EXP3++ Algorithm – review

Yevgeny Seldin, Gabor Lugosi

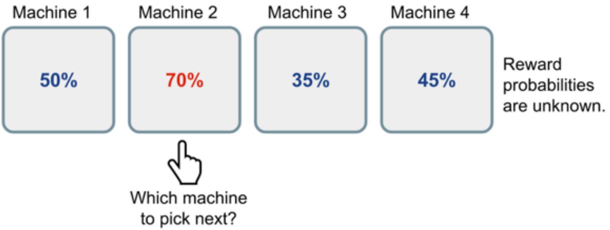
Institute of Information Science

Academia Sinica

Taiwan

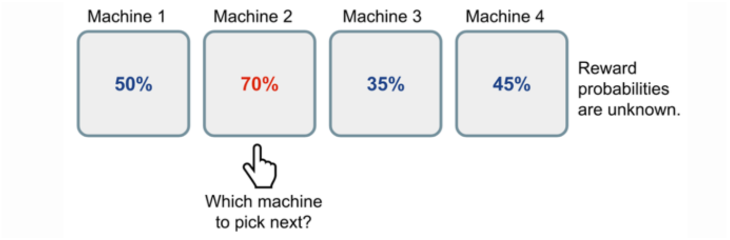
Nov 30, 2018

Multi-armed Bandit Problem



In our multi-armed bandit problem, there are K arms. At round t of the game, we choose an action A_t among K possible arms and observe the corresponding reward $r_t(A_t)$. Note that the rewards of other arms are not observed.

conti



Regret :

$$R(t) = \max_a \sum_{s=1}^{t} r_s(a) - \sum_{s=1}^{t} r_s(A_s)$$

The goal of the problem is to minimize the (expected) regret.

Loss generation models

1. Stochastic :
   The rewards $\{r_t(a)\}_{t,a}$ are sampled independently from an unknown distribution that depends on a, but not on t.
   - We use $\mu(a) = \mathbb{E}[r_t(a)]$ to denote the expected reward of an arm a.
   - Let $a^* = arg\min_a\{\mu(a)\}$ denote some best arm.
   - We define the gap $\Delta(a) = \mu(a) - \mu(a^*)$.
2. Adversarial :
   We consider that the reward sequences $\{r_t(a)\}_{t,a}$ are generated by an oblivious adversary.

Known algorithms and results

Usually, we have EXP3 algorithm work for adversarial regime to obtain

$$\mathcal{O}(\sqrt{KT \log K})$$

regret bound.
On the other hand, we have UCB algorithm work for stochastic regime to obtain

$$\mathcal{O}(\frac{\log T}{\Delta(a)})$$

bound for each suboptimal a.
However, is it possible to use a "single" algorithm to reach optimal regret bounds for both regime ?

$\Rightarrow$ best of both world ?

Algorithm

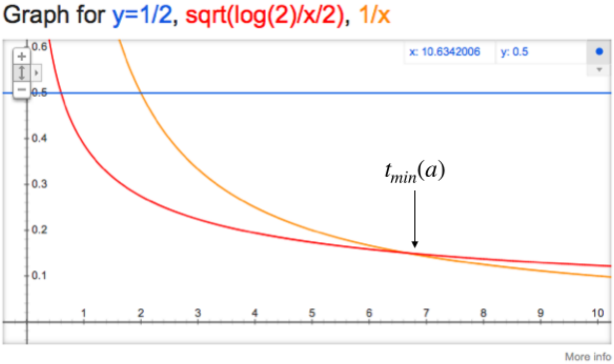
At every step, play action A_t according to $\tilde{\rho}_t$, where

$$\tilde{\rho}_t(a) = (1 - \sum_{a'} \epsilon_t(a'))\rho_t(a) + \epsilon_t(a)$$

and

$$\epsilon_t(a) = \min\{\frac{1}{K}, \sqrt{\frac{1}{tK}}, \frac{1}{t\hat{\Delta}_t(a)^2}\}$$

conti



Graph for y=1/2, sqrt(log(2)/x/2), 1/x
x: 10.6342006
y: 0.5
$t_{min}(a)$
0.6
0.5
0.4
0.3
0.2
0.1
1
2
3
4
5
6
7
8
9
10
More info