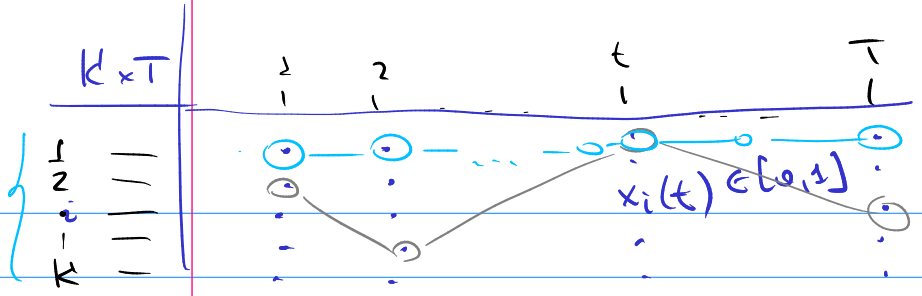
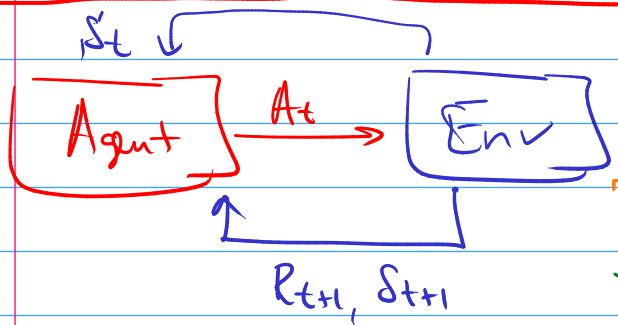



$$\mathbb{E}_{x \sim p(x)}[f(x)] = ? \qquad x \sim q(x)$$

$$\mathbb{E}_{x \sim q(x)}[g(x)] = \int g(x) q(x) dx$$

$$g(x) = \frac{f(x)\, p(x)}{q(x)}$$

---




$$p(r, s' \mid s, a)$$

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots$$

$\pi$ $\quad \pi(A_1 \mid S_1) \quad \pi$

**Return**

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

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

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

$$R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \ldots$$

**State value function**

$$V_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\right]$$

$A_t, R_{t+1}, S_{t+1}, A_{t+1}, \ldots$

**Action value function** Q-function

$$Q_\pi(s,a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a\right]$$

$R_{t+1}, S_{t+1}, A_{t+1}, R_{t+2}, \ldots$

$$Q_\pi(s,a) = \mathbb{E}_\pi\left[R_{t+1} + \gamma \cdot \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \mid S_t = s, A_t = a\right] =$$

$$\mathbb{E}_{R_{t+1}, S_{t+1}}\left[\mathbb{E}_\pi[\ \ldots\ ]\right]$$

$A_{t+1}, R_{t+2}, S_{t+2}, \ldots$

$$= \sum_{r,s'} p(r,s'|s,a)\left(r + \mathbb{E}_\pi\left[\gamma \cdot \sum \gamma^k R_{t+k+2} \mid S_t = s, A_t = a, S_{t+1} = s'\right]\right)$$

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

$$V_\pi(s) = \mathbb{E}_\pi\left[\sum \gamma^k R_{t+k+1} \mid S_t = s\right] = \mathbb{E}_{a\sim\pi}\left[\mathbb{E}_\pi\left[G_t \mid S_t = s, A_t = a\right]\right]$$

$A_t, R_{t+1}, S_{t+1}, A_{t+1} \ldots$

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

---

$\pi' \geq \pi \iff \forall s \quad V_{\pi'}(s) \geq V_\pi(s)$

---

$\pi_* = ?$ $\qquad V_*(s) = \max_\pi V_\pi(s)$

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

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

$$Q_{\pi_*}(s,a) = \mathbb{E}_{\pi_*}\left[G_t \mid S_t = s, A_t = a\right] =$$

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

$$V_*(s) = \max_\pi V_\pi(s) = \max_a \left(\max_\pi \sum \pi(a|s) \cdot Q(s,a)\right) =$$

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

$$V_\pi(s) = \mathbb{E}_\pi\left[G_t \mid S_t = s\right] = \mathbb{E}_\pi\left[R_{t+1} + \gamma \cdot G_{t+1} \mid S_t = s\right] =$$

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

$$= \mathbb{E}_{A_t, R_{t+1}, S_{t+1}} \left[ R_{t+1} + \gamma \underbrace{\mathbb{E}_\pi [G_{t+1} | S_{t+1}]}_{= V_\pi(S_{t+1})} \right]$$

$$\boxed{V_\pi(s) = \sum_a \pi(a|s) \sum_{r,s'} p(r,s'|s,a) \cdot \left( r + \gamma V_\pi(s') \right)}$$ Bellman equations

Сжим. ур. на $V_\pi(s)$, $s \in \mathcal{S}$ $\quad \bar{x} = f(\bar{x})$

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

$$= \mathbb{E}_{R_{t+1}, S_{t+1}} \left[ R_{t+1} + \gamma \mathbb{E}_\pi [G_{t+1} | S_{t+1}] \right] =$$

$$= \sum_{r,s'} p(r,s'|s,a) \left( r + \gamma \cdot \underbrace{V_\pi(s')}_{= \mathbb{E}_{A_{t+1}} [Q_\pi(s,a)]} \right)$$

$$\boxed{Q_\pi(s,a) = \sum_{r,s'} p(r,s'|s,a) \left( r + \gamma \cdot \sum_{a'} \pi(a'|s') Q_\pi(s',a') \right)}$$ Bellman equations

1. Перебирать $\pi$ нереально $\quad \pi : \mathcal{S} \to \mathcal{A} \quad |\mathcal{S}|^{|\mathcal{A}|}$
2. Мы редко знаем $p(r,s'|s,a)$
3. Всё равно уравнений слишком много — их $|\mathcal{S}|$ или $|\mathcal{S}| \cdot |\mathcal{A}|$

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

$$= \max_a \sum \left( \quad \cdot \quad \gamma \max_\pi V_\pi(s') \quad \right)$$

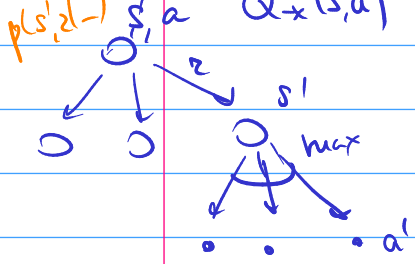
$$\boxed{V_*(s) = \max_a \sum_{r,s'} p(r,s'|s,a) \left( r + \gamma V_*(s') \right)}$$ Bellman equations

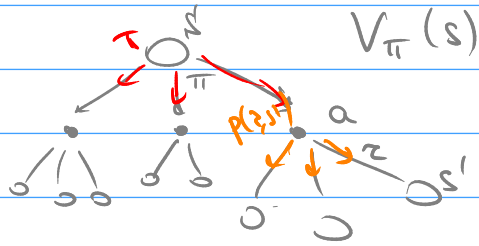
$$Q_*(s,a) = \max_\pi Q_\pi(s,a) = \max_\pi \sum_{r,s'} p(r,s'|s,a)\left(r + \gamma \sum_{a'} \pi(a'|s') Q_\pi(s',a')\right)$$

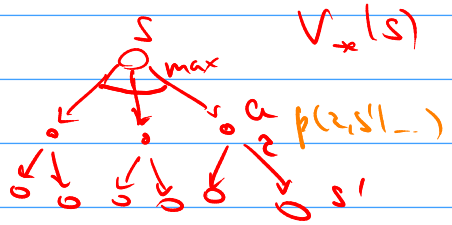
$$\boxed{Q_*(s,a) = \sum_{r,s'} p(r,s'|s,a)\left(r + \gamma \max_{a'} Q_*(s',a')\right)}$$

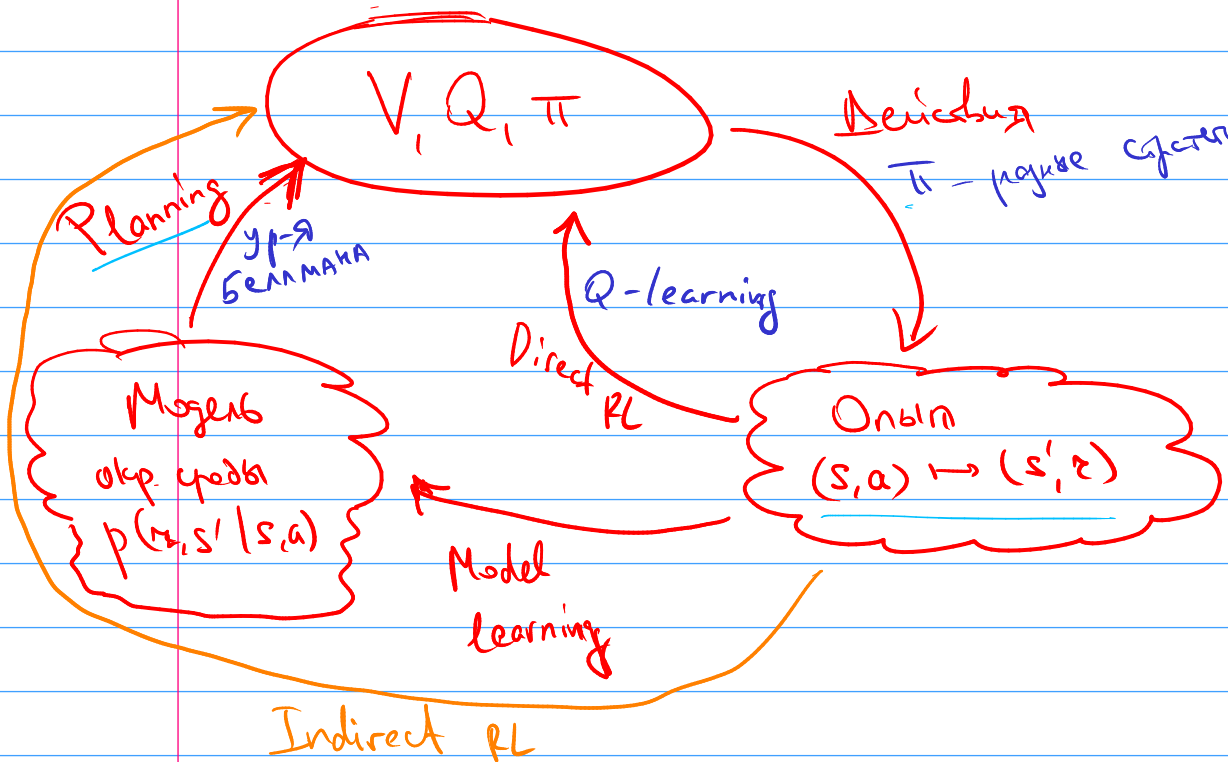
$V^{(0)}(s)$ $\quad V^{(k+1)}(s) := \max_{a}\left( \quad -- \quad V^{(k)}(s')\right)$

backup diagram




Policy improvement $\quad \pi \rightsquigarrow \pi'$ ?

$\pi \rightsquigarrow V_\pi(s)$, $Q_\pi(s,a)$

**Policy impr. thm.** Если для двух стратегий $\pi, \pi'$

$$\forall s \in S \quad Q_\pi(s, \pi'(s)) \geq V_\pi(s) = Q_\pi(s, \pi(s))$$

то $\forall s \in S \quad V_{\pi'}(s) \geq V_\pi(s)$

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

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

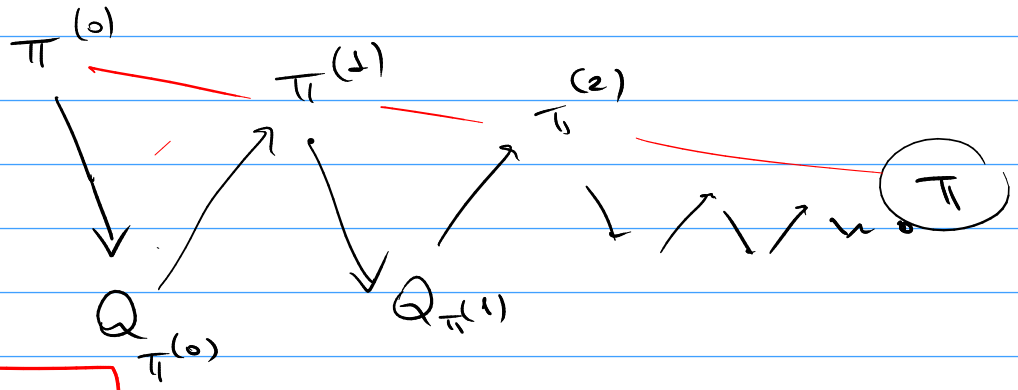
$$\leq \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma Q_\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s\right] =$$

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

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

$\pi \longrightarrow V_\pi, Q_\pi$ $\qquad \pi'(s) := \underset{a}{\operatorname{argmax}}\, Q_\pi(s, a)$

**Policy Iteration**




$$\pi(s) = \underset{a}{\operatorname{argmax}}\, Q_\pi(s, a)$$