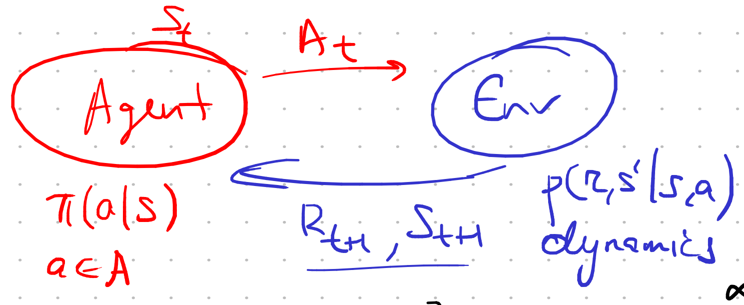
① **RL**



policy $\pi(a|s)$, $a \in A$

MDP: $S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, \ldots$

Return $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

$$\begin{cases} V_\pi(s) = E_\pi[G_t \mid S_t = s] \\ Q_\pi(s,a) = E_\pi[G_t \mid S_t = s, A_t = a] \end{cases}$$

$V_*(s) = \max_\pi V_\pi(s)$

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

Bellman equations

V_π, Q_π, V_*, Q_*

lin | max(–)

~~1)~~

2) Много S, пар (s,a) ← Approximate RL

3) Не знаем $p(r,s'|s,a)$

② **Monte-Carlo RL**

$V_\pi(s) = E_\pi[G_t \mid S_t = s] \approx \frac{1}{R}\sum_{i=1}^{R} G_t^{(i)}$

Agent; π:

1) $S_{10}, A_{10}, R_{11}, S_{11}, A_{11}, R_{12}, \ldots \overset{s}{=} S_{1t}, A_{1t}, R_{1,t+1}, S_{1,t+1}, \ldots$

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

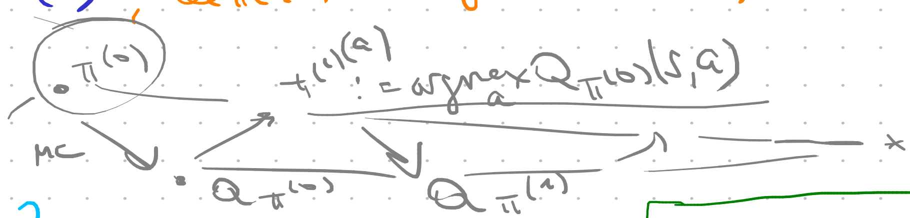
2) $S_{20}, A_{20}, \ldots$

$G_t \mid s,a \rightarrow Q(s,a)$

Monte Carlo estimation (π)

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

1) Policy improvement



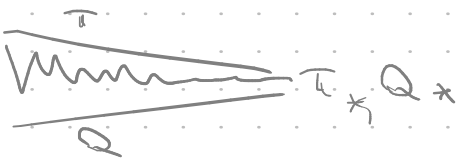
2) Returns(s,a) ≠ ∅ ?

exploring starts / soft strategies

3) first visit vs. every visit MC.

Sutton, Barto
RL: an introduction
1998, 2018

Monte Carlo control w/ expl. starts



- init π, Q
- repeat:
  - sample S_0, A_0
  - generate $S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, \ldots, S_T$ ~ π
  - for $t = T-1, \ldots, 0$:
    - $G := \gamma G + R_{t+1}$
    - add G to $Returns(S_t, A_t)$
    - $Q(S_t, A_t) := Avg(Ret(S_t, A_t))$
    - $\pi(S_t) := \underset{a}{\operatorname{argmax}}\, Q(S_t, a)$

(3) On-policy vs. off-policy MC control

учимся по π, оценив. Q_π

учимся по π', оценив. Q_π

- on-policy MC control

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

π'

$$\pi(a|S_t) = \frac{\varepsilon}{|A|}, \; a \neq a_* \quad ; \quad (1-\varepsilon) + \frac{\varepsilon}{|A|}, \; a = a_*$$

Пусть $\forall s \; Q_\pi(s, \pi'(s)) \geq V_\pi(s)$, где π, π' — ε-soft

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

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

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

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

$$\sum_a = 1 \qquad \leq \max_a Q_\pi(s,a)$$

- off-policy mc control

- generate episodes with π', π' - ε-soft

- learn Q_π for $\pi \neq \pi'$, e.g., π - greedy

importance sampling: $\mathbb{E}_p[f(\bar{x})] = \int \frac{p(\bar{x})}{q(\bar{x})} q(\bar{x}) f(\bar{x}) d\bar{x} = \mathbb{E}_{q \sim \pi'}\left[f(\bar{x}) \frac{p(\bar{x})}{q(\bar{x})} \right]$

$$p(\text{Traj}(\pi)^{S_t=s}) = p(A_t, R_{t+1}, S_{t+1}, A_{t+1}, \dots \mid S_t = s, \pi) =$$

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

$V_\pi(S_t)$ — $\pi(a|s)$ — $p(r, s'|s, a)$

$$\mathbb{E}_\pi[G_t | S_t] = \mathbb{E}_{\pi'}\left[G_t \cdot \frac{p(\text{Traj}|\pi)}{p(\text{Traj}|\pi')} \right] = \mathbb{E}\left[G_t \cdot \frac{\pi'(A_t|S_t) \dots}{\pi'(A_t|S_t) \dots} \right]$$

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

$\pi \neq A_{T-1} \sim \pi'$

$S_{T-3} \to S_{T-2} \to S_{T-1} \to S_T$

- init

- repeat:

  - gen. with π': $S_0, A_0, R_1, S_1, \dots$

  - $G \leftarrow 0$, $w \leftarrow 1$

  - for $t = T-1, \dots, 0$:

    - $G \leftarrow \gamma \cdot G + R_{t+1}$

    - $w \leftarrow w \cdot \frac{\pi(A_t|S_t)}{\pi'(A_t|S_t)}$

  - add $w \cdot G$ to Returns(S_t, A_t), update Q

  - $\pi'(a|S_t) \leftarrow \begin{cases} a_*, & 1-\varepsilon \\ \text{uni}, & \varepsilon \end{cases}$, $\pi(a|S_t) = a_*$

remark: each $\pi(a|s) > 0$ so $\pi'(a|s) > 0$

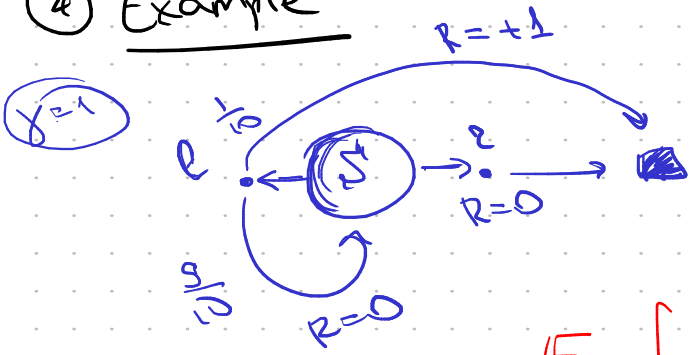
$S_0 \xrightarrow{A_0} S_1 \xrightarrow{A_1} S_2 \to \dots \xrightarrow{A_{t-1}} S_t \to S_T$

$\frac{\pi \cdot \pi \cdot \pi \cdot \dots}{\pi' \cdot \pi' \cdot \pi' \cdot \dots} \quad \frac{1}{\pi'} \cdot \frac{\pi}{\pi} \cdot 1$

$$\text{Avg}(\text{Returns}(\cdot,\cdot)) = \frac{1}{|\text{Ret}|} \cdot \sum_t G_t \, W_t^{\pi,\pi'}$$

$$= \frac{\sum_t G_t \cdot W_t^{\pi,\pi'}}{\sum W_t^{\pi,\pi'}}$$

④ Example



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

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

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

$$\mathbb{E}_{\pi'}\left[G_t \cdot W_t^{\pi,\pi'}\right] = 1$$

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

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

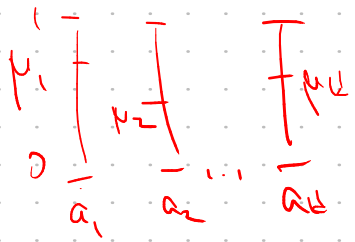
$$= \sum \left(W_0^{\pi,\pi'}\right)^2 \cdot p(\text{traj}|\pi') = \quad [\text{traj}: l\text{-}a,\ l\text{-}l\text{-}a,\ l\text{-}l\text{-}l\text{-}a, \dots]$$

traj: $G = 1$

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

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

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



ε-Greedy, $\hat{\mu}_i^{(0)} := 10$

---

⑤ TD-learning (temporal difference)

$G_t^{(2)}$

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

$$= \mathbb{E}_\pi[R_{t+1} | S_t = s] + \gamma \mathbb{E}_\pi[R_{t+2} + \gamma^2 R_{t+3} + \dots | S_t = s]$$

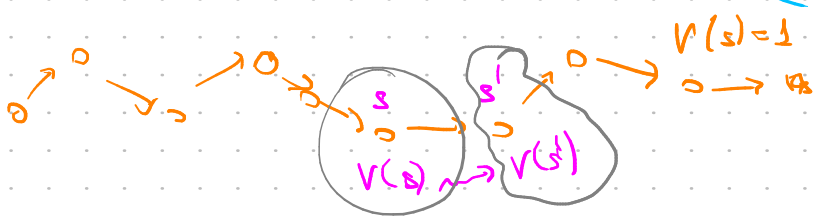
$\approx R_{t+1}$ $\qquad$ $\approx G_{t+1}$

$S_0, A_0, R_1, S_1, A_1, \ldots S_t, A_t, R_{t+1}, S_{t+1}, \ldots$ — bootstrap

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

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

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



TD estimation

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

target — old value

$$Q_\pi(s,a) := Q_\pi(s,a) + \alpha (r + \gamma \cdot Q_\pi(s',a') - Q_\pi(s,a))$$

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

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

$V(s) \approx 3/5$

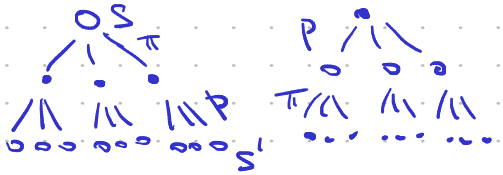
MC: $V(s') = 3/5$

$V(s) = 0$

Backup diagrams

TD:

Bellman equations:



MC

n-step TD:

$$Q_\pi(S_t, A_t) = \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots | S_t, A_t]$$

$$Q_\pi(S_t, A_t) := Q_\pi(S_t, A_t) + \alpha (R_{t+1} + \gamma R_{t+2} + \gamma^2 Q_\pi(S_{t+2}, A_{t+2}) - Q_\pi(S_t, A_t))$$

TD: 1-step → 2-step → 3-step → ... → MC

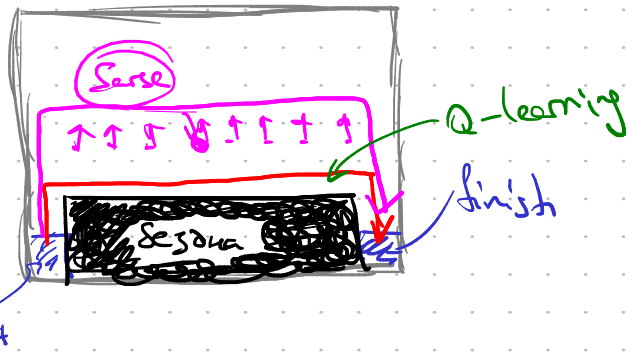
⑥ On-policy TD control / Sarsa

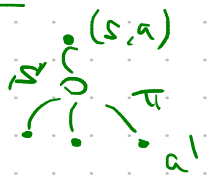
Sarsa: ... — выбираю a по $\pi(a|s) = \begin{cases} \text{argmax}_a Q(s,a), & 1-\varepsilon \\ \text{Unif}, & \varepsilon \end{cases}$

— наблюдаю r, s', выб. a'

— $Q(s,a) := Q(s,a) + \alpha(r + \gamma \cdot Q(s',a') - Q(s,a))$

exp. tuple (s, a, r, s', a')





Expected Sarsa

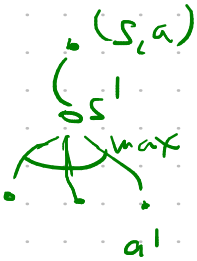
$$Q(s,a) := Q(s,a) + \alpha\left(r + \gamma \cdot \sum_{a'} \pi(a'|s') Q(s',a') - Q(s,a)\right)$$

target

$$\text{Target} = r + \gamma \cdot \sum_{a_1 \neq a'} \pi(a_1|s') Q(s', a_1) + \gamma \cdot \pi(a'|s')\left(r' + \gamma \cdot \sum_{a_2} \pi(a_2|s'') Q(s'', a_2)\right)$$

⑦ Q-learning / off-policy TD control (Watkins, 1989)

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



$Q(s,a) = \mathbb{E}_\pi[G_t | s,a]$ $\quad \mathbb{E}[\max] \neq \max \mathbb{E}[\cdot]$

$\approx$ max

$Q(s,a) \rightsquigarrow \max Q(s',a')$

Winner's curse — Аукционы

$\hat{x}_1, \hat{x}_2, \dots, \hat{x}_k$

$\hat{x}_k \sim N(\hat{x}_k | x, \sigma^2)$

x

Double Q-learning

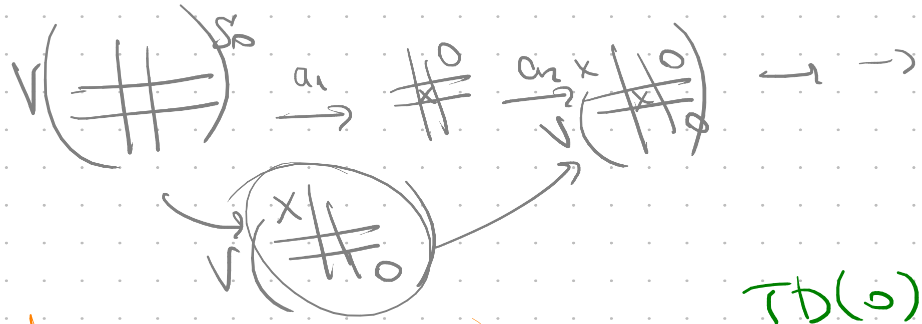
$Q_1(s,a)$ $\quad$ $Q_2(s,a)$

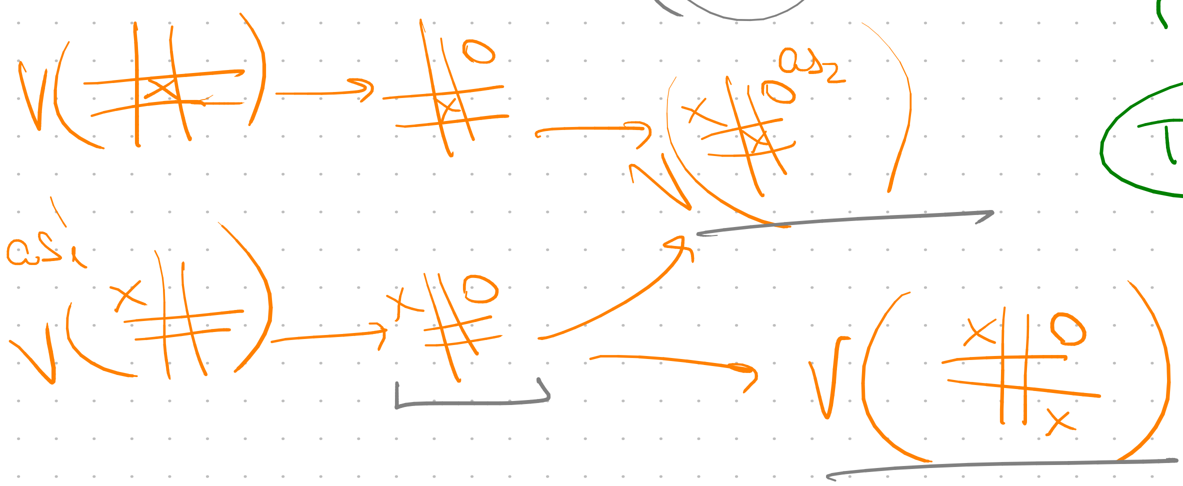
(s,a,r,s')

$p = \frac{1}{2}:\quad Q_1(s,a) := Q_1(s,a) + \alpha\left[r + \gamma\, Q_2\left(s', \arg\max_{a'} Q_1(s',a')\right) - Q_1(s,a)\right]$

$p = \frac{1}{2}:\quad Q_2(s,a) := \text{—}\,''\,\text{—}\; Q_1\left(s', \arg\max_{a'} Q_2(s',a')\right) \text{—}\,''\,\text{—}$

Afterstates





TD(0)

TD(λ)

8 Approximate RL

$s \to [\;] \to \hat{V}(s,\overline{w}) \approx V(s)$, with parameters $\overline{w}$

$s, a \to [\;] \to \hat{Q}(s,a,\overline{w}) \approx Q(s,a)$, with parameters $\overline{w}$

$$L(\overline{w}) = \sum_{s} \left(V(s) - \hat{V}(s,\overline{w})\right)^2 \xrightarrow[\overline{w}]{} \min$$

$$L(\overline{w}) = E_{s}\left[\ell(s,\overline{w})\right]$$

GD: $\quad \overline{w} := \overline{w} - \alpha \cdot \nabla_{\overline{w}} L(\overline{w})$

SGD: $\quad \overline{w} := \overline{w} + \alpha \cdot \sum_{i=1}^{m} 2\left(V(s_i) - \hat{V}(s_i,\overline{w})\right) \cdot \nabla_{\overline{w}} \hat{V}(s_i,\overline{w})$

$$L(\overline{w}) = \sum_{s \in S} \mu(s)\left(V(s) - \hat{V}(s,\overline{w})\right)^2 \xrightarrow[\overline{w}]{} \min$$

$\mu(s) = Pr_{\pi}[S_t = s] =$ „доля врем., когд. в s"

Gradient MC estimation

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

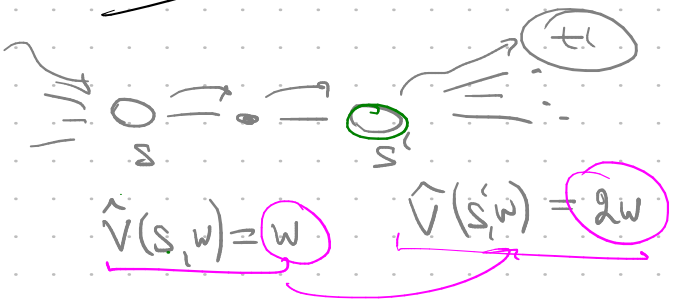
Semi-gradient TD estimation

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

Semi-gradient Sarsa / on-policy TD control

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

~~Off-policy semi-gradient TD control~~



Deadly triad
- bootstrap (TD)
- off-policy
- approximation (semi-gradient)

Batch DQN deep Q-network

experience replay $\{(s, a, ?, s')\}$