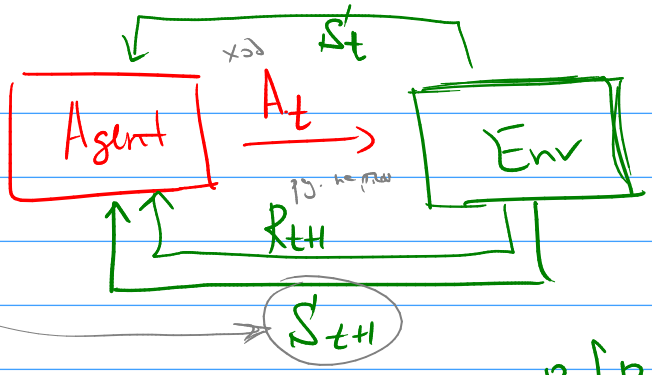
MDP - Markov Decision Process

$S \rightsquigarrow A$

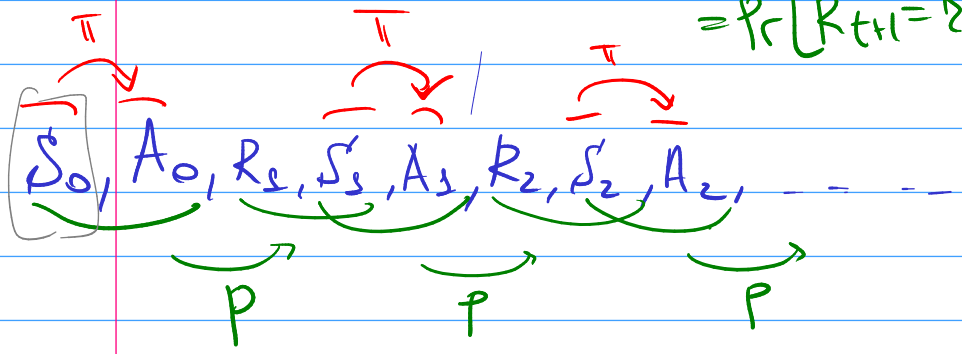


Стратегия:

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

Dynamics:

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



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

Episodic tasks

$S_0, A_0, R_1, \ldots, R_T, S_T$

$G = R_1 + R_2 + \ldots + R_T \rightarrow \max$

Continuous tasks

$S_0, A_0, R_1, \ldots, S_t, A_t, \ldots$

$G = R_1 + \gamma R_2 + \gamma^2 R_3 + \ldots$

$$= \sum_{k=1}^{\infty} \gamma^{k-1} R_k, \quad \gamma < 1$$

$\gamma < 1$ $\rightarrow \max$

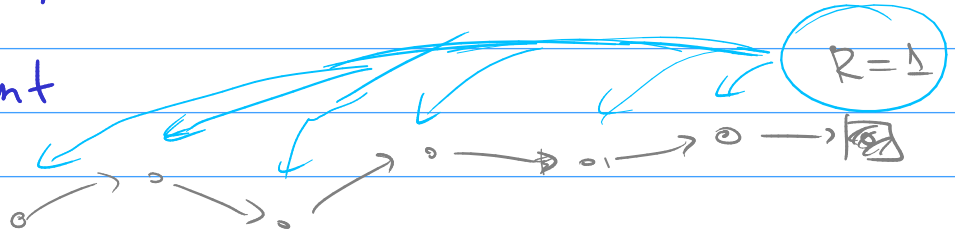
$\gamma = 1$

$$G = \sum_{k=1}^{\infty} \gamma^{k-1} R_k$$

$$\mathbb{E}_{\pi, p}\left[\sum_{k=1}^{\infty} \gamma^{k-1} R_k\right] \xrightarrow[\pi]{} \max$$

1) Exploration vs. exploitation

2) Credit assignment



Multiarmed bandits

$|\mathcal{A}| = k$

$a_1, a_2, \ldots, a_n$

$\pi(a) = Pr[A_t = a]$

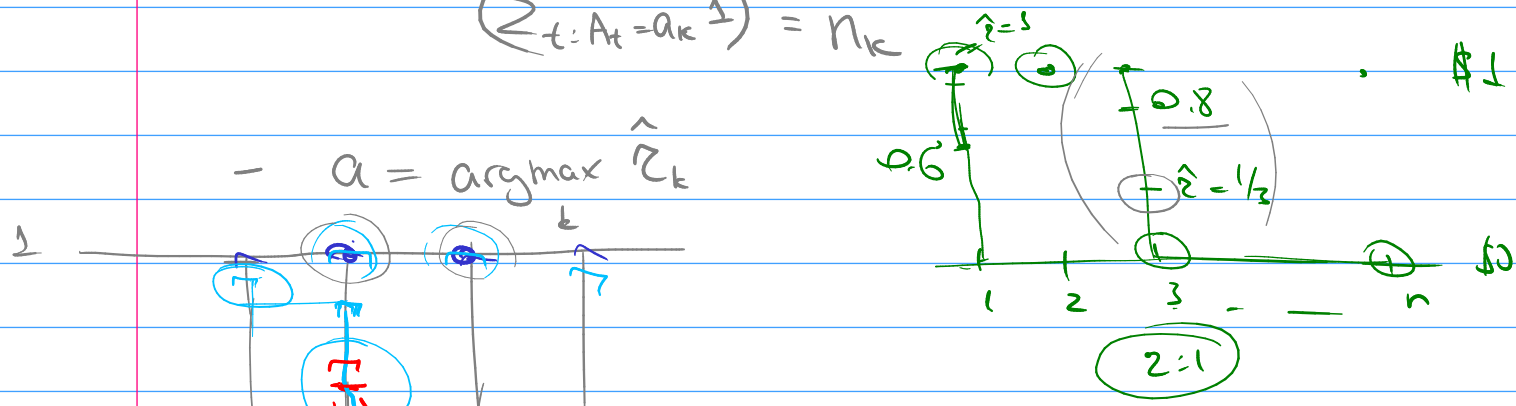
$r = \$1 / 10$

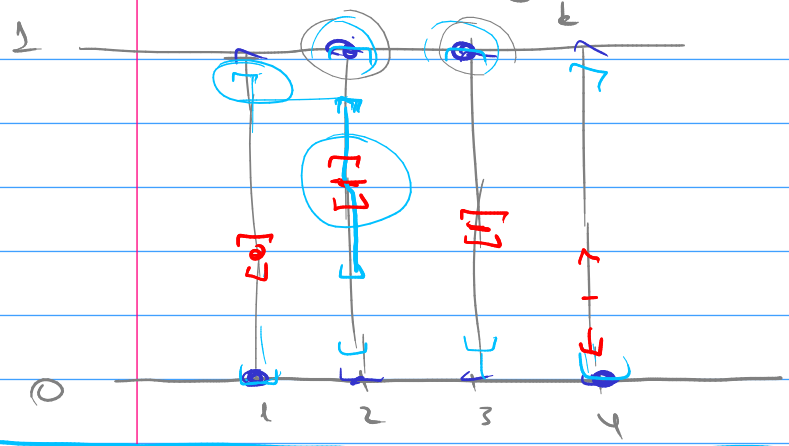
$r_k = Pr[R = 1 | A = a_k]$

Greedy:

- $\hat{r}_k = \dfrac{\sum_{t: A_t = a_k} r_t}{\left(\sum_{t: A_t = a_k} 1\right)} = n_k$

- $a = \operatorname{argmax}_k \hat{r}_k$





$$\hat{r}_k = \frac{\sum r_t}{n_k} = \frac{r_1 + r_2 + \ldots + r_{n_k}}{n_k}$$

! r

$$\hat{r}_k' = \frac{(r_1 + \ldots + r_{n_k}) + r}{n_k + 1} = \frac{1}{n_k + 1}\left(\left(\hat{r}_k n_k + r\right) - \hat{r}_k + \frac{1}{n_k + 1}\left(r - \hat{r}_k\right)\right)$$

$$\boxed{\hat{r}_k' = \hat{r}_k + \frac{1}{n_k + 1}\left(r - \hat{r}_k\right)}$$

Новая оценка = Старая оценка + Коэф. обуч. (Цель − Стар. оценка)

$\sum_{t=1}^{\infty} h_t = \infty$

$\sum h_t^2 < \infty$

Коэф. обуч. $\sim \frac{1}{t} \Rightarrow \hat{r}$

$$\hat{a}_{t+1} = \hat{a}_t + h \cdot (r_t - \hat{a}_t) =$$

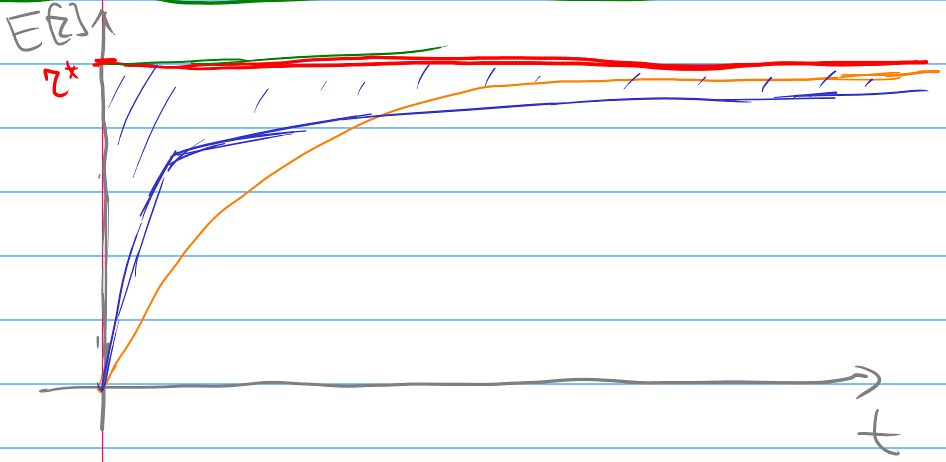
Nonstationary bandits

$$= h r_t + (1-h)\hat{a}_t =$$

$$= h r_t + (1-h)\left(h r_{t-1} + (1-h)\hat{a}_{t-1}\right) = \_\_$$

$$= h r_t + h(1-h) r_{t-1} + h(1-h)^2 r_{t-2} + \dots$$

$$= \sum_{k=t}^{0} h(1-h)^{k-t} r_k$$



Regret

$$r^* - \mathbb{E}_\pi[r]$$

$$\sum_{i=1}^{K} n_i$$

UCB1

upper confidence bound

$$\text{Priority}_k = \hat{r}_k + c\sqrt{\frac{\log n}{n_k}}$$

c " $\sqrt{2}$ "

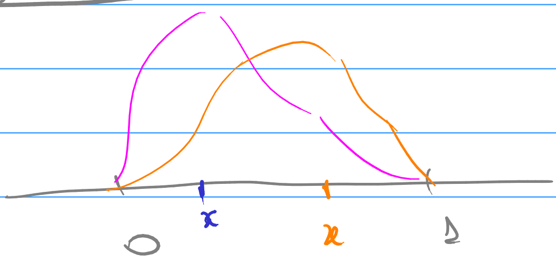
n_k ∈ [illegible] је k

$T \quad O(\log T)$ log добијаш је губит. функц

$$\text{regret} \le O\left(\sqrt{K \cdot T \log T}\right)$$

Thompson sampling

$$\pi(a) = p(a = a^* \mid D)$$



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

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

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

return

Value function

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

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

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

$$V_\pi(s) = \mathbb{E}_{a\sim\pi}[Q_\pi(s,a)] = \sum_a \pi(a|s)\, Q_\pi(s,a)$$

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

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

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

Bellman equations

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

$\pi(a|s)$

$\bar{x} = f(\bar{x})$

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

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

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

Blackjack

$S = ?$ ($\Sigma_{...}$, usable Ace, карта дилера)

$\pi: s \to a$

Hit / Stand

1

(Hand, Ace, Dealer card, ...)

Score

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

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

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

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

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

$\bar{x} = f(\bar{x})$

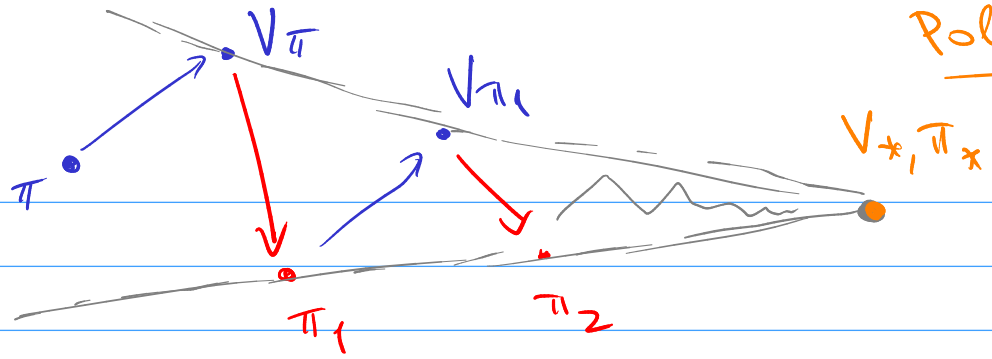
$$\pi^*(a|s) = \begin{cases} 1, & a = \arg\max_a Q_*(s,a) \\ 0, & - \text{---} \end{cases}$$

Policy improvement

$\pi \Longrightarrow V_\pi, Q_\pi$

~~V_*~~

Policy iteration



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

Thm. (Policy improvement thm) Пусть π, π' – дет. стр., и

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

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

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

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

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

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

$$= \dots \leq \dots \qquad \mathbb{E}_{\pi'}\left[G_t \mid S_t = s\right] = V_{\pi'}(s)$$

$$\pi^{(k+1)}(s) = \arg\max_a Q_{\pi^{(k)}}(s, a)$$

Если $\exists \pi'$: $\forall s \quad V_{\pi'}(s) \geq V_\pi(s)$

$\exists s_0 \quad V_{\pi'}(s_0) > V_\pi(s_0)$

$$\Rightarrow Q_\pi(s_0, \pi'(s_0)) > V_\pi(s_0)$$