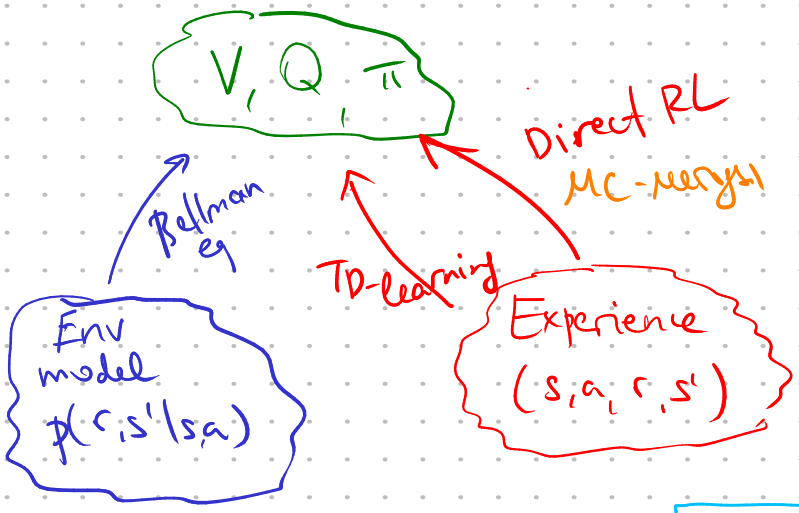
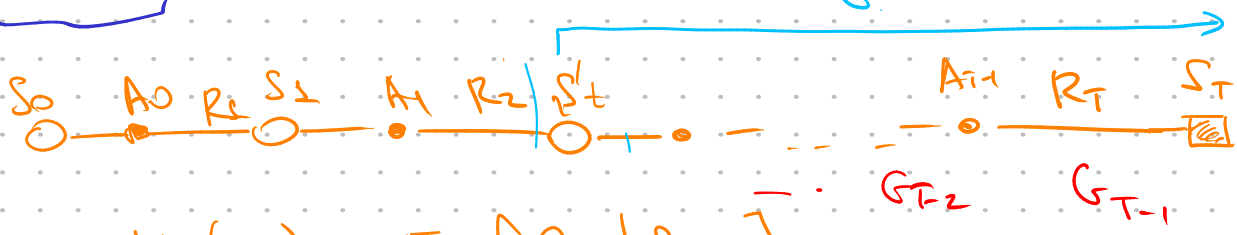
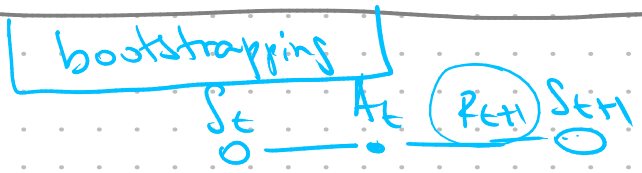




$$V_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s]$$

$V_{\pi'}$, $Q_{\pi'}$

Temporal difference learning



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

$$= \mathbb{E}_\pi[R_{t+1} + \gamma \cdot G_{t+1} \mid S_t = s]$$

$$G_t = R_{t+1} + \gamma G_{t+1} \approx R_{t+1} + \gamma V_\pi(S_{t+1})$$

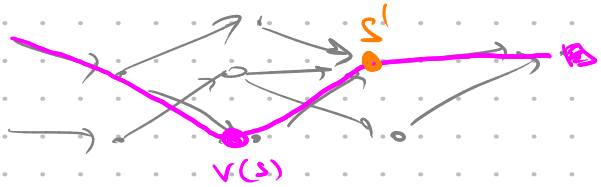
TD(0) for estimation

Update ← Step · (New − Old)

- init V_π
- repeat:
  - сделаем A_t в сост. S_t по π, наблюд. R_{t+1}, S_{t+1}
  - обновляем:

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

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



$V(S_1) = \frac{1}{2}$

$S_1 \to$ ▪ $\quad R = 1, 0, 1, 1, 0$

$S_0 \to S_1 \to$ ▪ $\quad R = 0$

$V^{MC}(S_0) = 0$

$V^{TD}(S_0) = 0 + \gamma \cdot V^{TD}(S_1) = \frac{1}{2}$ (γ = 1)

TD control

Sarsa (s, a, r, s', a')

On-policy TD control

- init
- repeat:
  - выбр. A_t для S_t по π, пол. R, S (будет из Q, напр. $\pi(A_t) = \begin{cases} 1-\varepsilon+\frac{\varepsilon}{|A|}, & \text{если это argmax } Q \\ \varepsilon/|A| \end{cases}$)
  - $Q(S_t, A_t) := Q(S_t, A_t) + \alpha\left(R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)\right)$

Expected Sarsa

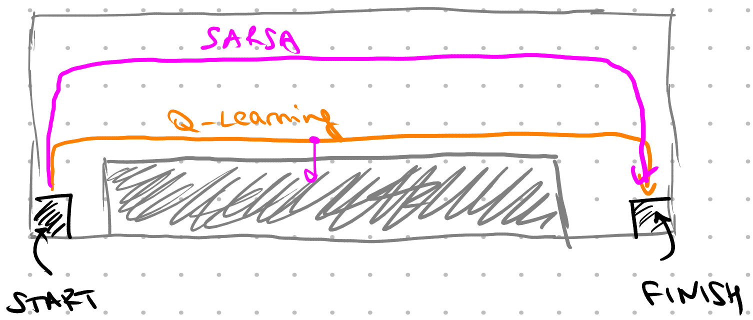
$\ldots + \gamma \cdot \mathbb{E}_{\pi(a|S_{t+1})}[Q(S_{t+1}, a)] = \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a)$

Off-policy TD control

Q-learning 1989

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

TD-Gammon

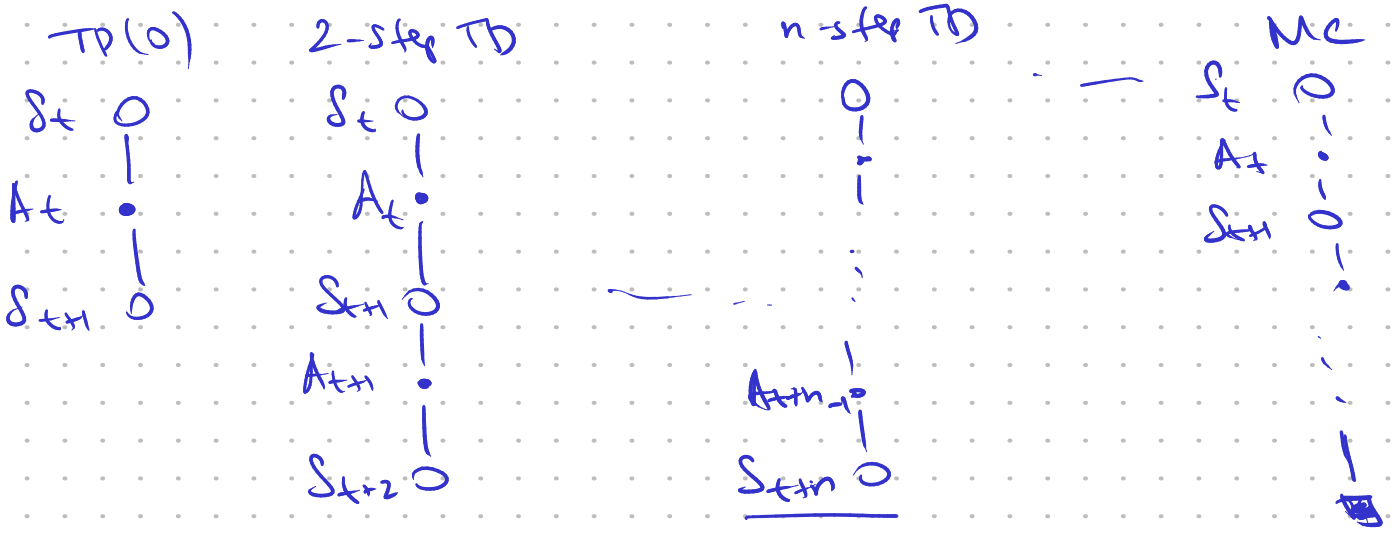


$G_{t:t+1} = R_{t+1} + \gamma V(S_{t+1})$

$G_{t:t+2} = R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2})$

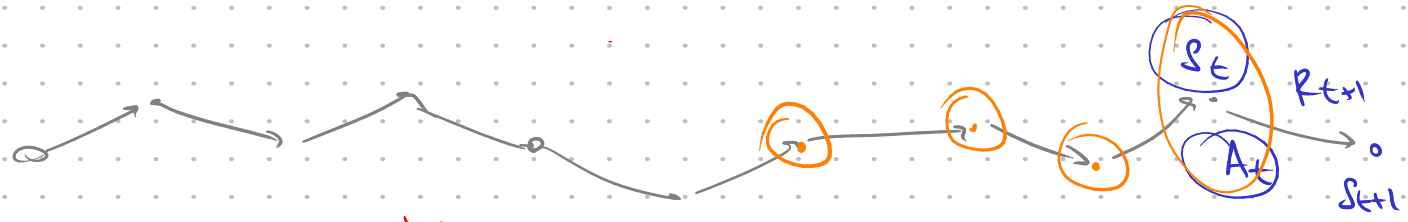
$G_{t:t+n} = R_{t+1} + \ldots + \gamma^{n-1} R_{t+n} + \gamma^n V(S_{t+n})$

$G_{t:\infty} = R_{t+1} + \ldots + \gamma^n R_{t+n-1} + \ldots = G_t$

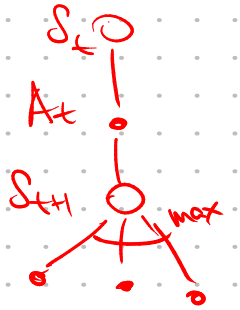


n-step SARSA:

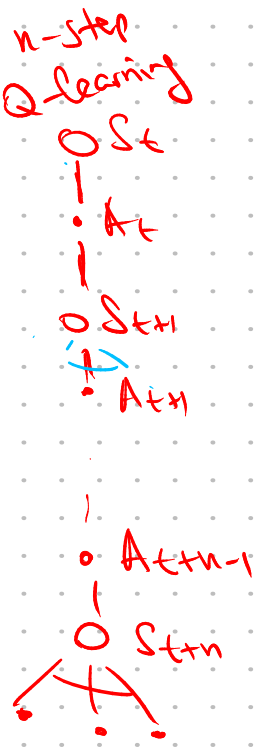
$$\alpha\left(R_{t+n} + \gamma R_{t+2} + \dots + \gamma^{n-1} R_{t+n} + Q(S_{t+n}, A_{t+n})\right)$$



Q-learning



n-step Q-learning



Exp. SARSA

n-step Exp. SARSA

Sutton Barto 1998/2018

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha \cdot \left(\hat{G} - Q(S_t, A_t)\right)$$

$$\hat{G} = R_{t+1} + \gamma \cdot \sum_{a \neq A_{t+1}} \pi(a|S_{t+1}) \cdot Q(S_{t+1}, a) +$$

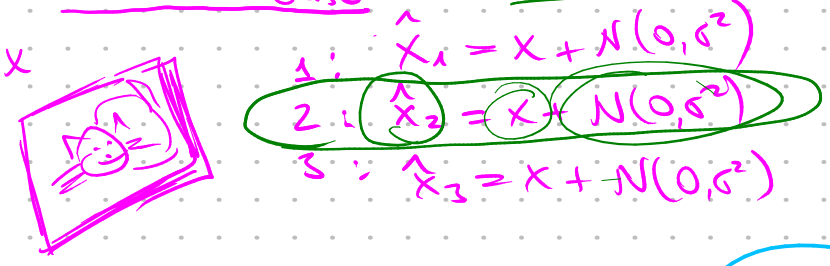
$$+ \gamma\, \pi(A_{t+1}|S_{t+1}) \cdot \left(R_{t+2} + \gamma \sum_{a \neq A_{t+2}} \pi(a|S_{t+2})\, Q(S_{t+2}, a) + \right.$$

$$\left. + \gamma\, \pi(A_{t+2}|S_{t+2}) \left(R_{t+3} + \gamma \cdot \max_a Q(S_{t+3}, a)\right)\right)$$

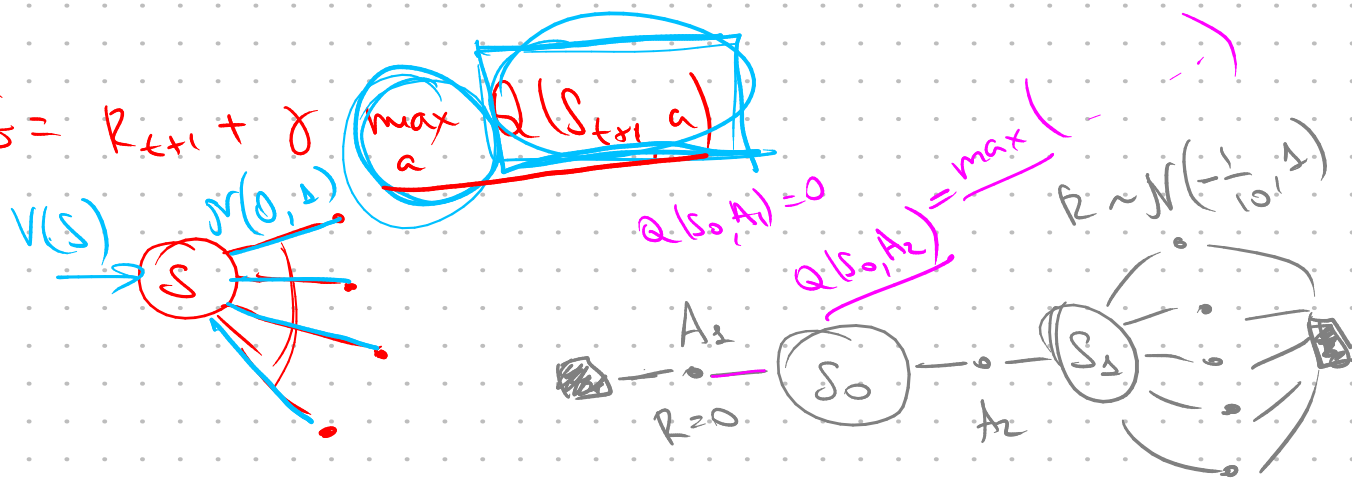
Double Q-learning

Winner's curse

Vickrey's auction



$$\hat{G} = R_{t+1} + \gamma \max_a Q(S_{t+1}, a)$$



$$\hat{G}_{\text{DOUBLE}} = R_{t+1} + \gamma Q_2\left(S_{t+1}, \operatorname*{argmax}_a Q_1(S_{t+1}, a)\right)$$

Double Q-learning:

- (s, a, r, s')
- Moneta:
  - $Q_1(s,a) := Q_1(s,a) + \alpha\left(r + \gamma Q_2\left(s', \operatorname*{argmax}_a Q_1(s',a)\right) - Q_1(s,a)\right)$

ИЛИ

- $Q_2(s,a) := \ldots$ — ½ —

Afterstates