$S_t,\ A_t$ $\quad$ $p(s',r|s,a) = \Pr[R_t = r, S_t = s' | S_{t-1} = s, A_{t-1} = a]$ $\quad$ $p(r|s,a)$ $\quad$ $p(s'|s,a)$

$\pi: S \to \Delta[A(s)]$, $\quad \pi(a|s)$

$$G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k = R_{t+1} + \gamma G_{t+1}$$

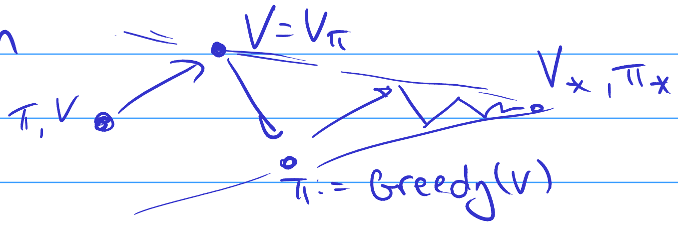
$$V_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s]$$
$$Q_\pi(s,a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a]$$

$\left.\right\}$ Bellman equations

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

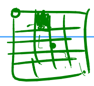
Value iteration $\quad V_*(s)$

Policy iteration



① Monte-Carlo estimation $\quad \pi,\ V_\pi\ Q_\pi \quad p(r, s'|s,a)$

- init π, $V_\pi(s)$, Returns$(s) := [\,]$ $\quad V(s)$



- loop:
  - пройд. эпизод по π, $\quad G := 0$
  - $\forall t = T-1, T-2, \dots, 0$
    - $G := \gamma G + R_{t+1}$
    - если надо:
      - Returns(S_t).append(G)
      - $V_\pi(S_t) :=$ Avg(Returns(S_t))

- init random (s_0, a_0) $\quad s_0$

Exploring starts

first-visit MC

every-visit MC

$V_\pi(S_t),\ N(S_t)$

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

$Q_\pi(s,a)$ $\quad$ — сокращать exploration

② On-policy MC control

- — // — $\quad \pi$ — ε-мягкой, $\quad \forall a, s \quad \pi(a|s) \geq \varepsilon / |A(s)|$
- loop:
  - — // — , $\quad G := 0$
  - $\forall t = T-1, T-2, \dots, 0.$
    - $G = \gamma G + R_{t+1}$

$\pi: S \to$ расп. на $A(s)$

$Q_*: (S, A) \to \mathbb{R}$

- если надо:
  - Returns(S_t, a_t).append(G)
  - $\boxed{Q(s_t, a_t)} := \mathrm{Avg}(\mathrm{Returns}(S_t, a_t))$
  - $\forall a: \quad \pi(a|s_t) = \begin{cases} 1 - \varepsilon + \varepsilon/|A(S_t)|, & a = a^* \\ \varepsilon/|A(S_t)|, & a \neq a^* \end{cases}$

$\underset{a}{\arg\max}\, Q(S_t, a)$
$\|$
$a = a^*$

(3) Off-policy MC control

? [- порождать эпизоды по b (behaviour)
- обучать не $V_b(s), Q_b(s,a)$, а $V_\pi(s), Q_\pi(s,a)$, $\pi \neq b$]

Importance sampling

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

$G_t \sim$ сгоп b

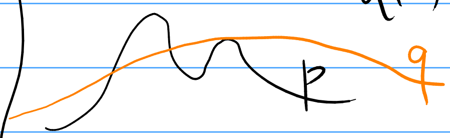
S_t: $\underbrace{A_t, S_{t+1}, A_{t+1}, \dots, A_{T-1}, S_T}_{\text{Traj}}$

$$\Pr[\text{Traj} \mid \pi, S_t] = \pi(A_t|S_t)\, p(S_{t+1}|S_t, A_t)\, \pi(A_{t+1}|S_{t+1})\, p(S_{t+2}|\dots) \dots p(S_T|S_{T-1}, A_{T-1}) = \prod_{k=t}^{T-1} \pi(A_k|S_k)\, p(S_{k+1}|S_k, A_k)$$

$\mathbb{E}_{p(x)}[f(x)]$, $x_n \sim q(x)$, $\int \dots q(x)dx$

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

imp. weights



Треб.:
$\forall x \quad p(x) > 0 \Rightarrow q(x) > 0$
$\forall x \quad q(x) = 0 \Rightarrow p(x) = 0$

$$\rho_{t:T-1} = \frac{\Pr[\text{Traj} \mid \pi, S_t]}{\Pr[\text{Traj} \mid b, S_t]} = \frac{\prod_k \pi(A_k|S_k)\, \cancel{p(S_{k+1}|S_k, A_k)}}{\prod_k b(A_k|S_k)\, \cancel{p(S_{k+1}|S_k, A_k)}}$$

$$\rho_{t:T-1} = \frac{\prod_{k=t}^{T-1} \pi(A_k|S_k)}{\prod_{k=t}^{T-1} b(A_k|S_k)} = \prod_{k=t}^{T-1} \frac{\pi_k}{b_k}$$

Coverage: $\forall s, a \quad \pi(a|s) > 0 \Rightarrow$
$\Rightarrow b(a|s) > 0$

$$\boxed{V_\pi(s) = \mathbb{E}_b[G_t \cdot \rho_{t:T-1} \mid S_t = s]}$$

$\rho_T = 1 \quad \rho_t = \rho_{t+1} \cdot \left(\frac{\pi_t}{b_t}\right)$

$t,\ G,\ \text{Returns}(S_t).\text{append}(G : \forall t : T-1)$

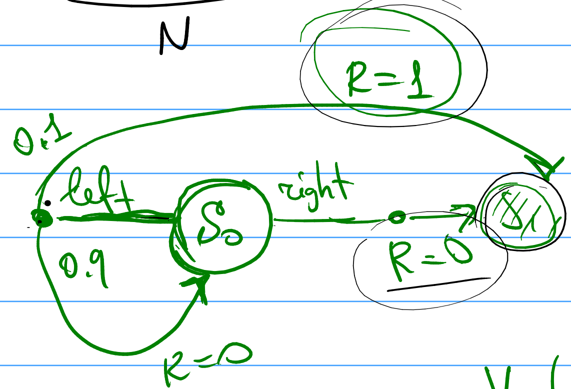
$\searrow$ weighted I.S.

imp. sampling

$$V_\pi(s) = \frac{\sum^N G_i W_i}{N}$$

$$V_\pi(s) = \frac{\sum G_i W_i}{\sum_i W_i}$$

$\boxed{\gamma = 1}$



$\pi(\text{left}\mid S_0) = 1$

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

$V_\pi(S_0) = 1$

$$V_\pi(S_0) = \mathbb{E}_b\left[G \cdot \frac{\prod_k \pi(a_k\mid S_k)}{\prod_k b(a_k\mid S_k)}\right]$$

$\mathrm{Var}(x) = \mathbb{E}[x^2] - (\mathbb{E}x)^2$

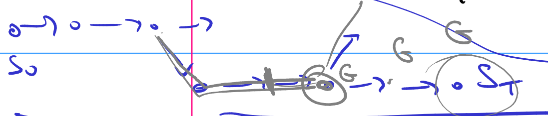
$$\mathbb{E}_b\left[\left(G \cdot \prod_{k=0}^{T-1} \frac{\pi(a_k\mid S_k)}{b(a_k\mid S_k)}\right)^2\right] = \left(\frac{1}{2}\cdot\frac{1}{10}\right)\cdot\left(\frac{1}{1/2}\right)^2 +$$

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

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

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

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



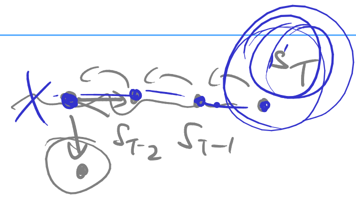
Off policy MC control:

- init
- loop:
  - эпизод чз b, $G := 0$, $p := 1$
  - $\forall t = T-1, \dots, 0$
    - $\hookrightarrow$ (*)

(*)
- $G = \gamma G + R_{t+1}$
- $C(S_t, a_t) = C(S_t, a_t) + p$
- $Q(S_t, a_t) := Q(S_t, a_t) + \frac{p}{C(S_t, a_t)}\cdot\left(G - Q(S_t, a_t)\right)$
- $\pi(S_t) := \arg\max_a Q(S_t, a)$
- $W := W \cdot \frac{\pi(a_t\mid S_t)}{b(a_t\mid S_t)}$

$\pi(S_t) = a_t$?

- если $\pi(S_t) \neq a_t$, то break



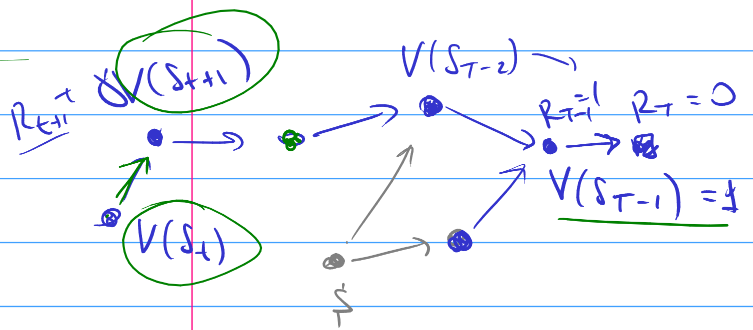
TD-learning (temporal difference) – bootstrapping

TD(1) $V(S_t) = V(S_t) + \alpha [R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2}) - V(S_t)]$

TD(0): $V(S_t) := V(S_t) + \alpha [\underbrace{R_{t+1} + \gamma V(S_{t+1})} - V(S_t)]$

$$V(S_t) \approx G_t = R_{t+1} + \gamma \big(\underbrace{R_{t+2} + (\gamma R_{t+3} + (\ldots + \gamma^{T-t-2} R_T))}\big)$$

$$\text{''} \quad G_{t+1} \approx V(S_{t+1})$$



S_0 S_1

1) $S_1 \to R = 1$
2) $S_1 \to R = 0$
3) $S_1 \to R = 1$
4) $S_1 \to R = 1$
5) $S_0 \to S_1 \to R = 0$
6) $S_1 \to R = 1$

MC: $V(S_1) = \frac{2}{3}$

$V(S_0) = 0$

TD: $V(S_1) = 2/3$

$V(S_0) = 2/3$

On-policy TD control – Sarsa

$Q_\pi(s,a)$

$\pi = \pi(Q)$

$— S_t — \underset{a_t}{\bullet} \xrightarrow{R_{t+1}} S_{t+1} — \underset{a_{t+1}}{\circ} \xrightarrow{R_{t+2}} S_{t+2} — \ldots$

- init
- Loop по шагам эпизода:
  - s, a, s', R $\quad (s, a, R, s', a')$
  - выберем a' в s' по $\pi(s') =$ ε-жадн. стр. по $Q(s', a)$
  - $Q(s,a) := Q(s,a) + \alpha [R + \gamma Q(s', a') - Q(s,a)]$
  - $s = s', a = a'$

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

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

$Q_*(s,a)$ — π_*