MDP
Markov decision process

$\pi: S \to \text{Prob}(A)$

Agent

$\pi(s_t, a_t)$

Environment

t: s_t $\quad a_t \in A_t \longrightarrow$

r_t - reward

$t+1$: s_{t+1}

$r_t \sim p(r_t | s_t, a_t)$

$s_{t+1} \sim p(s_{t+1} | s_t, a_t)$

Episodic tasks

$1, \dots, T$

$$\sum_{t=1}^{T} r_t \to \max$$

Continuous tasks

$$\sum_{t=1}^{\infty} r_t \to \max$$

$$\sum_{t=1}^{\infty} r_t \cdot \gamma^t \to \max$$

$D = \{(\bar{x}_n, y_n)\}_{n=1}^{N}$

$p(y|\bar{x})$

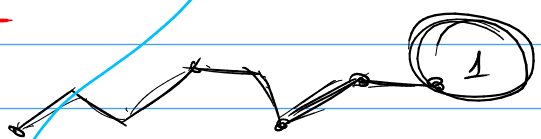
$D = \{\bar{x}_n\}_{n=1}^{N}$

$p(\bar{x})$

$p(\bar{x} | \bar{\alpha}, \bar{\mu}_1, \dots \bar{\mu}_k, \Sigma_1 \dots \Sigma_k)$

1) Exploration vs. exploitation

2) Credit assignment



Reward - $r_t \in \{-1, 0, 1\}$

$V(s) \approx \mathbb{E}[R \mid \text{начинаем в позиции } s]$

$Q(s,a) \approx \mathbb{E}[R \mid \text{начинаем в } s \text{ и делаем действие } a]$

$$a^* = \arg\max_a Q(s,a)$$

Multiarmed bandits

t, $a_i \rightsquigarrow r_t(i) \sim p_i(r)$, $\mathbb{E}_{p_i} r = \boxed{R_i}$

$i^* = \underset{i}{\operatorname{argmax}} R_i$

$\hat{R}_i = \frac{1}{n_i} \sum_{t=1}^{n_i} r_t$

Greedy: — Дёрнуть 1 раз за каждую ручку (0/1)

— $\forall t = 1, \dots, T$:

$I_t = \underset{i}{\operatorname{argmax}} \hat{R}_i$

$\hat{R}_{it} = \frac{1}{n_i} \sum r_t$

$r_i \in [0, 1]$

— «Выр. эксп.» — $n_i = 1$, $\hat{R}_i = 5$

— Greedy

---

t_i — время, когда зёрнышки a

t: $Q_t(a) = \frac{\sum_{i=1}^{n_a} r_{t_i}}{n_a}$

r_{t+1}

$Q_t(a) = \frac{1}{n_a}(r_{t_1} + \dots + r_{t n_a})$

$Q_{t+1}(a) = \frac{1}{n_a+1}\left(\underbrace{(r_{t_1} + \dots + r_{t n_a})}_{= n_a \cdot Q_t(a)} + r_{t+1}\right) =$

$= \frac{r_{t_1} + \dots + r_{t n_a}}{n_a+1} + \frac{r_{t+1}}{n_a+1} =$

$1 - \frac{1}{n_a+1}$

$= \frac{n_a}{n_a+1} Q_t(a) + \frac{1}{n_a+1} r_{t+1}$

α_t

$Q_{t+1}(a) = Q_t(a) + \frac{1}{n_a+1}\left[r_{t+1} - Q_t(a)\right]$

Новая оценка = Старая оценка + Шаг · [r − Старая оценка]

$\|F(x^*) - F(x_t)\| \le \frac{R^2 + G^2 \cdot \sum \alpha_t^2 \not\to \infty}{2 \cdot \sum \alpha_t \to \infty}$

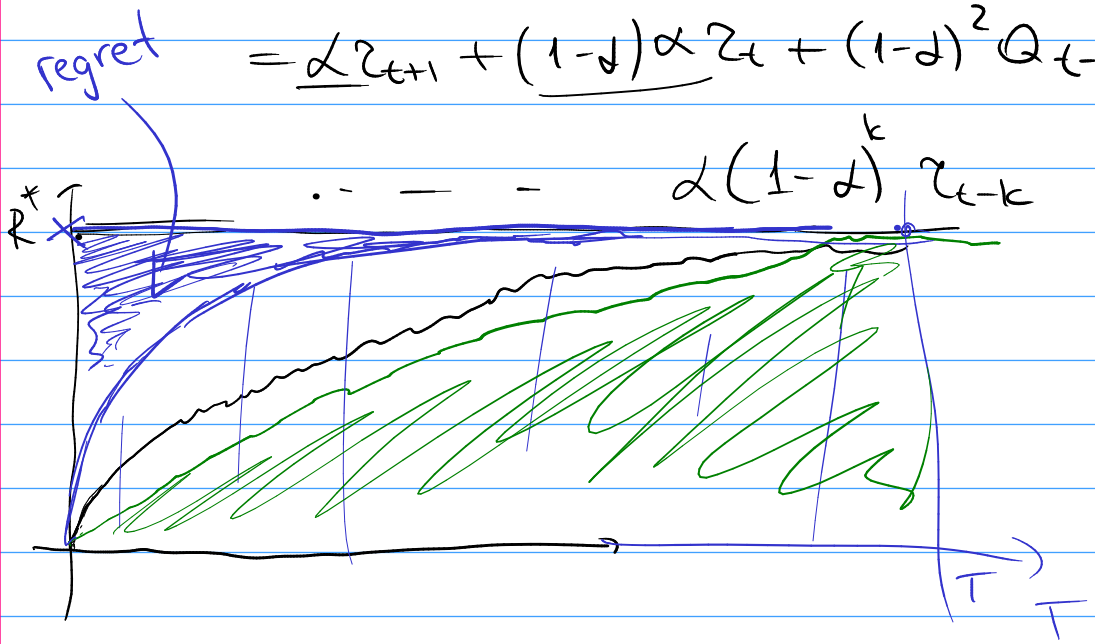
$$\sum_{t=1}^{\infty} \alpha_t = \infty \qquad \sum \alpha_t^2 < \infty$$

$$Q_{t+1}(a) = Q_t(a) + \alpha \cdot \left[r_{t+1} - Q_t(a) \right] =$$

$$= \alpha r_{t+1} + (1-\alpha) Q_t(a) =$$

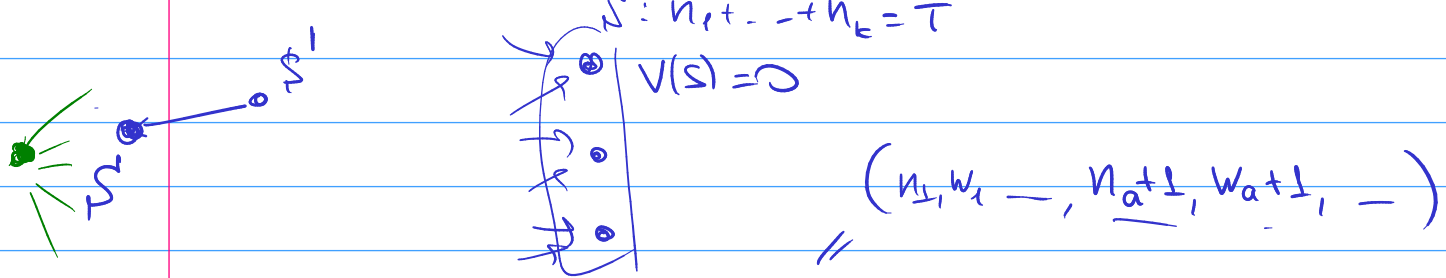
$$= \alpha r_{t+1} + (1-\alpha)\alpha r_t + (1-\alpha)^2 Q_{t-1}(a)$$

$$\ldots - - - \alpha(1-\alpha)^k r_{t-k}$$



$$S = \{ n_1, w_1, n_2, w_2, \ldots, n_k, w_k \} \qquad n = n_1 + \ldots + n_k$$

$V(S) = \mathbb{E}[\text{награда до конца } T]$



$$(n_1, w_1 -, n_a+1, w_a+1, -) \quad // \quad (n_1, w_1 -, n_a+1, w_a, \ldots)$$

$t=0$

$$Q(S,a) = p_a \cdot (1 + V(S')) + (1-p_a) V(S'')$$

$$V(S) = \max_a Q(S,a)$$

$$p_a = \frac{w_a+1}{n_a+2}$$

Gittins indices

Thompson sampling

$r_t \in \{0,1\}$, $\theta_i = p(r_t = 1 | I_t = i)$

t: $i \to D_{i,t}$ $\{n_{i,t}, w_{i,t}\}$

$p(\theta_i) = B(\theta_i | 1,1)$

$p(\theta_i | D_{i,t}) =$

$$p(\theta_i | w_{i,t}, n_{i,t}) = B(\theta_i | w_{i,t}+1, n_{i,t}-w_{i,t}+1)$$

- $t = 1, \dots, T$:
  - $\theta_{i,t} \sim p(\theta_i | w_{i,t}, n_{i,t})$
  - $I_t = \arg\max_i \theta_{i,t}$

↑ что мы дёргаем на шаге t