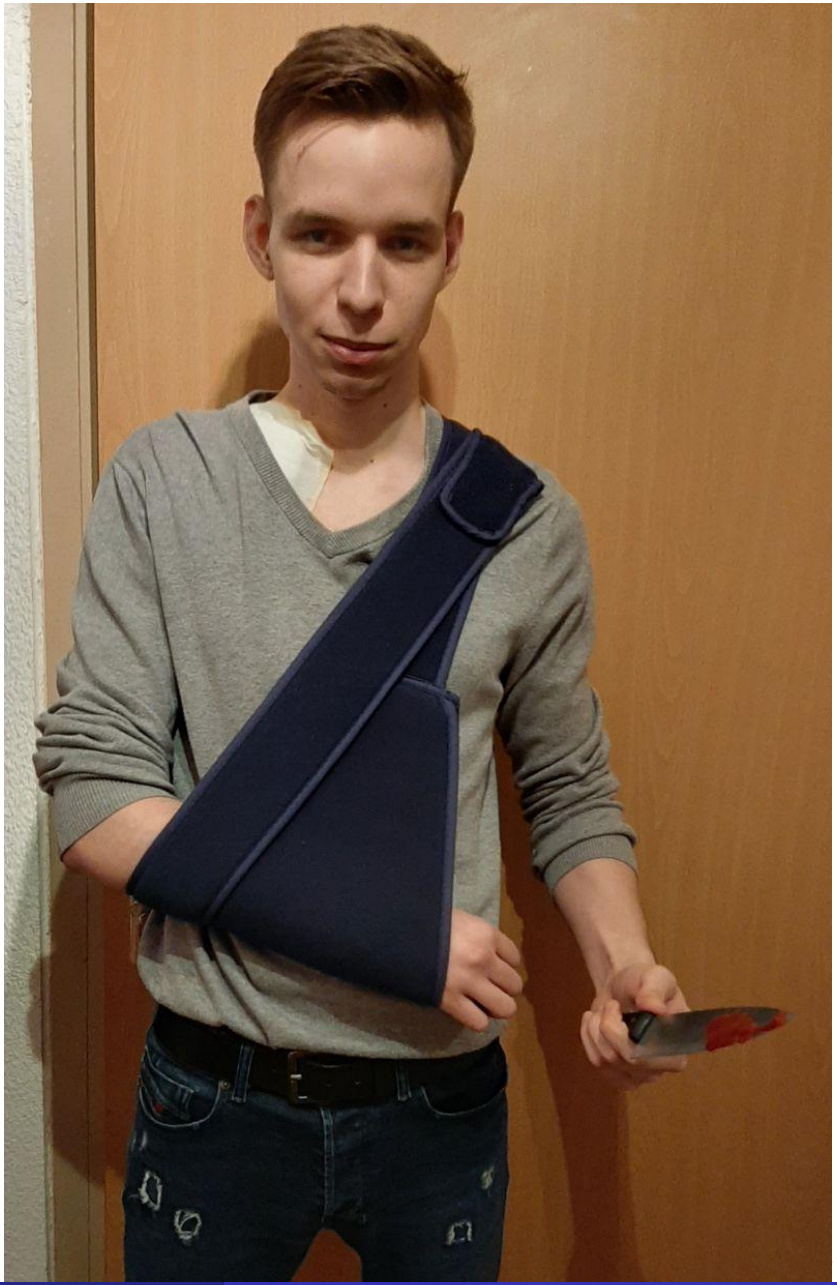
Multi-armed bandit problem

Шабалин Александр

17 февраля 2020 г.

Постановка задачи

- У нас есть K машин с распределениями наград $\{Y_1, \ldots, Y_K\}$.
- На каждом шаге t мы выбираем действие $a \in A$ и получаем награду r.
- $\Delta_a = r(a^*) - r(a)$
- $\text{regret}(T) = \sum_{\tau=1}^{T} \Delta_{a_\tau}$
- Задача: максимизировать $\sum_{t=1}^{T} r_t$ (минимизировать regret).



Greedy

$$\hat{Q}_t(a) = \frac{1}{N_t(a)} \sum_{t=1}^{T} r_t[a_t = a]$$

greedy

$$a_t = \underset{a \in A}{argmax}\, \hat{Q}_t(a)$$

ϵ-greedy

С вероятностью $1 - \epsilon$ выбирает $a_t = \underset{a \in A}{argmax}\, \hat{Q}_t(a)$

С вероятностью ϵ выбирает случайное действие

Томпсовское семплирование [reward$_i \sim$ Bernoulli(μ_i)]

Beta

$$\rho x(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$
$$\mathrm{B}(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$$
$$X \sim \mathrm{Beta}(\alpha,\beta) \quad \mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$$

Algorithm 1 Thompson Sampling for Bernoulli bandits

For each arm $i = 1, \ldots, N$ set $S_i = 0, F_i = 0$.

foreach $t = 1, 2, \ldots,$ **do**

  For each arm $i = 1, \ldots, N$, sample $\theta_i(t)$ from the Beta$(S_i + 1, F_i + 1)$ distribution.

  Play arm $i(t) := \arg\max_i \theta_i(t)$ and observe reward r_t.

  If $r = 1$, then $S_{i(t)} = S_{i(t)} + 1$, else $F_{i(t)} = F_{i(t)} + 1$.

end

Томпсовское семплирование [reward$_i \in [0, 1]$]

Algorithm 2 Thompson Sampling for general stochastic bandits

For each arm $i = 1, \ldots, N$ set $S_i(1) = 0, F_i(1) = 0$.

foreach $t = 1, 2, \ldots,$ **do**

For each arm $i = 1, \ldots, N$, sample $\theta_i(t)$ from the Beta$(S_i + 1, F_i + 1)$ distribution.

Play arm $i(t) := \arg\max_i \theta_i(t)$ and observe reward $\tilde{r}_t$.

Perform a Bernoulli trial with success probability $\tilde{r}_t$ and observe output r_t.

If $r_t = 1$, then $S_{i(t)} = S_{i(t)} + 1$, else $F_{i(t)} = F_{i(t)} + 1$.

end

$$\text{regret}^{TS}(T) = O\left(\left(\sum_{a \neq a^*} \frac{1}{\Delta_a^2}\right)^2 \ln T\right)$$

Upper Confidence Bounds

- $Q(a) = \mathbb{E}[r|a]$ - ожидаемая награда за действие a
- $\hat{Q}_t(a)$ - средняя награда за действие a за $t-1$ шаг
- $U_t(a) : \mathbb{P}[Q(a) > \hat{Q}_t(a) + U_t(a)]$ маленькая
- $a_t^{UCB} = \underset{a \in A}{argmax}(\hat{Q}_t(a) + U_t(a))$

UCB1

- $X_1, \ldots, X_n$ независимые случайные величины.
- $X_i \sim \text{Bernulli}(p)$.
- $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

Тогда для $u > 0$ из неравенства Чернова следует:

$$\mathbb{P}[\mathbb{E}[X] > \overline{X} + u] \leq e^{-2nu^2}$$

UCB1

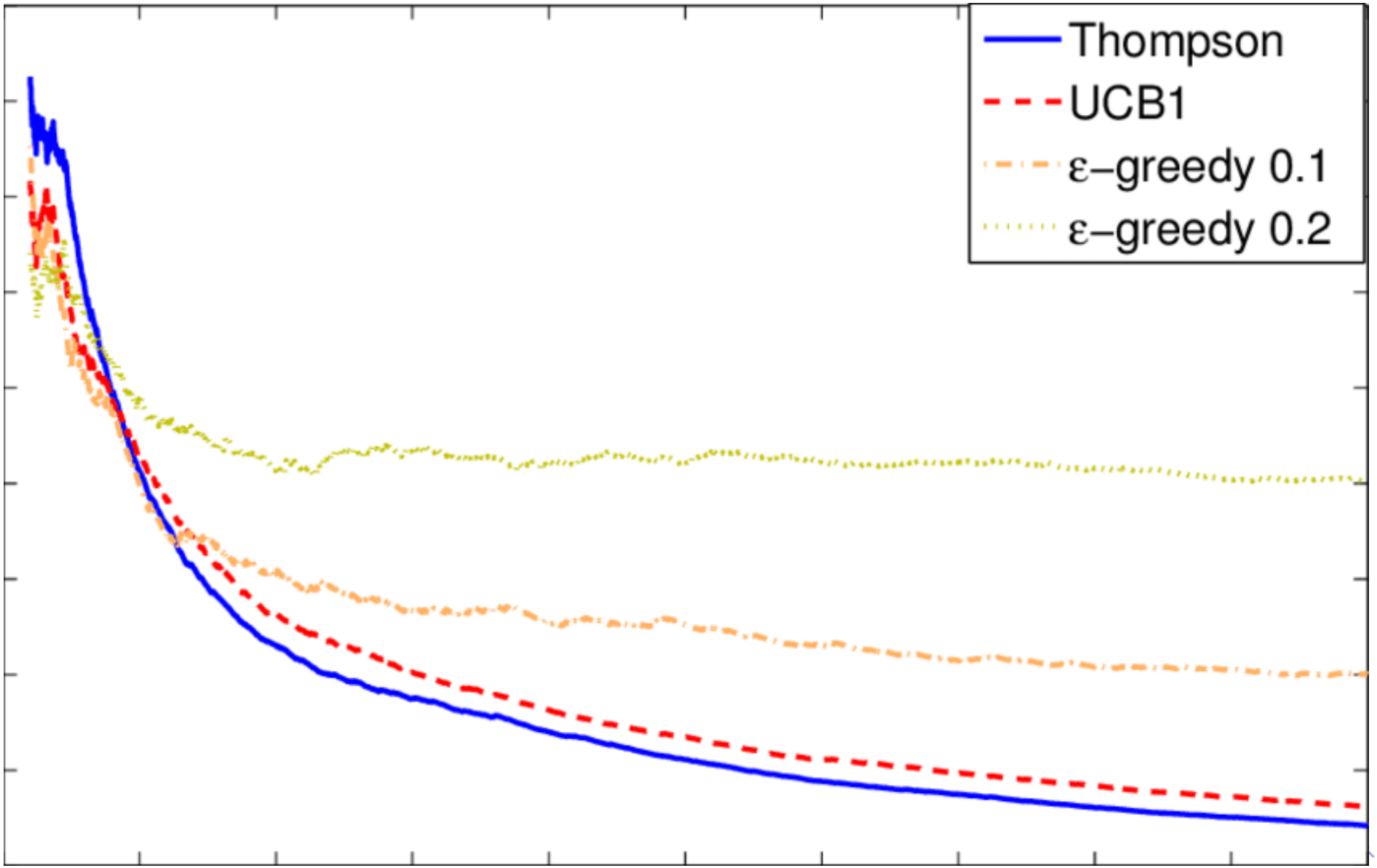
$$\mathbb{P}[Q(a) > \hat{Q}_t(a) + U_t(a)] \leq e^{-2N_t(a)U_t(a)^2}$$

- Мы хотим, чтобы $e^{-2N_t(a)U_t(a)^2}$ было как можно меньше.
- $e^{-2N_t(a)U_t(a)^2} = p \;\Rightarrow\; U_t(a) = \sqrt{\frac{-\ln p}{2N_t(a)}}$
- $p = t^{-4} \;\Rightarrow\; U_t(a) = \sqrt{\frac{2 \ln t}{N_t(a)}}$
- $a_t^{UCB1} = \underset{a \in A}{argmax}(\hat{Q}_t(a) + \sqrt{\frac{2 \ln t}{N_t(a)}})$

$$\text{regret}^{UCB1}(T) = O(\sqrt{KT \ln T})$$

Результаты

Regret comparison of TS, UCB1 and ϵ-greedy



Вопросы

- Постановка задачи multi-armed bandit problem
- В чем заключается адаптация томпсоновского семплирования для распределения в общем случае?
- Формула для выбора автомата в алгоритме UCB1 с объяснениями обозначений

Список литературы

- Analysis of Thompson Sampling for the Multi-armed Bandit Problem. Shipra Agrawal, Navin Goyal (2012)
- Optimism in the Face of Uncertainty: the UCB1 Algorithm (2013)