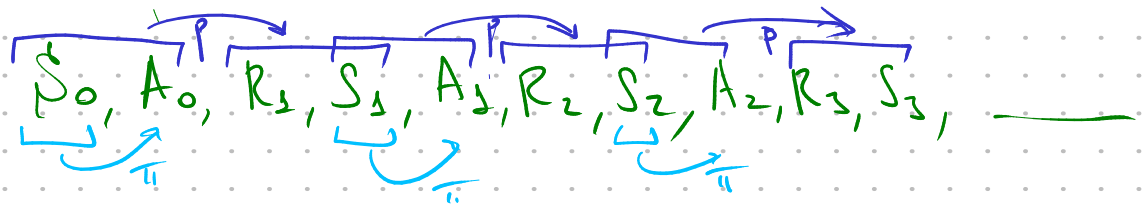


$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$$

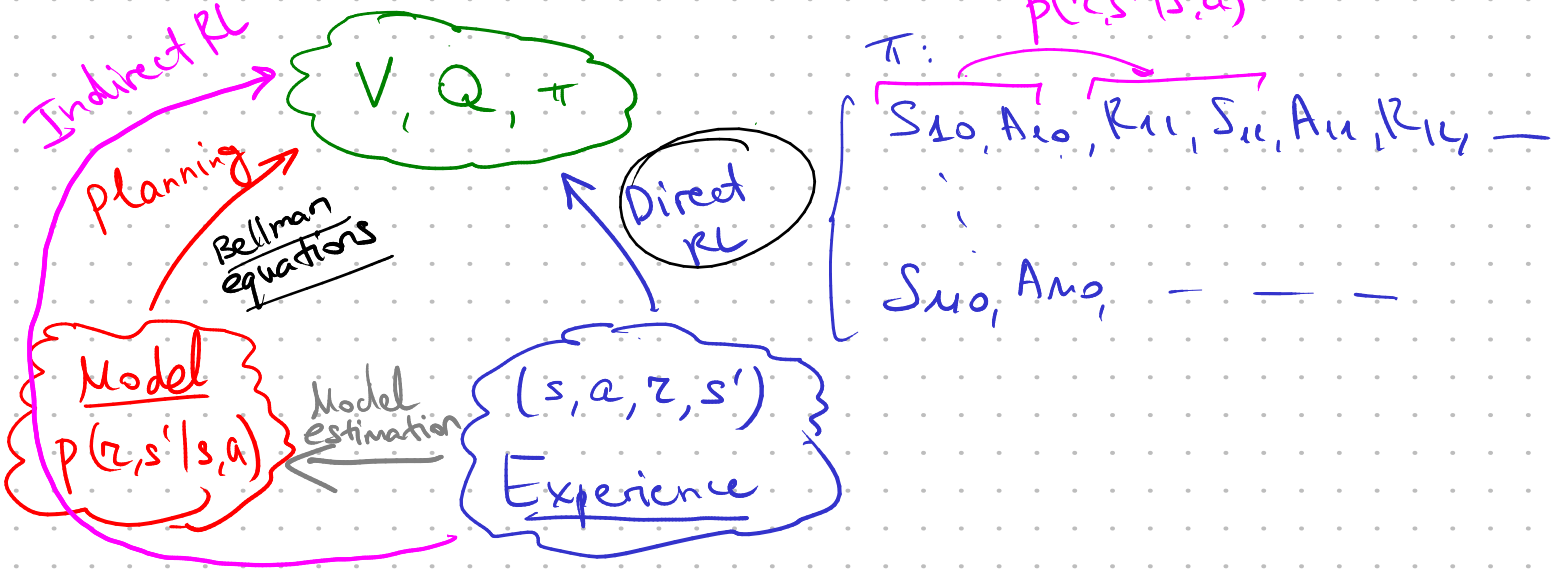
$$V_\pi(s) = E_\pi[G_t \mid S_t = s]$$

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

π^*

Vs $V_{\pi^*}(s) = V_*(s) = \max_\pi V_\pi(s)$

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



$$V_\pi(s) = E_\pi[G_t \mid S_t = s] \approx \text{Avg}(G_t \mid S_t = s, \pi)$$

first-visit MC

① Monte Carlo estimation for V

G_1

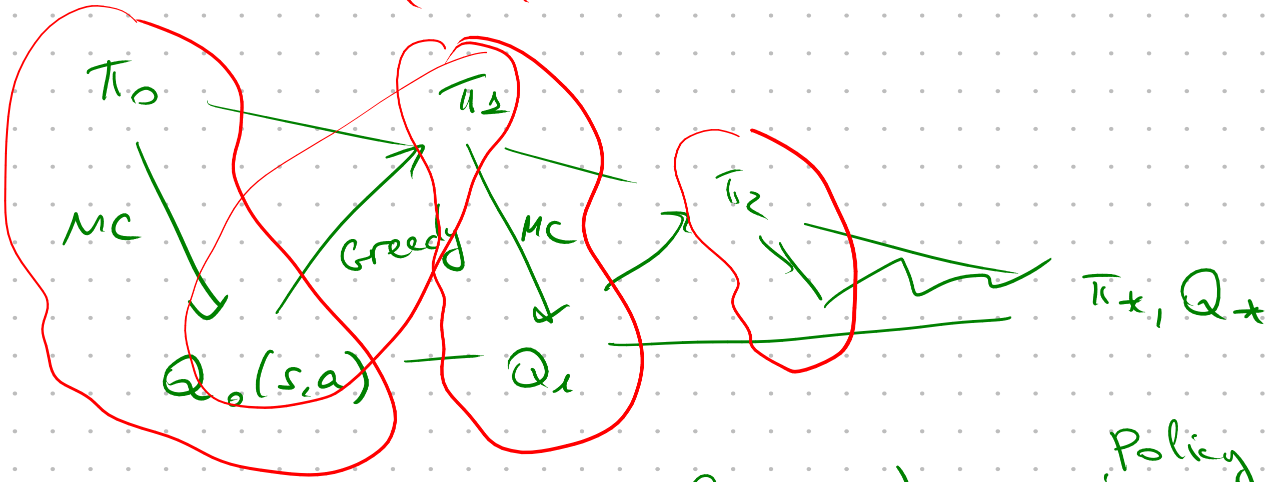
- init
- repeat:
  - gen $S_0, A_0, R_1, S_1, A_1, \ldots S_t, A_t, R_{t+1}, S_{t+1}, \ldots S_T$ по сценарию π
  - $G := 0$
  - for $t = T-1, T-2, \ldots, 0$
    - $G := \gamma G + R_{t+1}$
    - если надо, то запомнить G в $\text{Returns}(S_t)$
      $V(S_t) := \text{Avg}(\text{Returns}(S_t))$

— " — for Q:

— " — $Q(S_t, A_t)$ · —

Exploring starts

$\forall (s,a) \quad p(s,a) > 0$



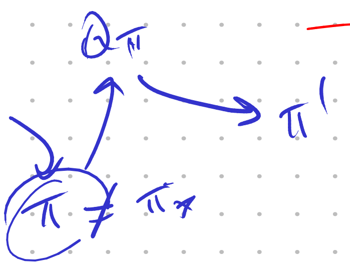
$$\pi_{k+1}(s) = \arg\max_a Q_k(s,a)$$

Policy improvement theorem

② MC control with exploring starts

— " — —

— " — —



- $Q(S_t, A_t) := Avg(Ret(S_t, A_t))$
- $\pi(S_t) := \arg\max_a Q(S_t, a)$

---

③ On-policy MC control

without Expl. starts

— " — —

— " — $Q(S_t, A_t) := Avg(Ret(\_ \ \_))$

ε-soft

$$\pi(S_t) = \begin{cases} \arg\max_a Q(S_t,a) & \text{с вер. } 1-\varepsilon \\ \text{uniform} & \text{— " — } \varepsilon \end{cases}$$

$$\pi(a|s) = \frac{\varepsilon}{|A|} + (1-\varepsilon)\cdot[a = \arg\max_{a'} Q(s,a')]$$

Упр. докажите Policy Improvement Thm для ε-soft стратегий

→ π_*-оптим. среди ε-soft

④ Off-policy MC control

порожд. эпизоды по π', оцениваем Q_π $(\pi \neq \pi')$

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

π' — неизвестная стратегия; наша стратегия

Importance sampling

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

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

$$S_t, A_t : \underbrace{R_{t+1}, S_{t+1}, A_{t+1}, R_{t+2}, \ldots, S_T}_{\text{Траектория Traj}}$$

$$p(\text{Traj} \mid \pi) = p(R_{t+1}, S_{t+1} \mid S_t, A_t) \cdot p(A_{t+1} \mid S_{t+1})\, p(R_{t+2}, S_{t+2} \mid S_{t+1}, A_{t+1}) \ldots p(R_T, S_T \mid S_{T-1}, A_{T-1})$$

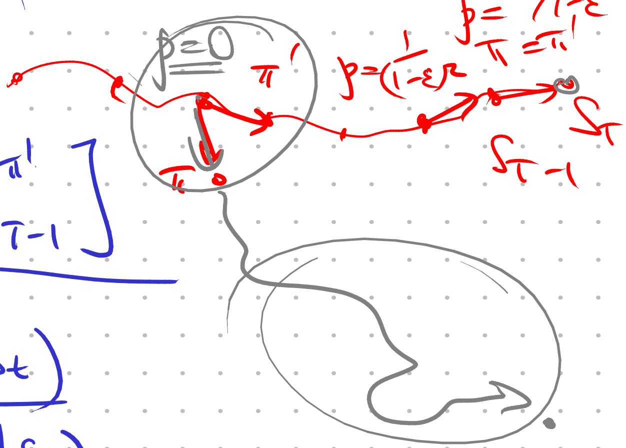
динамика MDP; $p(A_{t+1} \mid S_{t+1}) = \pi(A_{t+1} \mid S_{t+1})$

$$p(\text{Traj} \mid \pi') = p(R_{t+1}, S_{t+1} \mid S_t, A_t)\, p(A_{t+1} \mid S_{t+1}) \ldots$$

динамика MDP; $p(A_{t+1} \mid S_{t+1}) = \pi'(A_{t+1} \mid S_{t+1})$

$$\rho^{\pi,\pi'}_{t+1:T-1} = \frac{p(\text{Traj} \mid \pi)}{p(\text{Traj} \mid \pi')} = \frac{\pi(A_{t+1} \mid S_{t+1})\,\pi(A_{t+2} \mid S_{t+2}) \ldots \pi(A_{T-1} \mid S_{T-1})}{\pi'(\ldots)\,\pi'(\ldots) \ldots \pi'(\ldots)}$$

Off-policy estimation:



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

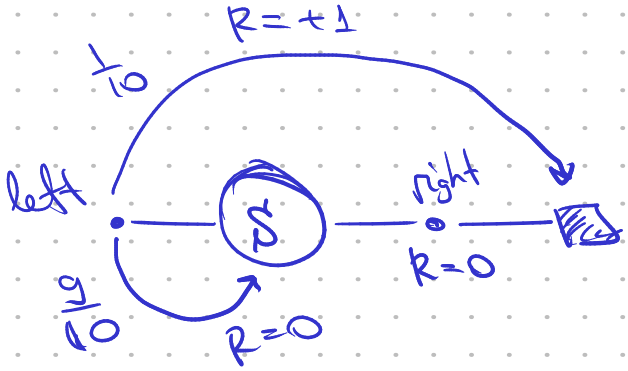
$$\rho^{\pi,\pi'}_{t:T-1} = \rho^{\pi,\pi'}_{t+1:T-1} \cdot \frac{\pi(A_t \mid S_t)}{\pi'(A_t \mid S_t)}$$

$$\text{Returns}(S_t) \leftarrow G_t \cdot \rho^{\pi,\pi'}_{t:T-1}$$

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

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

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



$\pi(\text{left}|s) = 1$

$\pi'(\text{left}|s) = \pi'(\text{right}|s) = \frac{1}{2}$

$$V_\pi(s) = \mathbb{E}_{\pi'}\left[G \cdot \frac{\prod_t \pi(A_t|S_t)}{\prod_t \pi'(A_t|S_t)} \right]$$

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

$$\mathbb{E}_{\pi'}\left[\left(G \cdot \prod_{t=0}^{T-1} \frac{\pi(A_t|S_t)}{\pi'(A_t|S_t)} \right)^2 \right] = \sum_{\text{Traj}} \pi'(\text{Traj}) \cdot (\quad)^2 =$$

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

$$+ \frac{1}{2} \cdot \frac{9}{10} \cdot \frac{1}{2} \cdot \frac{9}{10} \cdot \frac{1}{2} \cdot \frac{1}{10} \cdot 1 \cdot \left(\frac{1}{(\frac{1}{2})^3} \right)^2 =$$

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