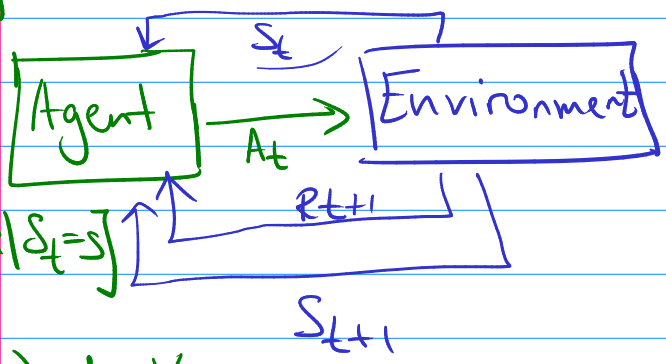
① MDP - Markov decision process

$\mathcal{S}$, $\mathcal{A}(s)$, $r \in \mathcal{R}$

strategy π

$\pi(a|s) = \Pr[A_t = a | S_t = s]$

$\sum_a \pi(a|s) = 1 \quad \forall s$



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

dynamics

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

$= \Pr[S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a]$

$\sum p(s', r | s, a) = 1 \quad \forall s, a$

$$p(s'|s,a) = \sum_r p(s', r | s, a)$$

$$r(s,a) = \mathbb{E}[R_t | S_{t-1} = s, A_{t-1} = a] =$$

$$= \sum_r r \cdot \sum_{s'} p(s', r | s, a)$$

$\sum_{s'} p(s', r | s, a) = p(r | s, a)$

$|S| = 1$ - bandity

$|S| > 1$ + kontext c

$r(a|c)$

---

② Returns

t: $R_{t+1}, R_{t+2}, R_{t+3}, \ldots$

return

Epizodic - T

Continuous

discount

$G_t = R_{t+1} + R_{t+2} + \ldots + R_T$

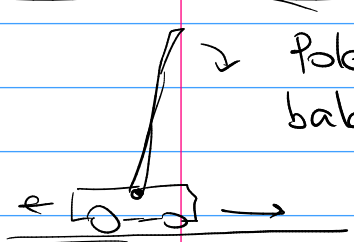
$G_t = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^k R_{t+k+1} + \ldots$

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

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

M.S. $T = \infty$

M.S. $\gamma = 1$



$\gamma = 1$

$\gamma < 1$

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

$= R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \ldots)$

$$G_t = R_{t+1} + \gamma \cdot G_{t+1}$$

③ Value functions

Agent $\pi(a|s) = \Pr[A_t = a | S_t = s]$

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

$$= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right]$$

$$Q_\pi(s,a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum \gamma^k R_{t+k+1} \,\middle|\, \begin{matrix} S_t = s \\ A_t = a \end{matrix}\right]$$

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

$$Q_\pi(s,a) = \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] =$$

$$= \underbrace{\mathbb{E}_\pi[R_{t+1} | S_t = s, A_t = a]}_{r(s,a)} + \gamma \mathbb{E}_\pi[G_{t+1} | S_t = s, A_t = a]$$

$S_{t+1} = s'$

$$Q_\pi(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) \cdot V_\pi(s')$$

$p(s'|s,a)$

④ Bellman equations

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

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

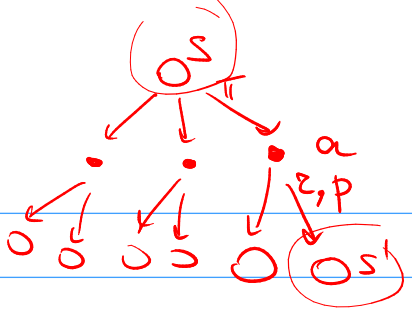
$$= \sum_a \pi(a|s) \cdot \sum_{s'} \sum_r p(s', r | s, a)\left[r + \gamma \mathbb{E}_\pi[G_{t+1} | S_{t+1} = s']\right]$$

$$\forall s \quad V_\pi(s) = \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r | s, a)\left[r + \gamma V_\pi(s')\right]$$

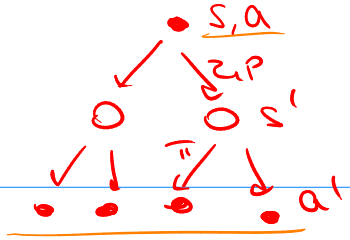
Bellman equations

$$\forall s, a \quad Q_\pi(s,a) = \sum_{s'} \sum_r p(s', r | s, a)\left[r + \gamma \cdot \sum_{a'} \pi(a'|s') Q_\pi(s', a')\right]$$

backup diagram



$V_\pi(s)$



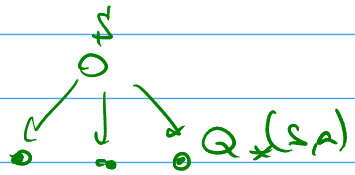
$Q_\pi(s,a)$

(5) Optimal strategies

π_* - оптим. страт., т.е. $\forall \pi \; \forall s \quad V_{\pi_*}(s) \geq V_\pi(s)$

$\pi : s \to \pi(a|s)$

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



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

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

$$V_*(s) = \max_\pi V_\pi(s) = \max_\pi \left[\sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[\dots] \right] =$$

$$a = \arg\max_a \max_\pi \sum_{s',r} p(\dots)\left[r + \gamma V_\pi(s')\right] =$$

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

$\forall s$

$$V_*(s) = \max_\pi V_\pi(s) = \max_a \left(\sum_{s',r} p(s',r|s,a)\left[r + \gamma \cdot V_*(s')\right] \right)$$

Bellman equations

$$Q_*(s,a) = \max_\pi Q_\pi(s,a) = \sum_{s',r} p(s',r|s,a)\left[r + \gamma \max_\pi V_\pi(s')\right]$$

$\max_{a'} \max_\pi Q_\pi(s',a')$

$$Q_*(s,a) = \sum_{s',r} p(s',r|s,a)\left[r + \gamma \cdot \max_{a'} Q_*(s',a')\right]$$

$\bar{x} = f(\bar{x})$ $\quad \bar{x}^0, \bar{x}^1 = f(\bar{x}^0), \bar{x}^2 = f(\bar{x}^1), \ldots$

$\bar{x}^{k+1} = f(\bar{x}^{(k)})$

$\bar{x} = A\bar{x} - (\lambda - 1)\bar{x}$

Policy evaluation π: Eval $\quad V_\pi^{(k+1)}(s) := \text{Bellman}\left(V_\pi^{(k)}(s)\right)$

π_* $\quad V_*^{(k+1)}(s) = \max$ ——

---

(6) Policy improvement $\quad \pi \xrightarrow{\text{Impr.}} V_\pi(s),\ Q_\pi(s,a)$

Policy impr. theorem

пусть π, π' - детерм. ст., и $\forall s \quad Q_\pi(s, \pi'(s)) \geq V_\pi(s) \Rightarrow$ тогда $\forall s \quad V_{\pi'}(s) \geq V_\pi(s)$

1-ый: $V_\pi(s) \leq Q_\pi(s, \pi'(s)) =$

$= \mathbb{E}\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s, A_t = \pi'(s)\right] =$

$= \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s\right] \leq$

$\leq \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma Q_\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s\right] = \ldots$

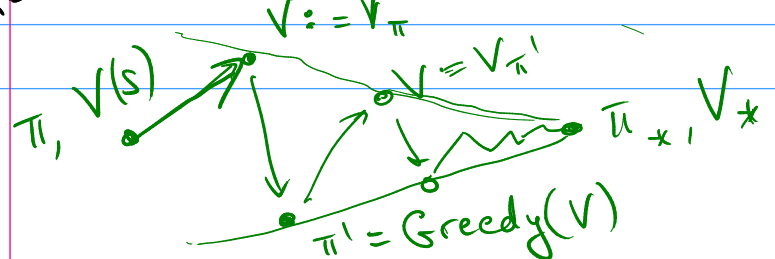
$\ldots \leq \ldots \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 Q_\pi(S_{t+2}, \pi'(S_{t+2})) \mid S_t = s\right]$

$\leq \ldots \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \ldots \mid S_t = s\right] = V_{\pi'}(s)$

$\pi'(s) = \arg\max_a Q_\pi(s,a) = \arg\max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma V_\pi(s')\right]$

$Q_\pi(s, \pi'(s)) \geq Q_\pi(s, \pi(s)) = V_\pi(s)$

Policy iteration



$\pi' = \pi \Rightarrow \pi' = \pi = \pi_*$

$\pi(s) = \arg\max_a Q_\pi(s,a)$

DP dynamic programming

Value iteration $V^{(k+1)}(s) := \max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma \cdot V^{(k)}(s')\right]$

$V_\pi(s)$ $p(\_,\_|\_,\_)$

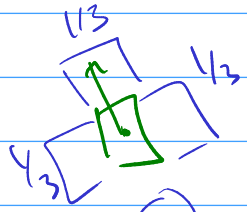
- $V(s) \leftarrow$ random
- Loop:
  - Играю по π, $S_0, A_0, R_1, S_1, A_1, \ldots, R_T$
  - $G \leftarrow 0$
  - for $t = T-1, T-2, \ldots, 0$:
    - $G \leftarrow \gamma G + R_{t+1}$
    - сохраняю G в список $Returns(S_t)$
    - $V(S_t) := Avg(Returns(S_t))$

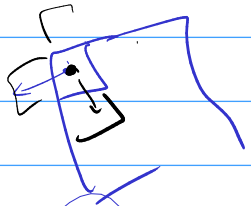
first-visit MC
every-visit MC

- $V_\pi(s) := V(s)$

$\pi \to \pi'$

$G = R_1 + \gamma R_2 + \gamma^2 R_3 + \ldots + \gamma^{T-1} R_T$





$G_T = 0$, $G_{T-1} = R_T$, $G_{T-2} = R_{T-1} + \gamma R_T$