ML: $D \rightsquigarrow ?$

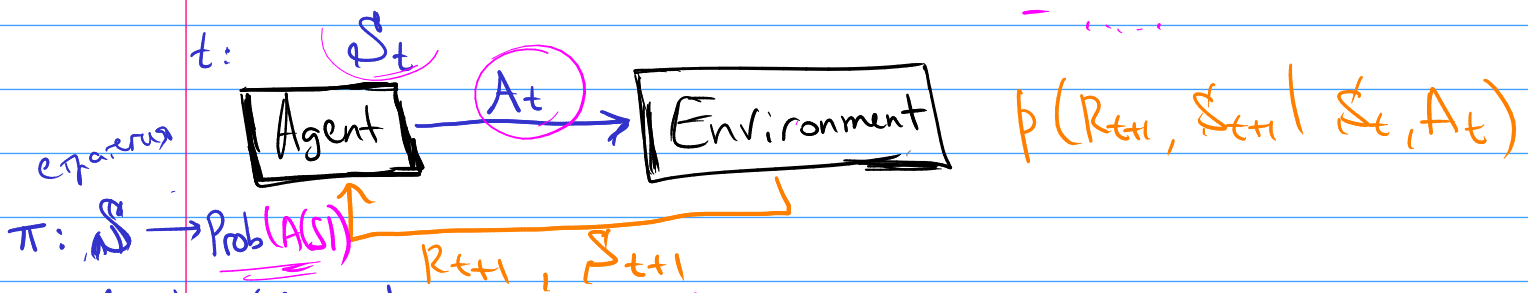
Supervised learning
$\bar{x} \rightsquigarrow y$?
$D = \{(\bar{x}, y)\}$
$p(y|\bar{x})$

Unsupervised learning
$\bar{x}$, $D = \{\bar{x}\}$
$p(\bar{x})$

Reinforcement learning

Markov decision process (MDP)

S_t в шахматах: – позиция – ....



$p(R_{t+1}, S_{t+1} | S_t, A_t)$

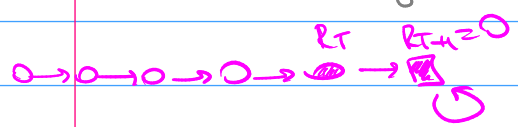
$\pi(s,a) = p(A_t = a | S_t = s)$

Episodic task

$1, \dots, T$

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

$\gamma = 1$



Continuous task

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

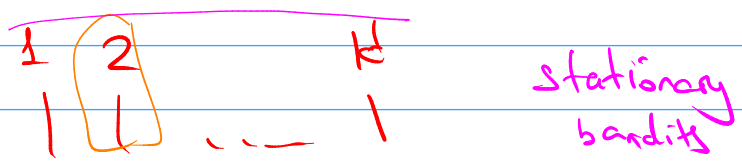
$\gamma < 1$

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

Sutton, Barto RL: An Introduction
1998 2018

1) Exploration vs. exploitation
2) Credit assignment

Multiarmed bandits



stationary bandits

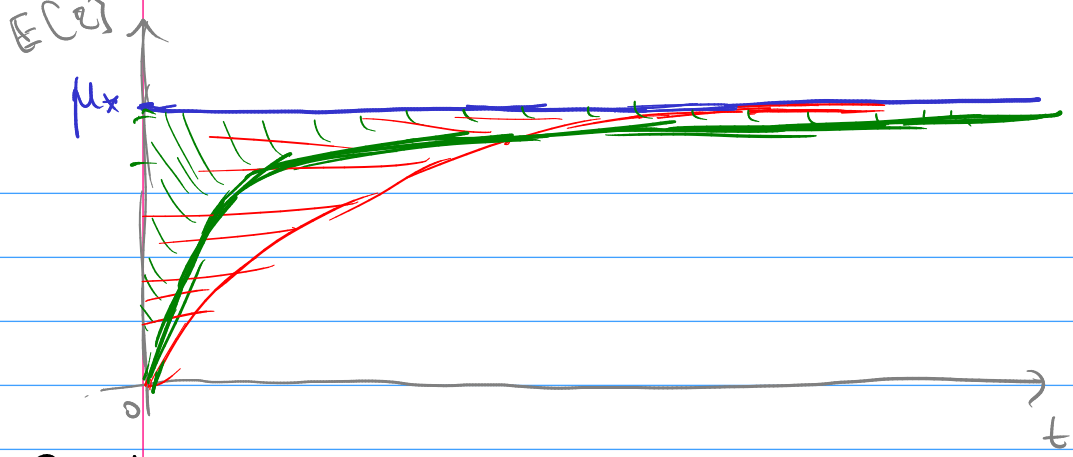
$|S| = 1$

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

t: $A_t \in \{1, 2, \dots k\}$

$R_{t+1} \sim p(R_{t+1} = r | A_t = a)$

$i_* = \underset{i=1..k}{\operatorname{argmax}} \mathbb{E}_{p(r|i)}[r]$



Regret

Greedy:
- Дёрнуть по 1 разу за каждую ручку
- for $t = k+1, \dots$:

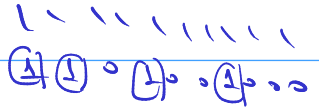
$$A_t = \arg\max_i \hat{\mu}_i(t)$$

$$\hat{\mu}_i(t) = \frac{1}{n_i} \sum_{t: A_t = i} R_t$$

Exploitation / ~~Exploration~~

binary bandit

$\mu_i = p(R=1 \mid i)$



$$\hat{\mu}_i(t+1) = \frac{1}{n_i+1} \sum_{t=1}^{n_i+1} R_t = \frac{1}{n_i+1} \left(R_{n_i+1} + \underbrace{\sum_{t=1}^{n_i} R_t}_{= n_i \cdot \hat{\mu}_i(t)} \right) =$$

$$= \frac{1}{n_i+1} \left(R_{n_i+1} + (n_i+1-1)\, \hat{\mu}_i(t) \right)$$

$$\hat{\mu}_i(t+1) = \hat{\mu}_i(t) + \frac{1}{n_i+1} \cdot \left(R_{n_i+1} - \hat{\mu}_i(t) \right)$$

Новая оценка = Старая оценка + α_t · (Новое знач. − Старая оценка)

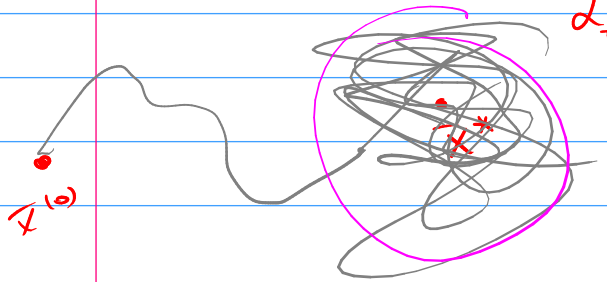
$R^2 \geq \|x^{(0)} - x^*\|^2$

$G^2 \geq \|g_t\|^2$

$$\|F(\bar{x}^*) - F(\bar{x}^{(k)})\| \leq \frac{R^2 + G^2 \cdot \sum_{t=1}^{k} \alpha_t^2}{2 \sum_{t=1}^{k} \alpha_t} \quad \left(\sum_{t=1}^{k} \alpha_t^2 < \infty \right)$$

$\alpha_t = h$

$$\frac{R^2}{2hk} + \frac{G^2 \cdot h}{2} \to \infty$$



Nonstationary bandits

$$\hat{\mu}_i(t+1) = \hat{\mu}_i(t) + \alpha\left(R_t - \hat{\mu}_i(t)\right) =$$

$$= \alpha R_t + (1-\alpha)\,\hat{\mu}_i(t) =$$

$$= \alpha R_t + (1-\alpha)\left(\alpha R_{t-1} + (1-\alpha)\,\hat{\mu}_i(t-1)\right) =$$

$$\ldots = \alpha R_t + \alpha(1-\alpha)R_{t-1} + \alpha(1-\alpha)^2 R_{t-2} + \alpha(1-\alpha)^3 R_{t-3} + \ldots$$

ε-Greedy:

$$A_t = \begin{cases} \operatorname{argmax}_i \hat{\mu}_i(t) & \text{с вер-ю } 1-\varepsilon, \\ \text{uniform} & \text{с вер-ю } \varepsilon. \end{cases}$$

Bingy $R_t \in \{0,1\}$ $\quad (n_1, w_1, n_2, w_2, \ldots, n_k, w_k)$, $0 \le w_i \le n_i$, $\sum n_i \le T$

$$V_*(n_1, \ldots, w_k) = \mathbb{E}\left[\sum_{\sum n_{i+1}}^{T} R_t\right]$$

Sga: $\sum n_i = T \Rightarrow V_* = 0$

Bepexy: $V_*(n_1, w_1 \ldots, n_k, w_k, \ldots) =$

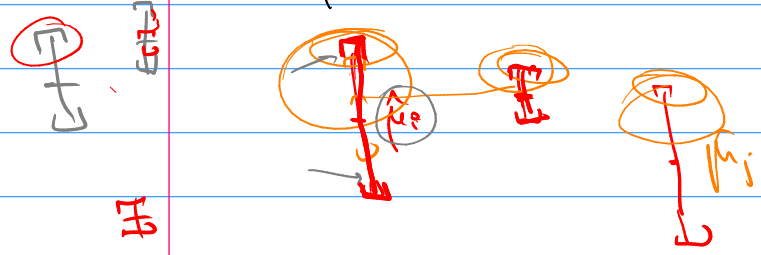
$$= \max_{i=1}^{k}\Big(\mu_i \cdot (1 + V_*(\ldots, n_i+1, w_i+1, \ldots)) + (1-\mu_i)\cdot V_*(\ldots, n_i+1, w_i, \ldots)\Big)$$

μ_i → $\frac{w_i+1}{n_i+2}$

Gittins index

$\operatorname{argmax}_i \; r(n_i, w_i)$

$\hat{\mu}_i$, $i = 1, \ldots, k$ $\qquad$ $\hat{\mu}_i \to$ softmax $\to$ —



$\operatorname{argmax}_i \left[\hat{\mu}_i + \ldots\right]$

Upper confidence bound

UCB1

$$\text{Priority}_i(t) = \hat{\mu}_i + c\cdot\sqrt{\frac{\log t}{n_i}}$$

ε-greedy $O(T)$

Thm. Dla $c=\sqrt{2}$ $\quad \hat{\mu}_i + \sqrt{\frac{2\log t}{n_i}}$ $\quad$ Regret $= O(\sqrt{KT\log T})$

$O(\log T)$ Lai, Robbins — optimal selection

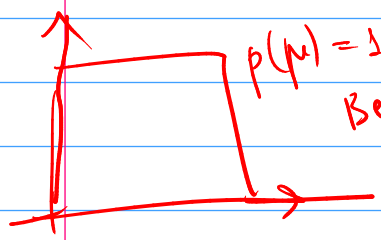
UCB2 ; UCB-Tuned $\quad \sqrt{\frac{\log t}{n_i}\cdot\min\left(\frac{1}{4}, V_j(n_j)\right)}$, gdzie

$$V_j(n_j) = \left(\frac{1}{n_j}\sum r^2\right) - \left(\frac{1}{n_j}\sum r\right)^2 + \sqrt{\frac{2\log t}{n_j}}$$

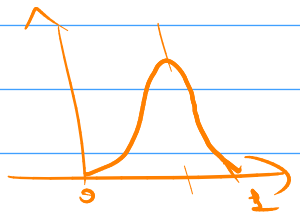
$a\cdot\sum_{i:\Delta_i\neq 0}\frac{\log T}{\Delta_i}$

$\Delta_i = \mu_* - \mu_i$

Thompson sampling: $\mu_1, \mu_2, \ldots, \mu_k \in [0,1]$

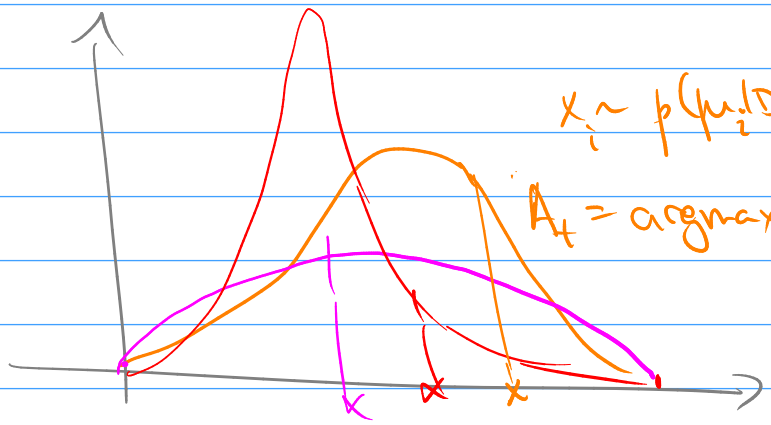


$\times \quad \mu^{w}(1-\mu)^{n-w} \quad \propto$



$\text{Beta}(w+1, n-w+1)$

$p(\mu)$ $\quad p(D|\mu)$ $\quad p(\mu|D)$



Probability matching: $\quad i \sim p\left(\mu_i = \max_j \mu_j \mid D\right)$