Machine Learning

Supervised learning

$D = \{d = (\bar{x}, y)\}$

$f: \bar{x} \mapsto y$

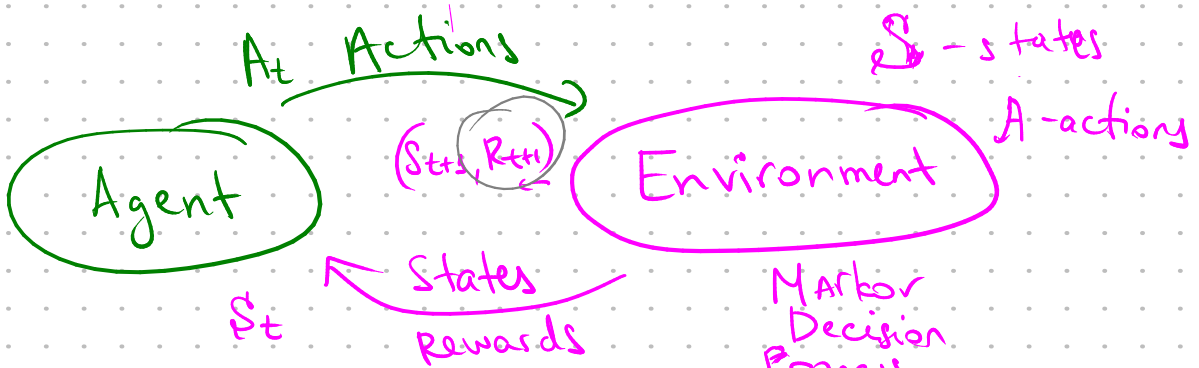
$p(y|\bar{x})$

Unsupervised learning

$D = \{\bar{x}\}$

$p(\bar{x})$

Reinforcement learning



Agent

strategy $\pi : S \to Prob[A]$

Environment

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

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

Шахматы

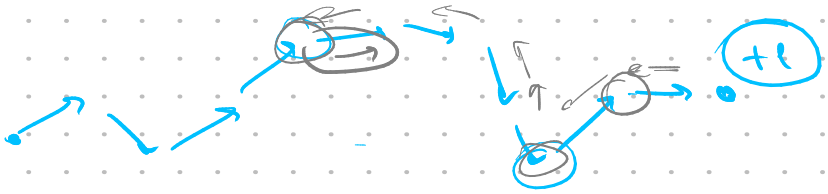
state – позиция с ходом игрока, время, рокировка, en passant, 50 ходов, троекр. повт.

action – ход

reward – +1 выигрыш, −1 проигрыш, 0 ничья

Episodic tasks

$R_t = 0$, $R_T \neq 0$?



Credit assignment

Exploration vs. exploitation

Multiarmed bandits

binary $R_t \in \{0, 1\}$

$R_t \in \mathbb{R}$

$|S| = 1 \quad A = (1, \dots, M) \quad (p_1, \dots, p_M)$

$A_t \longrightarrow \text{Env} \longrightarrow R_t$

$a_1, a_2, a_3, \dots, a_t, \dots$

$r_1, r_2, r_3, \dots, r_t, \dots$

$$a^* = \operatorname{argmax} \mathbb{E}[R_t]$$

Greedy:

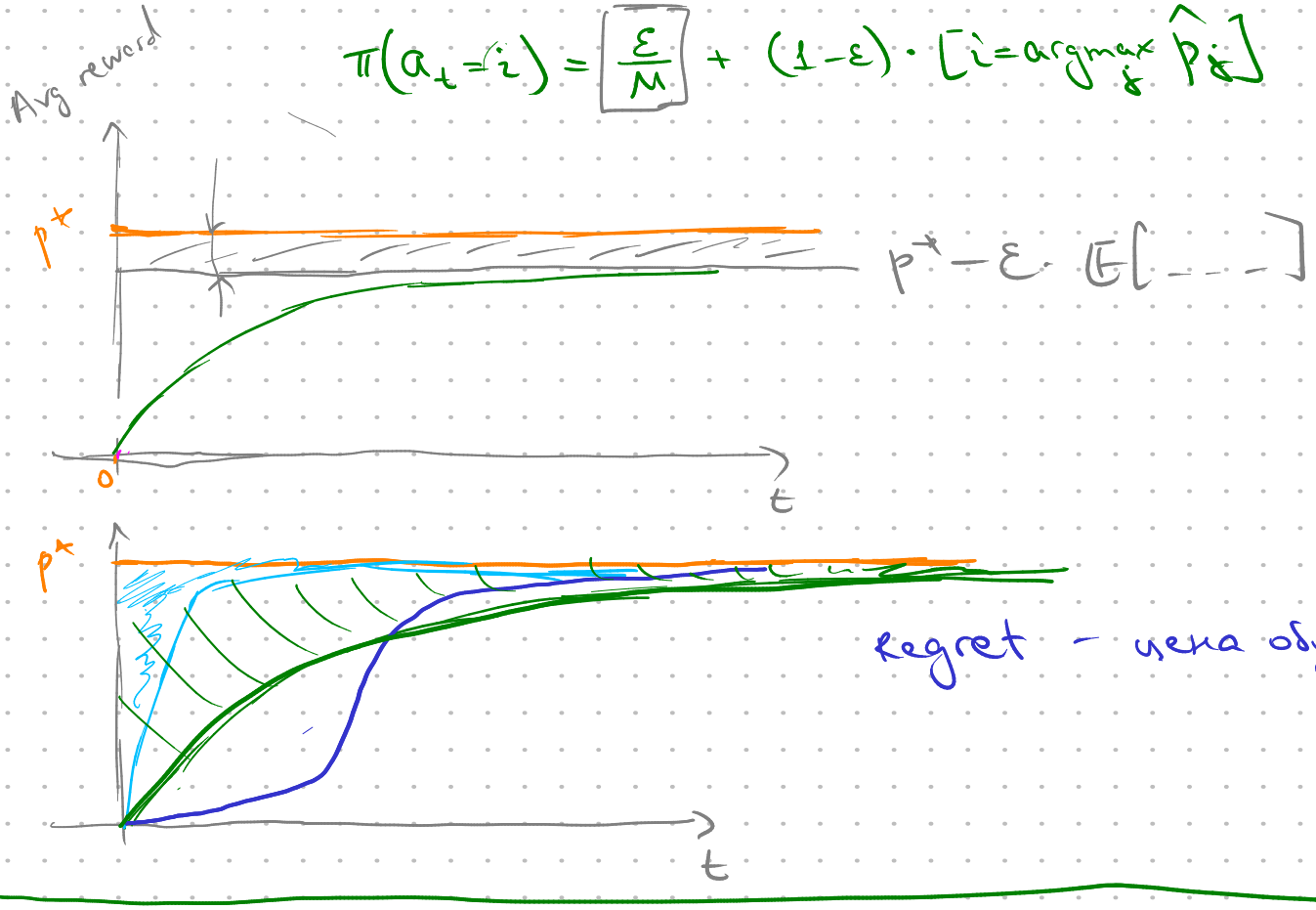
- $\hat{p}_i = \frac{1}{n_i} \sum_{t: a_t = i} r_t$
- $a_t = \operatorname{argmax}_i \hat{p}_i$

ε-Greedy:

- $a_t = \begin{cases} \operatorname{argmax}_i \hat{p}_i, & 1-\varepsilon \\ \text{uniform}, & \varepsilon \end{cases}$

$\varepsilon_t \to 0$ for $t \to \infty$

$$\pi(a_t = i) = \frac{\varepsilon}{M} + (1-\varepsilon) \cdot [i = \operatorname{argmax}_i \hat{p}_i]$$



$p^* - \varepsilon \cdot \mathbb{E}[\dots]$

regret – цена обучения

$$\hat{p}_i = \frac{r_1 + r_2 + \dots + r_n}{n} = \frac{1}{n}\left(r_n + \sum_{1}^{n-1} r_i\right) = \frac{1}{n}\left(r_n + (n-1)\hat{p}_i^{\,old}\right)$$

$$\hat{p}_i = \hat{p}_i^{\,old} + \frac{1}{n}\left(r_n - \hat{p}_i^{\,old}\right)$$

$$\text{New} = \text{Old} + \text{Stepsize}\left(\text{Target} - \text{Old}\right)$$

$$\hat{p}_{t+1} = \hat{p}_t + \alpha_t\left(r_t - \hat{p}_t\right) \qquad \alpha_t = \frac{1}{t} \Rightarrow \text{Avg}$$

$\alpha_t = \alpha = \text{const}$ ⇐ nonstationary bandits

$$\hat{p}_{t+1} = \hat{p}_t + \alpha(r_t - \hat{p}_t) = \alpha r_t + (1-\alpha)\hat{p}_t =$$

$$= \alpha r_t + (1-\alpha)\left(\alpha r_{t-1} + (1-\alpha)\hat{p}_{t-1}\right) = \dots$$

$$\dots = \alpha\left(r_t + (1-\alpha) r_{t-1} + (1-\alpha)^2 r_{t-2} + \dots + (1-\alpha)^k r_{t-k}\right)$$

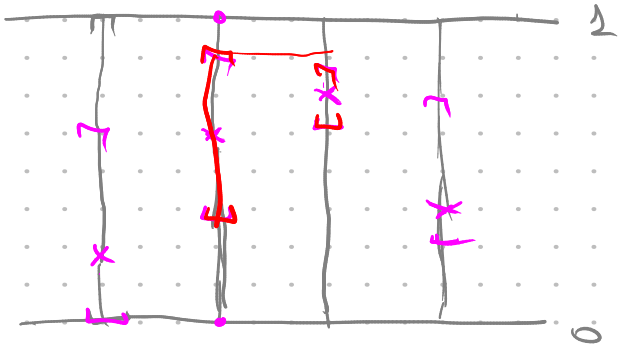
Thm Пусть $\hat{p}_{t+1} = \hat{p}_t + \alpha_t(r_t - \hat{p}_t)$ сход. с вер. 1,

если $\sum_{t=1}^{\infty} \alpha_t = \infty$ и $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$

Optimism under uncertainty

- доверительные интервалы – верхние границы
- optimistic initial values

$r_0 = 10$



- UCB upper confidence bound

$$\text{Priority}[i] = \hat{p}_i + (\dots)$$

UCB1: $a_t = \underset{i}{\operatorname{argmax}}\left[\hat{p}_i + c\sqrt{\frac{\ln t}{n_i}}\right]$

$n_i \approx \#[a_t = i]$

Thm. $c = \sqrt{2}$, $\sqrt{\frac{2\ln t}{n_i}}$ ⇒ UCB1 выбирает субоптимальное действие $O(\log T)$ раз за T шагов,

и $\text{regret} = O(\sqrt{MT\log T})$ $\Delta_i = p^* - p_i$

UCB – – – – –

Thompson sampling

- sample $p_i' \sim p(p_i|D)$
- $a_t = \underset{i}{\text{argmax}}\ p_i'$