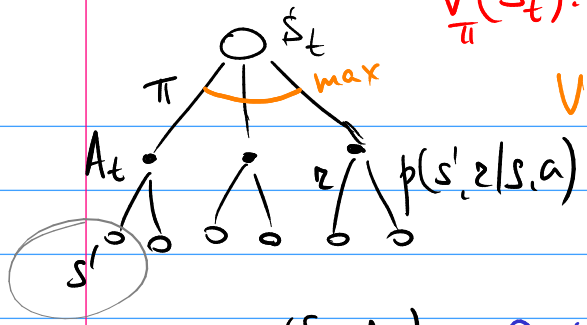
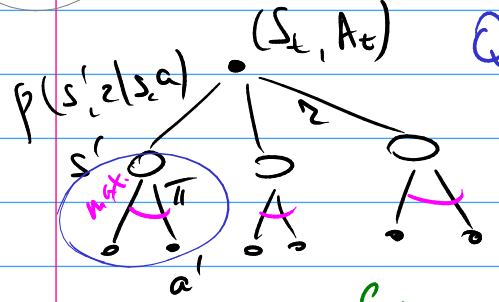
V



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

$$V_*(S_t) = \max_a \sum_{s',r} p(s',r|s,a)\left(r + \gamma V_*(s')\right)$$

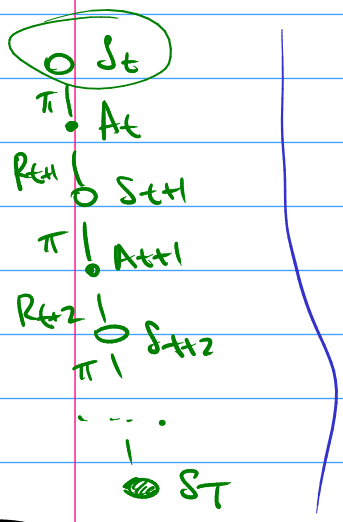


$$Q_\pi(S_t,A_t) = \sum_{s',r} p(s',r|s,a)\left(r + \sum_{a'} \pi(a'|s') Q_\pi(s',a')\right)$$

$$Q_*(S_t,A_t) = \sum_{s',r} p(s',r|s,a)\left(r + \max_{a'} Q_*(s',a')\right)$$

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

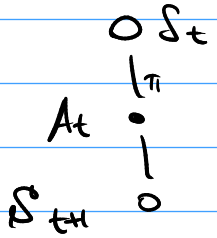
MC



TD

TD(0)

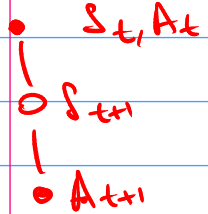
$$V_\pi(S_t) := V_\pi(S_t) + \alpha\left(R_{t+1} + \gamma V_\pi(S_{t+1}) - V(S_t)\right)$$



On-policy TD — Sarsa

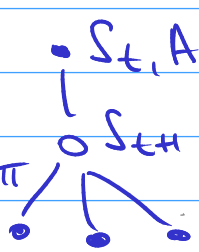
off-policy TD — Q-learning

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

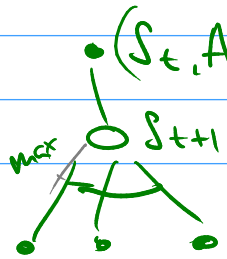


Expected Sarsa

$$Q := Q + \alpha\left(R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1},a) - Q\right)$$



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



n-step

TD: $\hat{G}_t = R_{t+1} + \gamma V(S_{t+1})$

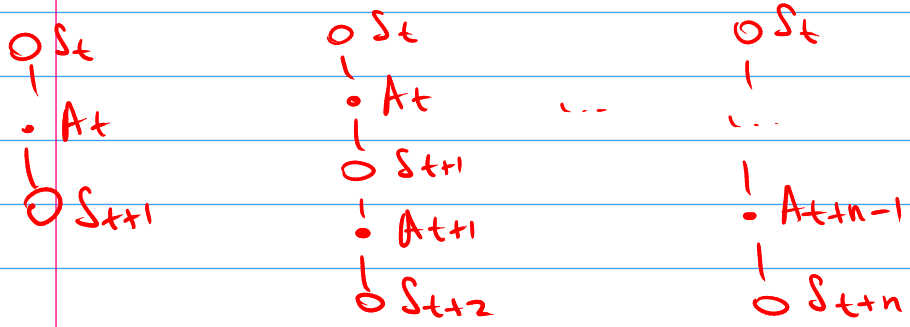
$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} = R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} =$

$= R_{t+1} + \gamma R_{t+2} + \gamma^2 \left(\sum_k \gamma^k R_{t+k+3} \right) = \dots$

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

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

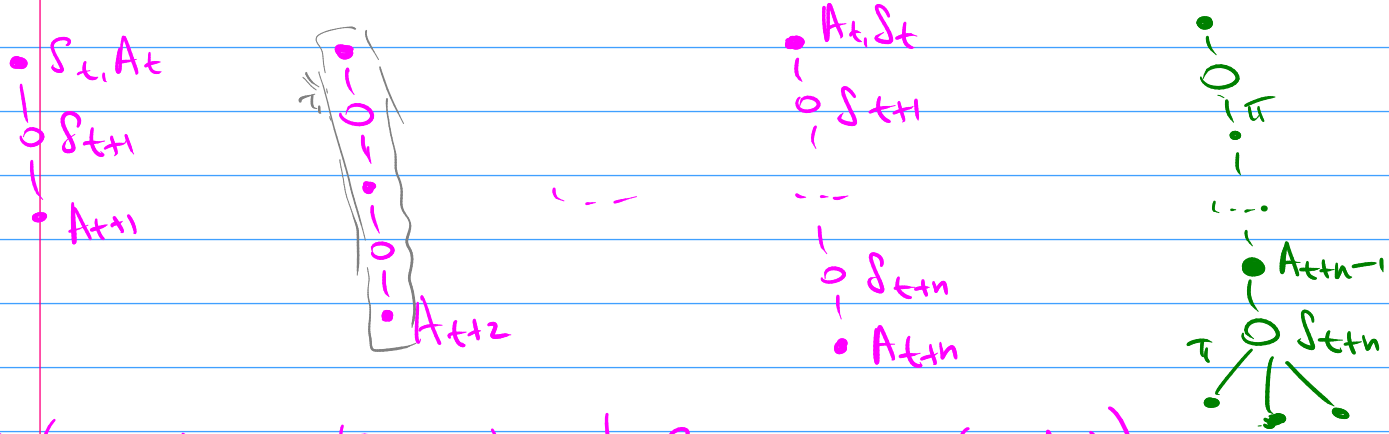
TD(0) | 2-step TD | n-step TD



$V_\pi(S_t) = V_\pi(S_t) + \alpha (G_{t:t+n} - V_\pi(S_t))$

n-step Sarsa:

(on-policy) expected Sarsa



$Q(S_t, A_t) = Q(S_t, A_t) + \alpha (G_{t:t+n} - Q(S_t, A_t))$

π - основная.

π' - поведенческая

$V(S_t) = V(S_t) + \alpha \rho_{t:t+n-1} (G_{t:t+n} - V(S_t))$

$\rho_{t:t+h} = \prod_{k=t}^{t+h} \frac{\pi(A_k|S_k)}{\pi'(A_k|S_k)}$

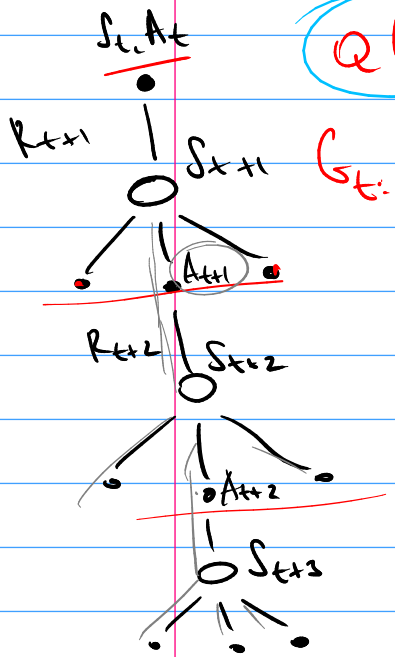
$$Q(S_t,A_t) = Q(S_t,A_t) + \alpha \rho_{t+1:t+n} \left[G_{t:t+n} - Q(S_t,A_t) \right]$$

n-step off-policy sarsa

regression

$$Q(S_t,A_t) = Q(S_t,A_t) + \alpha \left(G - Q(S_t,A_t) \right)$$

Tree backup



$$G_{t:t+3} = R_{t+1} + \gamma \sum_{a \neq A_{t+1}} \pi(a|S_{t+1}) Q(S_{t+1},a) +$$

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

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

TREE

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

$$= R_{t+1} + \gamma \underbrace{\sum_{a} \pi(a|\_)\ \_\_}_{\tilde{V}(S_{t+1})} + \gamma \pi(A_{t+1}|S_{t+1}) \left(G_{t+1:t+n} - Q(S_{t+1},A_{t+1}) \right)$$

$$= R_{t+1} + \gamma \tilde{V}(S_{t+1}) + \gamma \pi(A_{t+1}|S_{t+1}) \left(G_{t+1:t+n} - Q \right)$$

OFF

$$G_{t:t+n} = R_{t+1} + \gamma \rho_{t+1} \left(G_{t+1:t+n} - Q \right) + \gamma \tilde{V}(S_{t+1})$$

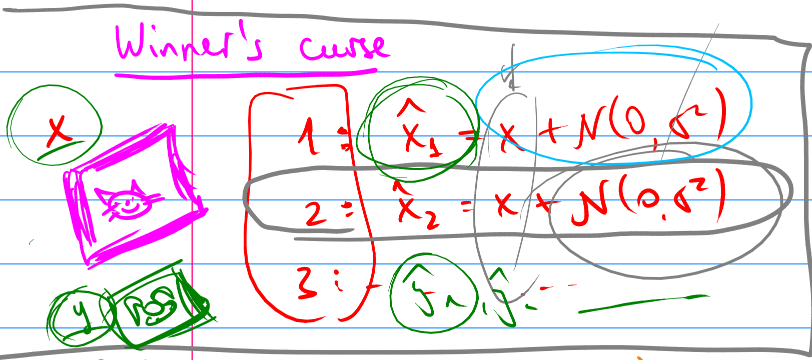
$Q(\sigma)$

σ_{t+1}

$$G^{\sigma} = R_{t+1} + \gamma \left(\sigma_{t+1} \rho_{t+1} + (1-\sigma_{t+1}) \pi(A_{t+1}|S_{t+1}) \right) \left(G_{t+1:t+n} - Q \right) +$$

$$+ \gamma \tilde{V}(S_{t+1})$$

Maximization bias, double learning



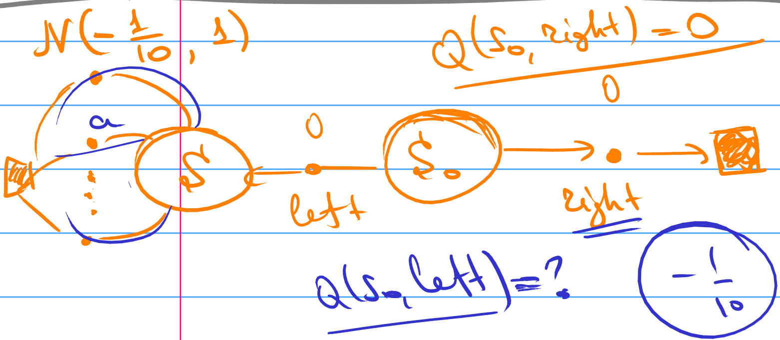
Q-learning

$\max_a Q(s,a)$

$\max_a Q(s,a)$

$\max_a [\hat{Q}(s,a)]$

Q_1 Q_2



$\max_a Q(s,a) > 0$

$a_* = \arg\max_a Q_1(s,a)$

$[Q_2(s, a_*)]$

Double Q-learning:

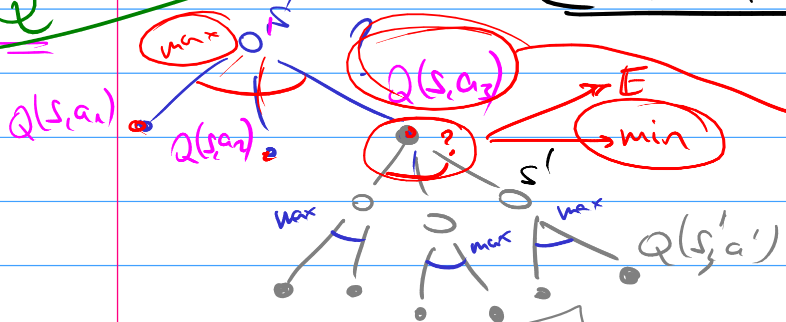
— " —

- Бросаем монетку

$$Q_1(S_t, A_t) := Q_1(S_t, A_t) + \alpha\Big(R_{t+1} + \gamma Q_2\big(S_{t+1}, \arg\max_a Q_1(S_{t+1}, a)\big) - Q_1(S_t, A_t)\Big)$$

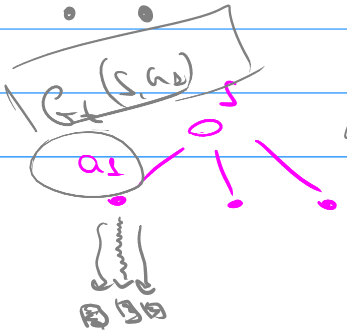
$Q = Q_*$

[Decision-time planning] ① Heuristic search



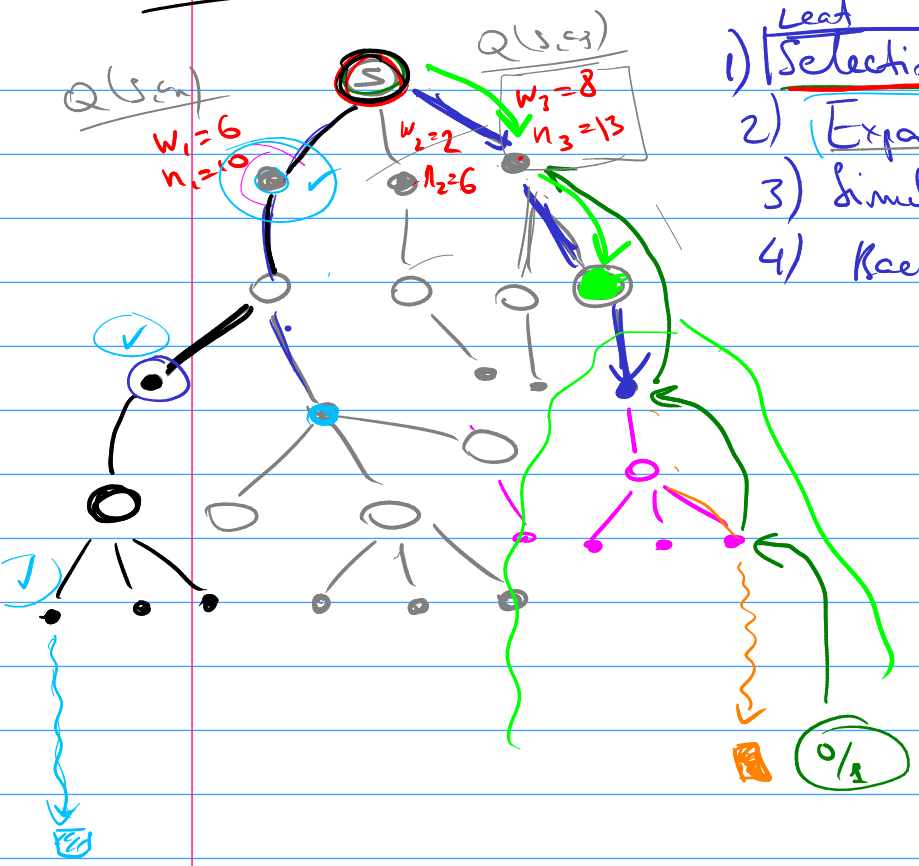
$\alpha Q(s,a_3) + (1-\alpha)\hat{Q}(s,a_3)$

② Rollouts

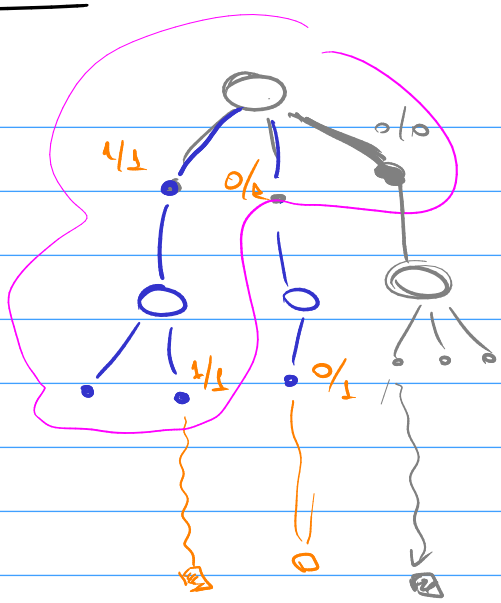


Тезауро — нарды

(3) MCTS – Monte Carlo Tree Search



1) Least Selection
2) Expansion
3) Simulation
4) Backup



UCT – upper conf. bound for trees

$$\text{argmax}\left[\frac{w_i}{n_i} + c\sqrt{\frac{\ln(\sum n_i)}{n_i}}\right]$$