ML

Supervised

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

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

Unsupervised

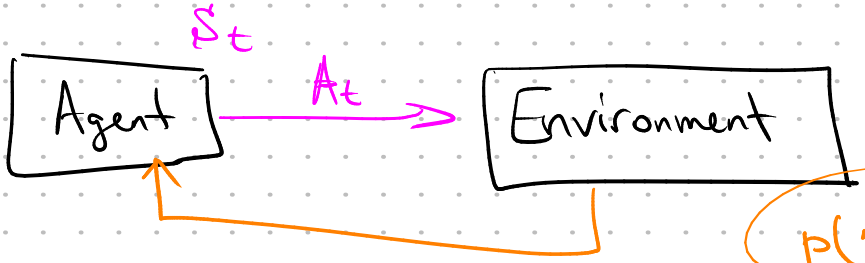
$D = \{\bar{x}\}$

$p(\bar{x})$

Reinforcement learning

Markov decision process

$t = 0, 1, \ldots$



dynamics

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

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

strategy

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

$\pi(a|s)$

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

Шахматы:

S – позиция + счётчики

A – ходы

R – ур-е партии

Episodic tasks

$t = 1, 2, \ldots, T$

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

$\gamma = 1$

Continuous tasks

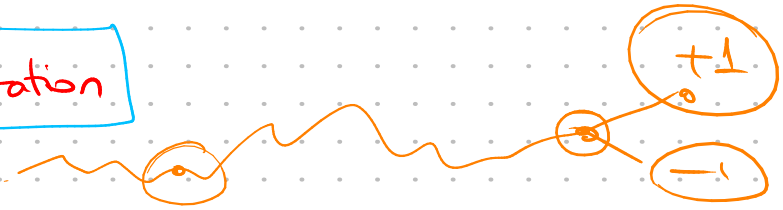
$t = 1, 2, \ldots, T, \ldots$

$\sum_{t=1}^{\infty} \gamma^{t-1} R_t \to \max_{\pi}$

$\gamma < 1$

1) Exploration vs. exploitation

2) ~~Credit assignment~~



Multiarmed bandits

$|S| = 1$

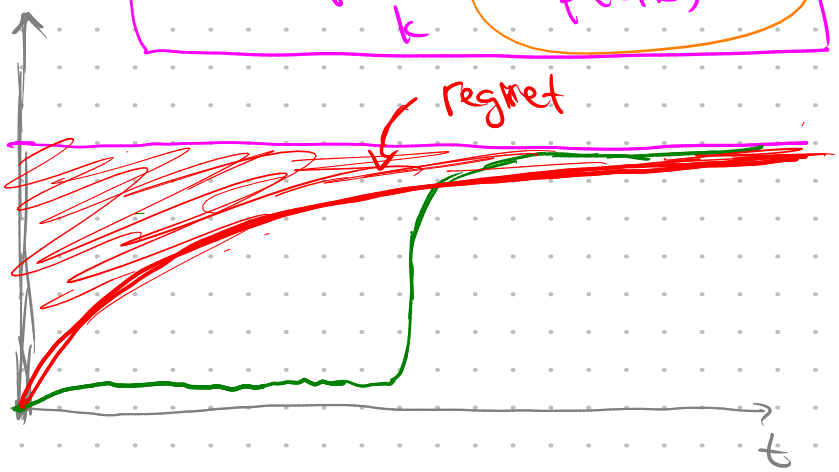
A: $a_1, a_2, \dots, a_K$

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

$p(a=a^*) = 1 - (\alpha?)$

$E[r] =$

$$a^* = \arg\max_k \; \mathbb{E}_{p(r|k)}[r]$$



Greedy:

- Sample 1 by ze $\forall a = 1, \dots, K$
- for $t = K+1, \dots$

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

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

binary bandits

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

ε-Greedy:

$\hat{\mu}_i(t)$

$$A_t = \begin{cases} \arg\max_i \hat{\mu}_i(t), & 1-\varepsilon \\ \text{uniform}, & \varepsilon \end{cases}$$

$$\pi(a) = (1-\varepsilon) \cdot [a = a^*] + \frac{\varepsilon}{K}$$

R_t $\quad \hat{\mu}_i(n+1) = \frac{1}{n+1} \sum_{t=1}^{n+1} R_t = \frac{1}{n+1} \left(R_{n+1} + \sum_{t=1}^{n} R_t \right) =$

$$= \frac{1}{n+1} R_{n+1} + \frac{n}{n+1} \cdot \hat{\mu}(n) \rightarrow \frac{1}{=} + \frac{1}{n+1}$$

$$\hat{\mu}(n+1) = \hat{\mu}(n) + \underbrace{\frac{1}{n+1}}_{\alpha_{n+1}}\left(R_{n+1} - \hat{\mu}(n)\right)$$

$$\sum_{n=1}^{\infty} \alpha_n = \infty \qquad \sum_{n=1}^{\infty} \alpha_n^2 < \infty$$

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

$\alpha_n = \alpha$ — Nonstationary bandits

$$\hat{\mu}(n+1) = \hat{\mu}(n) + \alpha\left(R_{n+1} - \hat{\mu}(n)\right) =$$

$$= (1-\alpha)\hat{\mu}(n) + \alpha R_{n+1} =$$

$$= \alpha R_{n+1} + (1-\alpha)\left(\alpha R_n + \hat{\mu}(n-1)\cdot(1-\alpha)\right) =$$

$$= \ldots = \alpha R_{n+1} + \alpha(1-\alpha)R_n + \alpha(1-\alpha)^2 R_{n-1} + \ldots$$

Binary $R_t \in \{0,1\}$ $t = 1, \ldots, T$ (n_k, w_k) $k = 1 \ldots K$ $0 \le w_k \le n_k$

$(n_1, w_1, n_2, w_2, \ldots, n_K, w_K)$, $\sum n_k \le T$

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

Gittins index

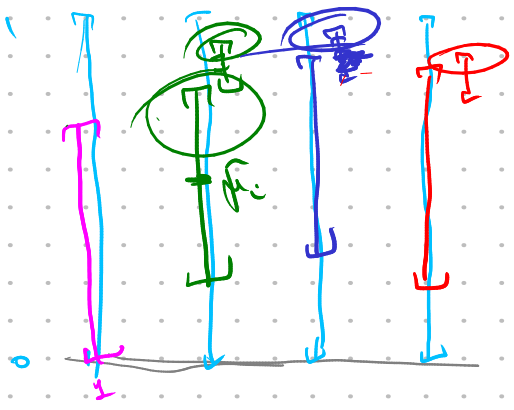
База: $\sum n_i = T \Rightarrow V_* = 0$

Переход: $V_*(n_1, \ldots, n_i, w_i, \ldots, w_K) =$

$$= \max_{i=1}^{K}\Big[\underbrace{\frac{w_i+1}{n_i+2}}_{\mu_i}\cdot\left(1 + V_*(n_1, \ldots, n_i+1, w_i+1, \ldots, w_K)\right)$$

$$+ (1-\mu_i)\cdot V_*(n_1, \ldots, n_i+1, w_i, \ldots, w_K)\Big)$$

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



UCB – upper confidence bound

UCB1 $$\text{Priority}_i(t) = \hat{\mu}_i(t) + c \cdot \sqrt{\frac{\log t}{n_i}}$$

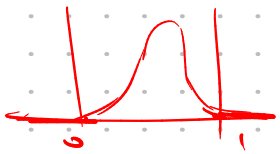
Thm $c = \sqrt{2} \Rightarrow$ Regret $= O(\sqrt{KT\log T})$

$\Delta_i = |\mu_* - \mu_i|$ $\sqrt{T\log T}$

UCB-Tuned $\hat{\mu}_i + \sqrt{\frac{\log t}{n_i} \cdot \min\left(\frac{1}{4}, V_i(n_i)\right)}$, где

$$V_i(n_i) = \frac{1}{n_i}\sum x^2 - \left(\frac{1}{n_i}\sum x\right)^2 + \sqrt{\frac{2\log t}{n_i}}$$

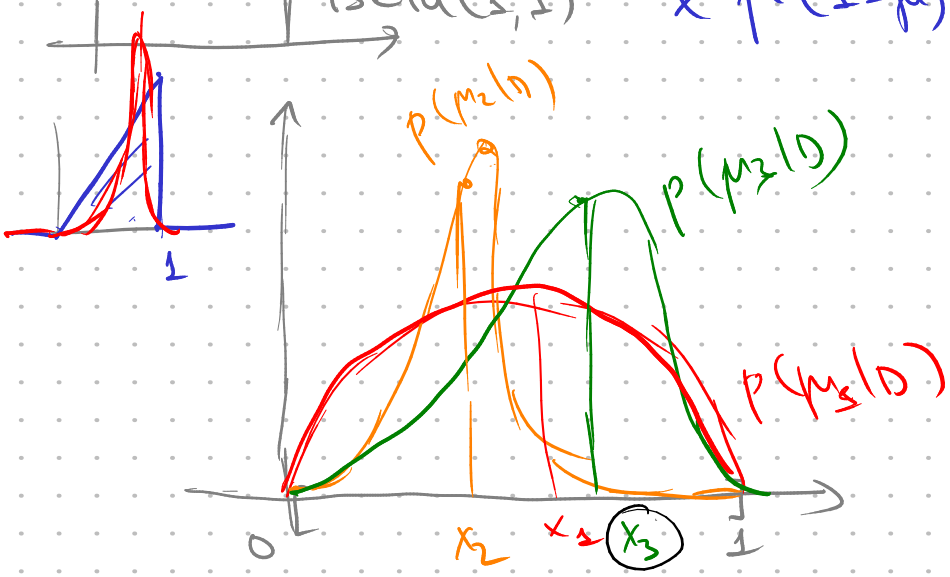
Thompson sampling binary $\mu_1, \mu_2, \ldots, \mu_K \in [0,1]$



$p(\mu) = \begin{cases} 1, & [0,1] \\ 0, & - \end{cases}$

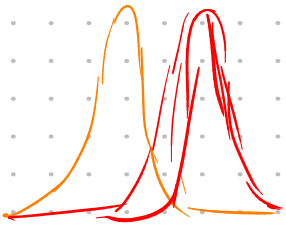
Beta(1,1)

$\times \mu^w (1-\mu)^{n-w} \propto p(\mu|D) = \text{Beta}(w+1, n-w+1)$



$x_i \sim p(\mu_i|D)$

$A_t = \arg\max x_i$



A/B testing

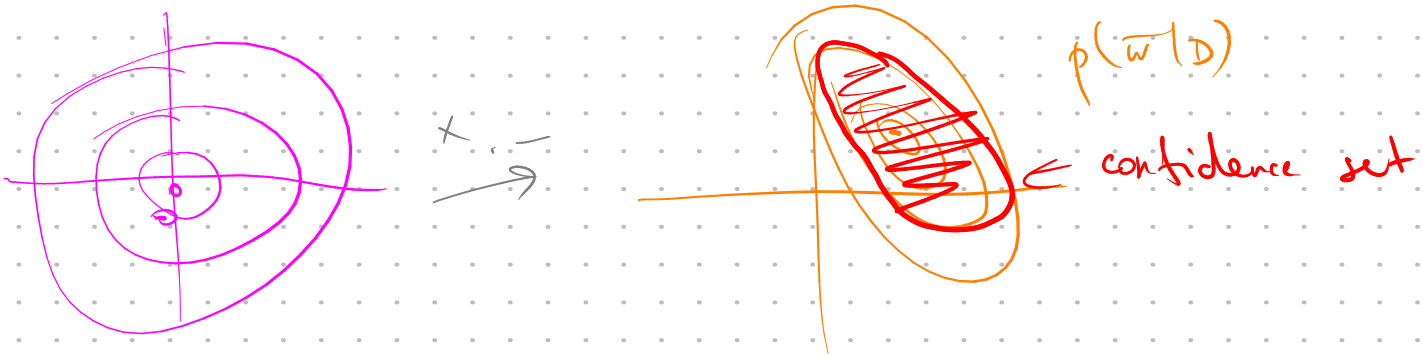
---

a_1, a_2, a_3 $t = 1, \ldots, T$ „$\{x_1, \ldots, x_M\}$

x_t – контекст $\pi : X \to \text{Prob}(A)$

$R_t \sim p(r_t | A_t, x_t)$ $\pi_1, \pi_2, \ldots, \pi_M$

$p(x_{t+1}|x_t)$ ~~$p(x_{t+1}|x_t)$~~


p(w|D)
confidence set