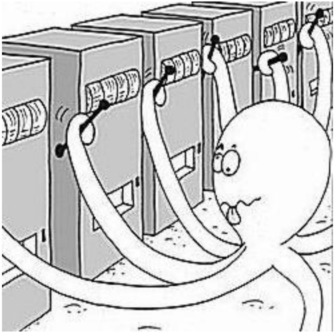
Recommender Systems
lecture 7: bandits for recommender systems

Alexey Grishanov

Moscow Institute of Physics and Technology

Spring 2023

Multi-armed bandits



Research question: how should I allocate my research time amongst my favorite open problems so as to maximize the value of my completed research?

source

Bandits overview

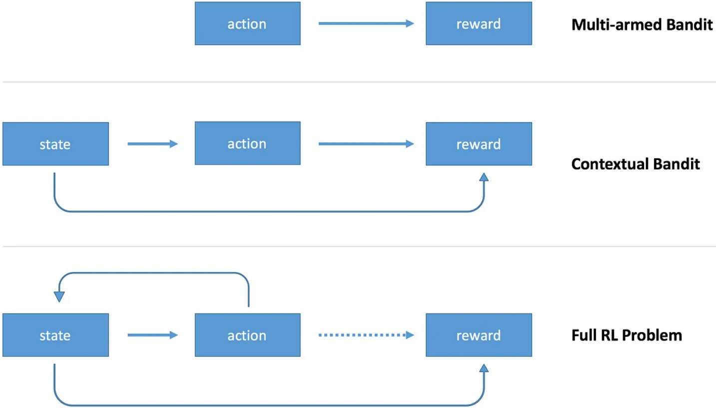


image source

Problem statement

Bernoulli multi-armed bandit

Can be described as a tuple of $\langle \mathcal{A}, \mathcal{R} \rangle$, where:

- We have K machines with reward probabilities $\{\theta_1, \ldots, \theta_K\}$.
- At time t, we take an action a on 1 machine and receive reward r.
- $\mathcal{A}$ is a set of actions. $Q(a) = \mathbb{E}[r \mid a] = \theta$. If action a_t at the time step t is on the i-th machine, then $Q(a_t) = \theta_i$.
- $\mathcal{R}$ is a reward function.
  $r_t = \mathcal{R}(a_t) = 1$ with probability $Q(a_t)$, 0 with probability $1 - Q(a_t)$

Objective: regret

$$\mathcal{L}_T = \mathbb{E}\left[\sum_{t=1}^{T} (\theta^* - Q(a_t))\right] \to \min$$

$$Q_t(a) = \frac{1}{N_t(a)} \sum_{\tau=1}^{t} r_\tau \mathbb{1}[a_\tau = a]$$

source

Exploration vs Exploitation

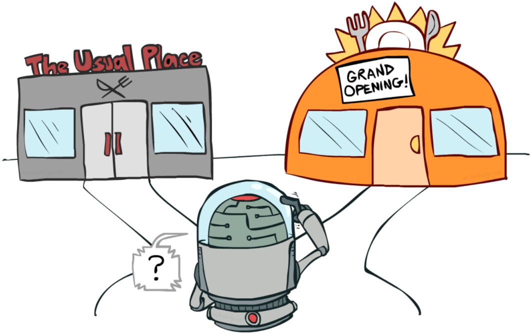
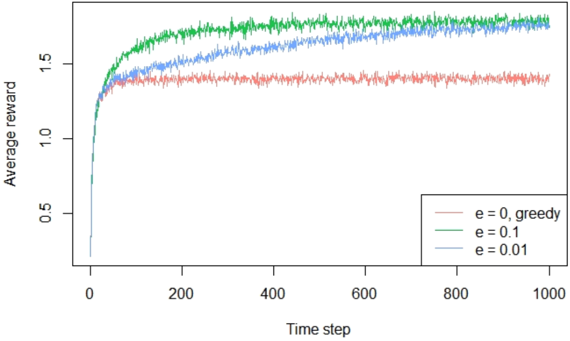


image source

ε-greedy

$$a_t = \begin{cases} \arg\max_{a \in \mathcal{A}} Q_t(a), \text{with probability } 1 - \varepsilon \\ \text{random}, \text{with probability } \varepsilon \end{cases}$$



Example from Sutton book (source)

Upper Confidence Bound (UCB)

$$a_t = \arg\max_{a \in \mathcal{A}} Q_t(a) + U_t(a)$$

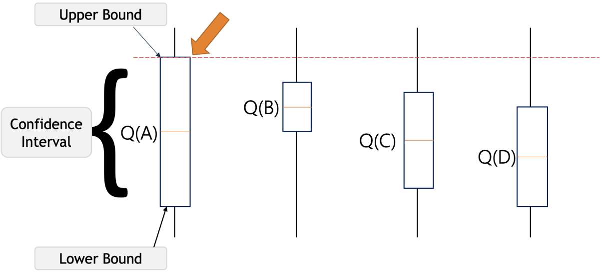


image credit: image source

Estimating confidence bounds

Hoeffding's Inequality

Let $X_1, \ldots, X_t$ be i.i.d. (independent and identically distributed) random variables and they are all bounded by the interval $[0, 1]$. The sample mean is $\bar{X}_t = \frac{1}{t}\sum_{\tau=1}^{t} X_\tau$. Then for $\mathrm{u} > 0$, we have:

$$\mathbb{P}\left[\mathbb{E}[X] > \bar{X}_t + u\right] \leq e^{-2tu^2}$$

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

$$e^{-2tU_t(a)^2} = p \Rightarrow U_t(a) = \sqrt{\frac{-\log p}{2N_t(a)}} \;\; \left(p = t^{-4} \text{ called } \mathrm{UCB}_1\right)$$

General UCB formula

$$a_t = \arg\max_{a \in \mathcal{A}} Q(a) + \alpha\sqrt{\frac{\log t}{N_t(a)}}$$

Thompson Sampling

Set of past observations $D = (a_i, r_i)_{i=1}^N$ modeled with $P(r|a, \theta)$.
Given $p(\theta)$, the posterior distribution is given by the Bayes rule:

$$P(\theta \mid D) \propto \prod P\left(r_i \mid a_i, \theta\right) P(\theta)$$

Algorithm 2 Thompson sampling for the Bernoulli bandit

Require: α, β prior parameters of a Beta distribution
  $S_i = 0, F_i = 0, \quad \forall i.$ {Success and failure counters}
  for $t = 1, \ldots, T$ **do**
    for $i = 1, \ldots, K$ **do**
      Draw θ_i according to Beta$(S_i + \alpha, F_i + \beta)$.
    end for
    Draw arm $\hat{\imath} = \arg\max_i \theta_i$ and observe reward r
    if $r = 1$ **then**
      $S_{\hat{\imath}} = S_{\hat{\imath}} + 1$
    else
      $F_{\hat{\imath}} = F_{\hat{\imath}} + 1$
    end if
  end for

image source

LinUCB (contextual bandits)

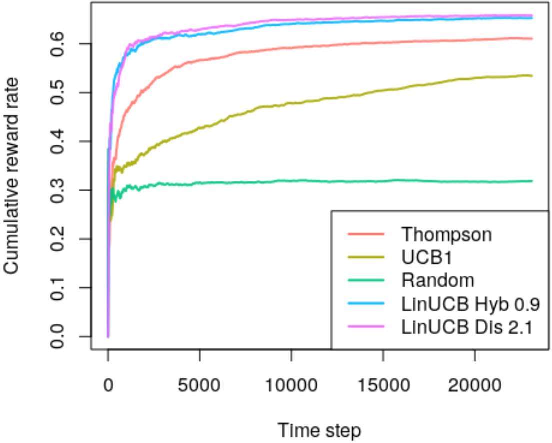
$$\mathbf{E}\left[r_{t,a} \mid \mathbf{x}_{t,a}\right] = \mathbf{x}_{t,a}^{\top}\boldsymbol{\theta}_a^*$$

Algorithm 1 LinUCB with disjoint linear models.

Inputs: $\alpha \in \mathbb{R}_+$
for $t = 1, 2, 3, \ldots, T$ **do**
  Observe features of all arms $a \in \mathcal{A}_t$: $\mathbf{x}_{t,a} \in \mathbb{R}^d$
  for all $a \in \mathcal{A}_t$ **do**
    if a is new **then**
      $\mathbf{A}_a \leftarrow \mathbf{I}_d$ (d-dimensional identity matrix)
      $\mathbf{b}_a \leftarrow \mathbf{0}_{d\times 1}$ (d-dimensional zero vector)
    end if
    $\hat{\boldsymbol{\theta}}_a \leftarrow \mathbf{A}_a^{-1}\mathbf{b}_a$ mean (to exploit)
    $p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^{\top}\mathbf{x}_{t,a} + \alpha\sqrt{\mathbf{x}_{t,a}^{\top}\mathbf{A}_a^{-1}\mathbf{x}_{t,a}}$ Variance (to explore)
  end for
  Choose arm $a_t = \arg\max_{a\in\mathcal{A}_t} p_{t,a}$ with ties broken arbitrarily, and observe a real-valued payoff r_t UCB style
  $\mathbf{A}_{a_t} \leftarrow \mathbf{A}_{a_t} + \mathbf{x}_{t,a_t}\mathbf{x}_{t,a_t}^{\top}$
  $\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r_t\mathbf{x}_{t,a_t}$
end for

image source

Algorithms comparison (Movielens-10M)



source

Can you beat the bandit?

1. https://iosband.github.io/2015/07/28/Beat-the-bandit.html
2. http://apbarraza.com/bandits_activity

Literature

1. *Richard S. Sutton, Andrew G. Barto* (2018). Reinforcement Learning: An Introduction
2. *O. Chapalle et al.* (2012). An Empirical Evaluation of Thompson Sampling
3. *D. Russo et al.* (2017). A Tutorial on Thompson Sampling
4. *L. Li et al.* (2010). A Contextual-Bandit Approach to Personalized News Article Recommendation