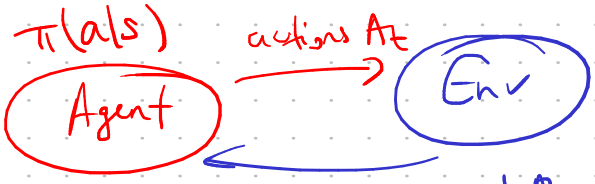
① MDP $\pi(a|s)$



$S_0, A_0, R_1, S_1, A_1, R_2, S_2 \dots$

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

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

$$V_\pi(s) = E_\pi[G_t | S_t = s] \qquad V_*(s) = \max_\pi V_\pi(s)$$

$$Q_\pi(s,a) = E_\pi[G_t | S_t = s, A_t = a] \qquad Q_*(s,a) = \max_\pi Q_\pi(s,a)$$

$$\pi_*(s) = \arg\max_a Q_*(s,a)$$

Bellman equations

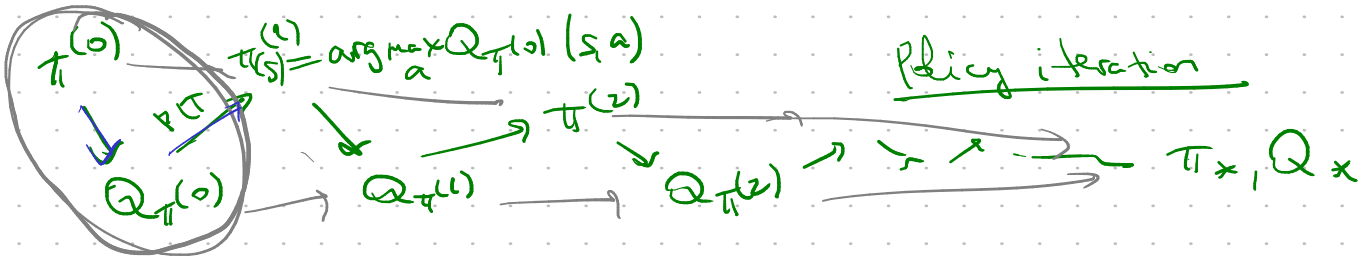
Policy improvement theorem

1) ~~Много π, $|A|^{|S|}$~~

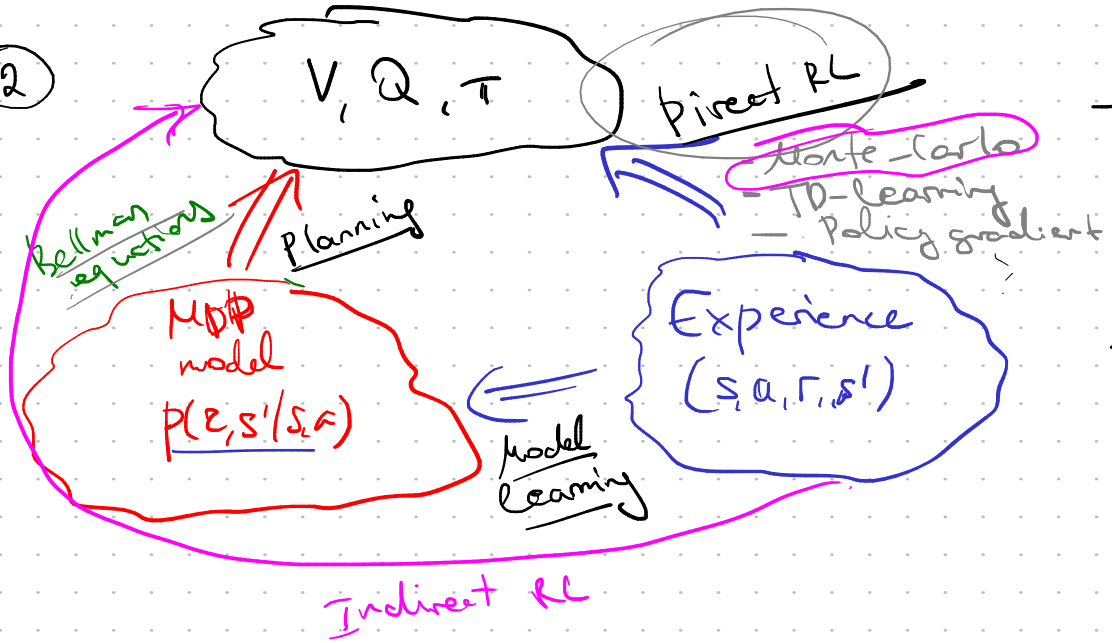
2) Много S, много (s,a)

3) Не знаем $p(r, s' | s, a)$

MP → Беллмана на V_*, Q_*



②



- Tabular RL

S	A	Q
s_1	a_1	-
s_1	a_2	-
...	...	...

- Approximate RL
  Deep RL

③ Monte-Carlo RL

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

Агент по π:

1) $S_{10}, A_{10}, R_{11}, S_{11}, A_{11}, R_{12}, S_{12}, \dots$, S_{1T}
2) $S_{20}, A_{20}, \dots$, S_{2T}
...
M) $S_{M0}, A_{M0}, R_{M1}, S_{M1}, \dots$, S_{MT}

$G_t = G_{12} = R_{t1} + \gamma R_{t+1} + \dots$

Monte-Carlo estimation ($\pi \rightsquigarrow V, Q$)

- init $V(s)$
- repeat:
  - generate w/ π $S_0, A_0, R_1, S_1, A_1, \ldots, S_{t-1}, A_t, R_{t+1}, S_{t+1}, \ldots, S_T$
  - $G := 0$
  - for $t = T-1, \ldots, 0$
    - $G := \gamma G + R_{t+1}$
    - add G to the list Returns(S_t) / Returns(S_t, A_t)
  - $V_\pi(s) := \text{Avg}(\text{Returns}(s))$, $Q_\pi(s,a) := \text{Avg}(\text{Returns}(s,a))$ $\neq \emptyset$

explore

G_t, $G_T = 0$

1) first-visit MC / every-visit MC

2) Exploring starts

Monte-Carlo control with exploring starts

- init π, Q
- repeat:
  - generate episode $(S_t, A_t, R_{t+1}, S_{T-1})$ $(S_0, A_0) \sim \text{Unif}(S, A)$
  - for $t = T-1, \ldots, 1, 0$:
    - add G_t to Returns(S_t, A_t)
    - $Q(S_t, A_t) := \text{Avg}(\text{Returns}(S_t, A_t))$
    - $\pi(S_t) := \underset{a}{\text{argmax}}\, Q(S_t, a)$

π_k, Q_k

Control

On-policy

off-policy

On-policy MC control

$S_0 \sim \text{Unif}(\ldots)$

—— « —— π — ε-soft

$$\pi(S_t) := \begin{cases} \text{Unif}(a) & , \varepsilon \\ \underset{a}{\text{argmax}}\, Q(S_t, a) = a_*, & 1-\varepsilon \end{cases}$$

$$\pi(a|S_t) = \frac{\varepsilon}{|A|} \text{ или } 1 - \varepsilon + \frac{\varepsilon}{|A|} \text{ для } a_*$$

$\forall s: \quad Q_\pi(s, \pi'(s)) \geq V_\pi(s)$ $\quad \pi_1(\pi^1)$ — ε-soft

$$Q_\pi(s, \pi'(s)) = \sum_a \pi'(a|s) Q_\pi(s,a) =$$

$$= \frac{\varepsilon}{|A|} \sum_a Q_\pi(s,a) + (1-\varepsilon) \cdot \max_a Q_\pi(s,a)$$

$$V_\pi(s) = \sum_a \pi(a|s) Q_\pi(s,a) = \frac{\varepsilon}{|A|} \sum_a Q_\pi(s,a) + \sum_a \left(\pi(a|s) - \frac{\varepsilon}{|A|}\right) Q_\pi(s,a)$$

$$\sum_a \left(\pi(a|s) - \frac{\varepsilon}{|A|}\right) = 1 \qquad 1-\varepsilon$$

$$(1-\varepsilon) \cdot \sum_a \frac{\pi(a|s) - \varepsilon/|A|}{1-\varepsilon} Q_\pi(s,a) \leq \max_a Q_\pi(s,a)$$

$$\leq Q_\pi(s, \pi'(s))$$

$Q_\pi = \mathbb{E}_\pi[\dots]$

④ Off-policy MC control

- generate episodes with π', e.g. ε-soft
- learn Q_π for $\pi \neq \pi'$, e.g. greedy

Importance sampling

$$\mathbb{E}_{p(\bar{x})}[f(\bar{x})] = \int p(\bar{x}) f(\bar{x}) d\bar{x} =$$

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

importance weights

$$p(\tau_{s_t:}|\pi) = p(A_t, R_{t+1}, S_{t+1}, A_{t+1}, R_{t+2}, S_{t+2}, \dots | S_t = s, A_t = a, \pi) =$$

$$= \underbrace{p(A_t|S_t)}_{\pi} \underbrace{p(R_{t+1}, S_{t+1}|S_t, A_t)}_{p(r,s'|s,a)} \underbrace{p(A_{t+1}|S_{t+1})}_{\pi} p(R_{t+2}, S_{t+2}|S_{t+1}, A_{t+1}) \dots$$

$$\mathbb{E}_\pi[G_t|S_t] = \mathbb{E}_{\pi'}\left[G_t \cdot \frac{p(\tau_{s_t:}|\pi)}{p(\tau_{s_t:}|\pi')}\right] =$$

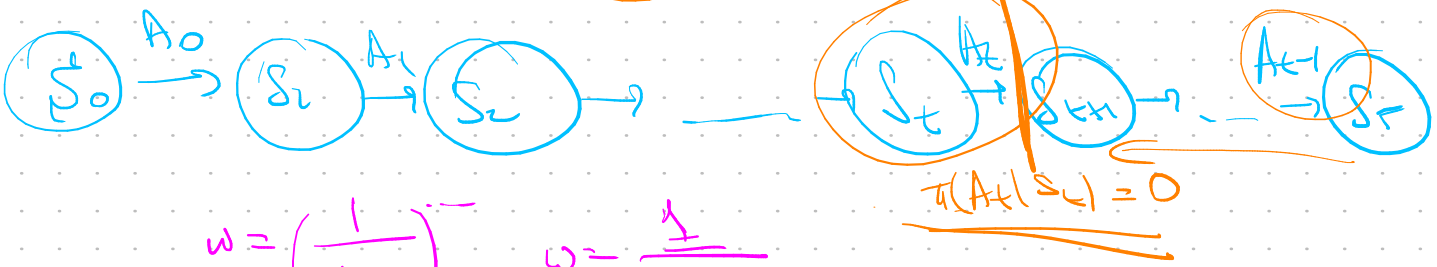
$$= \mathbb{E}_{\pi'}\left[G_t \cdot \frac{\pi(A_t|S_t) \cdot p(R_{t+1}, S_{t+1}|S_t, A_t) \, \pi(A_{t+1}|S_{t+1}) \, p(\dots|\dots) \dots}{\pi'(A_t|S_t) \, p(R_{t+1}, S_{t+1}|S_t, A_t) \, \pi'(A_{t+1}|S_{t+1}) \, p(\dots|\dots) \dots}\right]$$

$$V_\pi(s_t) = \mathbb{E}_{\pi'}\left[G_t \cdot \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{\pi'(A_k|S_k)} \right] = W_t^{\pi,\pi'}$$

Off-policy MC control

$\pi'(a|s) = 0 \Rightarrow \pi(a|s) = 0$

- init
- repeat:
  - generate w/π': $S_0, A_0, R_1, S_1, A_1, \dots$
  - $G := 0$, $W := 1$
  - for $t = T-1, \dots, 0$:
    - $G := \gamma \cdot G + R_{t+1}$
    - $W := W \cdot \frac{\pi(A_t|S_t)}{\pi'(A_t|S_t)}$ } если $\pi(A_t|S_t) = 0$, то $W_{t'}^{\pi,\pi'} = 0 \ \forall t' \le t$
    - add $G \cdot W$ to Returns(S_t, A_t)
    - $\pi' := \varepsilon\text{-soft}(Q_\pi)$, $\pi'(a|S_t) = \begin{cases} \varepsilon/|A|, & \varepsilon \\ 1-\varepsilon, & a = \operatorname{argmax} Q_\pi \end{cases}$



$$w = \left(\frac{1}{1-\varepsilon}\right)^{\cdot} \qquad w = \frac{1}{\prod_t \pi'(a_t|s_t)}$$

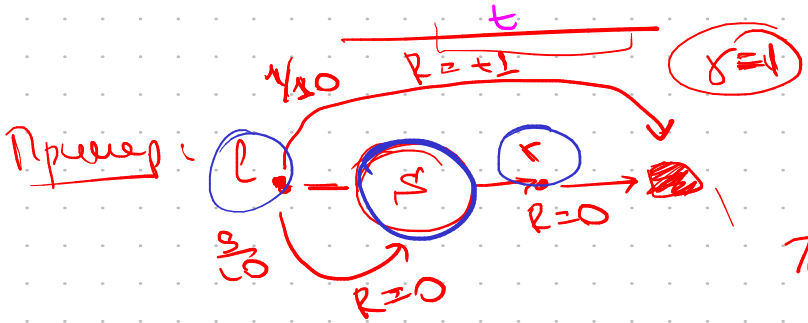
1) Ordinary imp. s.

$$V_\pi(s) := \text{Avg}(\text{Returns}(s)) = \frac{1}{|\text{Ret}|} \cdot \sum_{t:S_t=s} G_t \cdot W_t^{\pi,\pi'}$$

2) Weighted imp. s.:

$$V_\pi(s) := \frac{\sum_t G_t \cdot W_t^{\pi,\pi'}}{\sum_t W_t^{\pi,\pi'}}$$

Пример:



$Q(S, \text{right}) = 0$

$Q_*(S, \text{left}) = 1$

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

$\mathbb{E}_{\pi'}[G_t \cdot w_t^{\pi,\pi'}] = 1$ $\quad Var[--] = ?$

$Var[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \mathbb{E}[x^2] - 1$

$$\mathbb{E}_{\pi'}\left[\left(G_0 \cdot \prod_{t=0}^{T-1} \frac{\pi(A_t|S_t)}{\pi'(A_t|S_t)}\right)^2\right] = \sum_{Traj} p(Traj|\pi') \cdot (W)^2 \cdot G_0^2$$

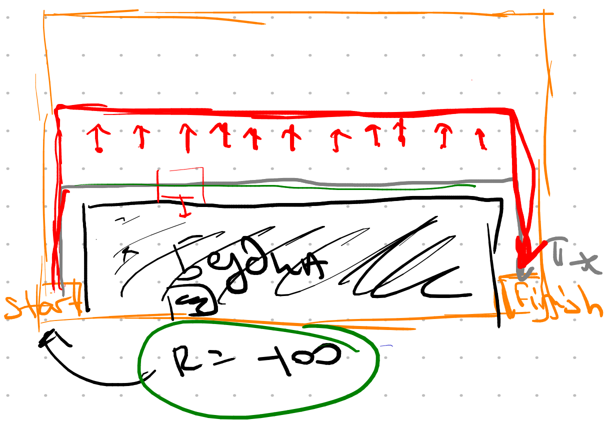
$$= \sum_{Traj:\, G=1} \left(w_0^{\pi,\pi'}\right)^2 p(Traj|\pi') =$$

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

l, +1 $\qquad$ l, l, +1

$$= \sum_{t=0}^{\infty} \frac{1}{10}\cdot\left(\frac{9}{10}\right)^t\cdot\left(\frac{1}{2}\right)^{t+1}\cdot 2^{2(t+1)} = \frac{2}{10}\cdot\sum_t\left(\frac{9}{10}\cdot\frac{1}{2}\cdot 4\right)^t =$$

$$= \frac{1}{5}\cdot\sum_t\left(\frac{9}{5}\right)^t$$



R = -1

$\bar{p} = (p_1 \ldots p_n)$ + softmax(Q)

$- i \rightarrow$ Win $\quad p_i \uparrow \quad p_s \downarrow$

(5) TD-learning (temporal difference)

$$V_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s] = \mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots | S_t = s]$$

$$= \mathbb{E}_\pi[R_{t+1}] + \gamma \cdot \mathbb{E}_\pi[V_\pi(S_{t+1})]$$

$\approx R_{t+1} \qquad \approx G_{t+1}$

G_t

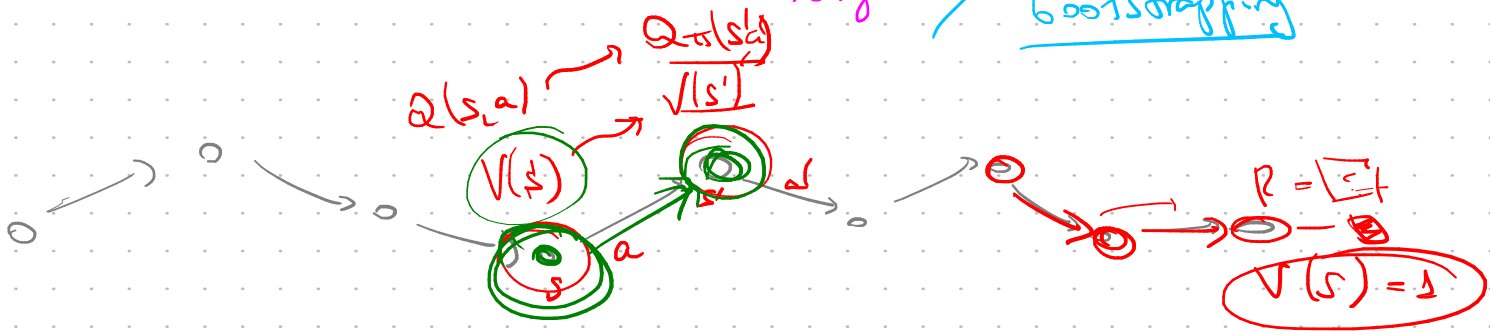
$S_0, A_0, R_1, S_1, A_1, R_2, S_2, \ldots, S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}, \ldots$

$$V_\pi(S_t) \approx R_{t+1} + \gamma V_\pi(S_{t+1})$$

Experience tuple: (s, a, r, s')

$$V_\pi(s) := V_\pi(s) + \alpha\left(r + \gamma V_\pi(s') - V_\pi(s)\right)$$

Target / Bootstrapping



TD estimation

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

$D =$ $(s' \to \blacksquare, +1)$, $(s' \to \blacksquare, +1)$, $(s' \to \blacksquare, 0)$, $(s' \to \blacksquare, +1)$, $(s \to s' \to \blacksquare, 0)$

MC: $V(s') \approx 3/5$

TD: $V(s') \approx 3/5$

$V(s) \approx 3/5$

MC: $V(s) = 0$

Backup diagrams

Bellman eq: $V(s) = \sum_a \pi(a|s) \sum_{r,s'} p(r, s' | s, a)\left(r + \gamma V(s')\right)$