Improved Parametrization and Analysis of the EXP3++ Algorithm

Yevgeny Seldin, Gabor Lugosi

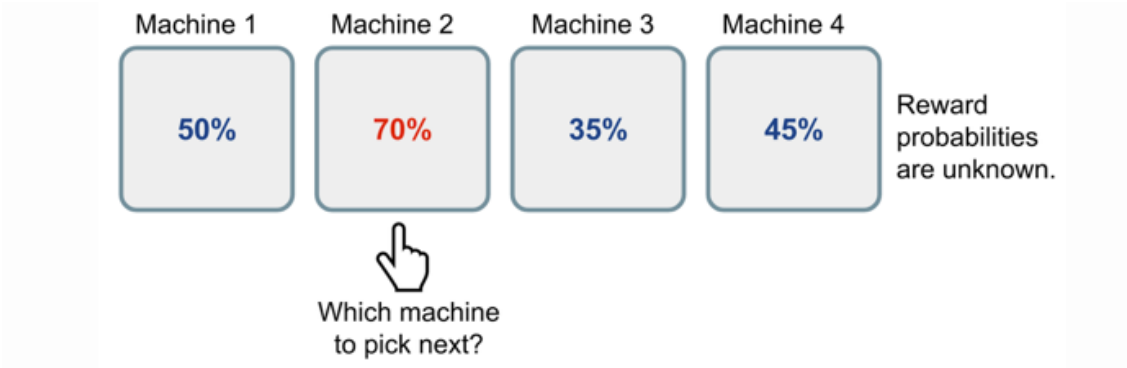
Institute of Information Science

Academia Sinica

Taiwan

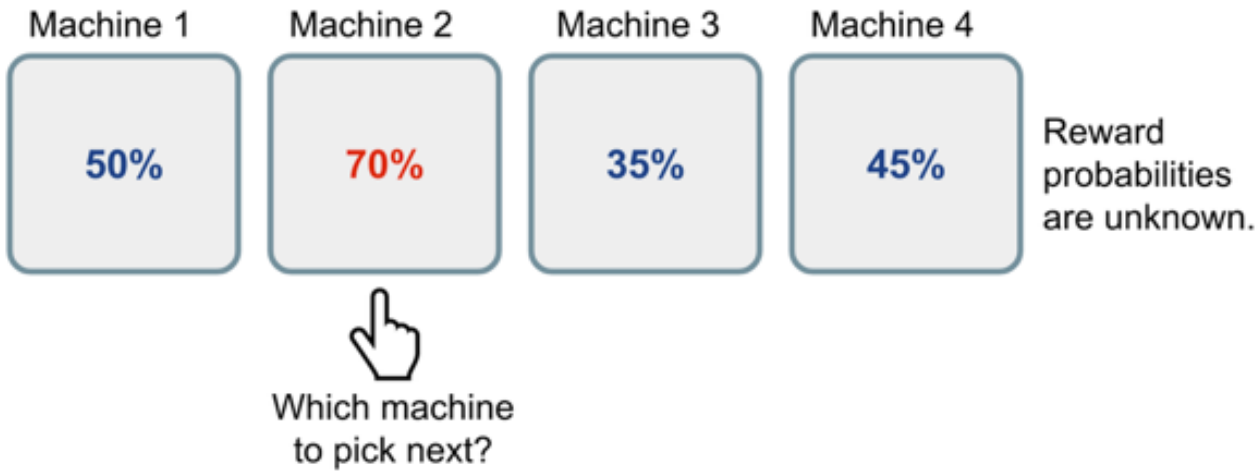
Oct 05, 2018

Explore v.s. Exploit



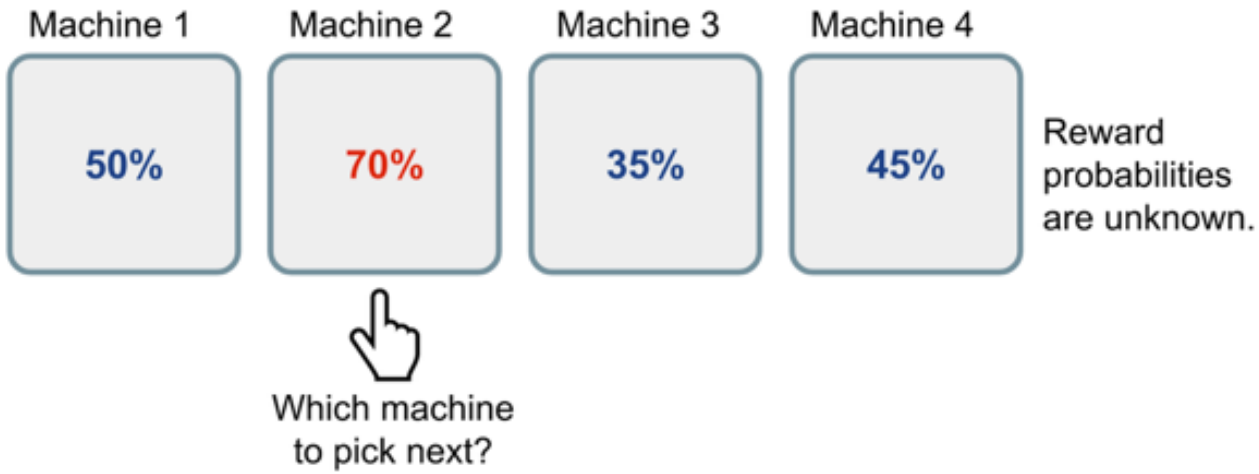
The exploration vs exploitation dilemma exists in many aspects of our life.
With exploitation, we take advantage of the best option we know.
With exploration, we take some risk to collect information about unknown options. The best long-term strategy may involve short-term sacrifices. For example, one exploration trial could be a total failure, but it warns us of not taking that action too often in the future.

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 loss $\ell_t(A_t)$. Note that the losses of other arms are not observed.

conti



Regret :

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

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

Loss generation models

1. Stochastic :
   The losses $\{\ell_t(a)\}_{t,a}$ are sampled independently from an unknown distribution that depends on a, but not on t.
   - We use $\mu(a) = \mathbb{E}[\ell_t(a)]$ to denote the expected loss 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 loss sequences $\{\ell_t(a)\}_{t,a}$ 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 ?

$$\Rightarrow \text{best of both world ?}$$

Results of EXP3++ Algorithm

Theorem 1 (2014)

The regret of the EXP3++ in the adversarial regime for any t satisfies :

$$R(t) \leq 4\sqrt{Kt \log K}$$

Theorem 2(2017)

The expected regret of EXP3++ in the stochastic regime satisfies :

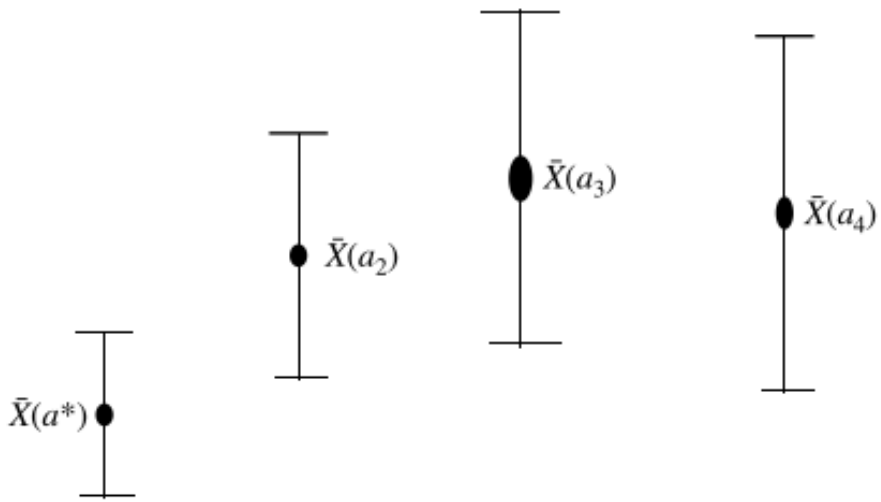
$$R(t) = \mathcal{O}\left(\sum_{a:\Delta(a)>0} \frac{(\ln t)^2}{\Delta(a)}\right) + \tilde{\mathcal{O}}\left(\sum_{a:\Delta(a)>0} \frac{K}{\Delta(a)^3}\right)$$

idea - UCB

For a stochastic arm, we can estimate the expected loss by observing its empirical loss with some confidence interval.

Theorem 10 (Hoeffding's inequality) *Let $X_1, \ldots, X_n$ be i.i.d. random variables, such that $0 \leq X_i \leq 1$ and $\mathbb{E}[X_i] = \mu$ for all i. Then*

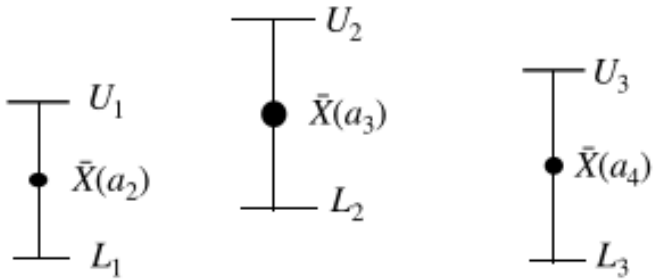
$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right\} \leq \delta,$$

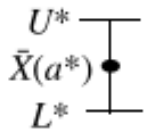


conti

As time getting longer, we can estimate the gap between these arms. For example, we have

$$\begin{aligned}
\hat{\Delta}_t(a_2) &= max\{0, LCB_t(a_2) - \min_{a'} UCB_t(a')\} \\
&= max\{0, L_1 - U^*\} \\
&= L_1 - U^* \qquad (1) \\
\hat{\Delta}_t(a^*) &= 0 \qquad (2)
\end{aligned}$$





Algorithm

Algorithm 1 EXP3++

for $t = 1, 2, \cdots$ **do**

$\forall a$, estimate upper bound ($UCB_t(a)$) and lower bound ($LCB_t(a)$)

$\forall a,\ \hat{\Delta}_t(a) = max\{0, LCB_t(a) - \min_{a'} UCB_t(a')\}$

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

$\forall a,\ \rho_t(a) = \exp\{-\eta_t \tilde{L}_{t-1}(a)\}/\mathcal{Z}_t$

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

Play action A_t according to $\tilde{\rho}_t$

Observe and suffer the loss $\ell_t(A_t)$

$\forall a,\ \tilde{\ell}_t(a) = \frac{\ell_t(a)}{\tilde{\rho}_t(a)} \mathbb{1}\{A_t = a\}$

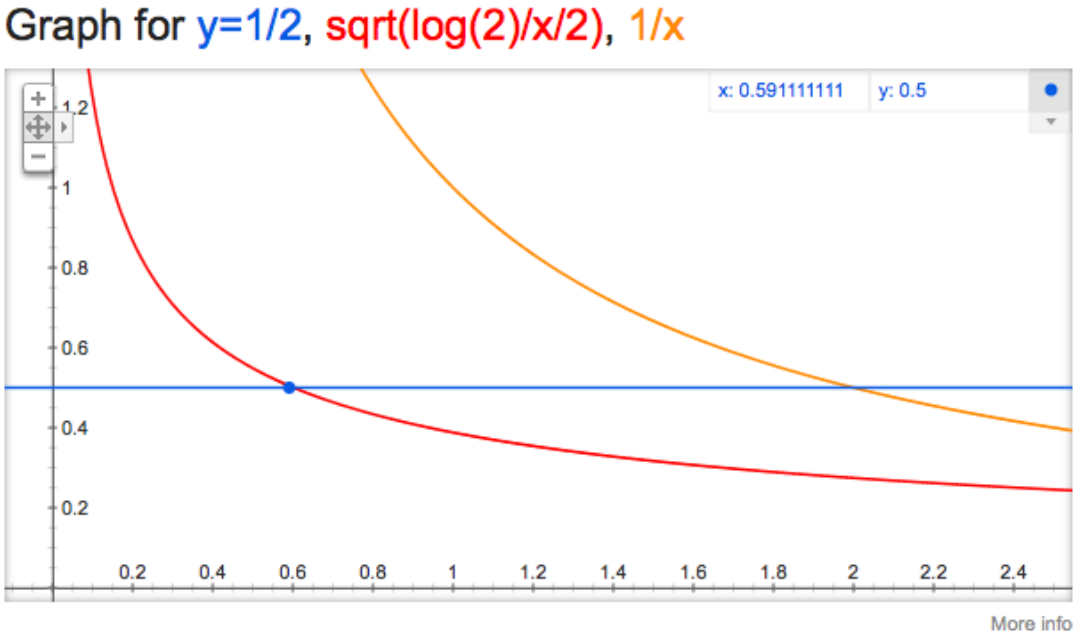
$\forall a,\ \tilde{L}_t(a) = \tilde{L}_{t-1}(a) + \tilde{\ell}_t(a)$

end for

idea - exploration rate

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

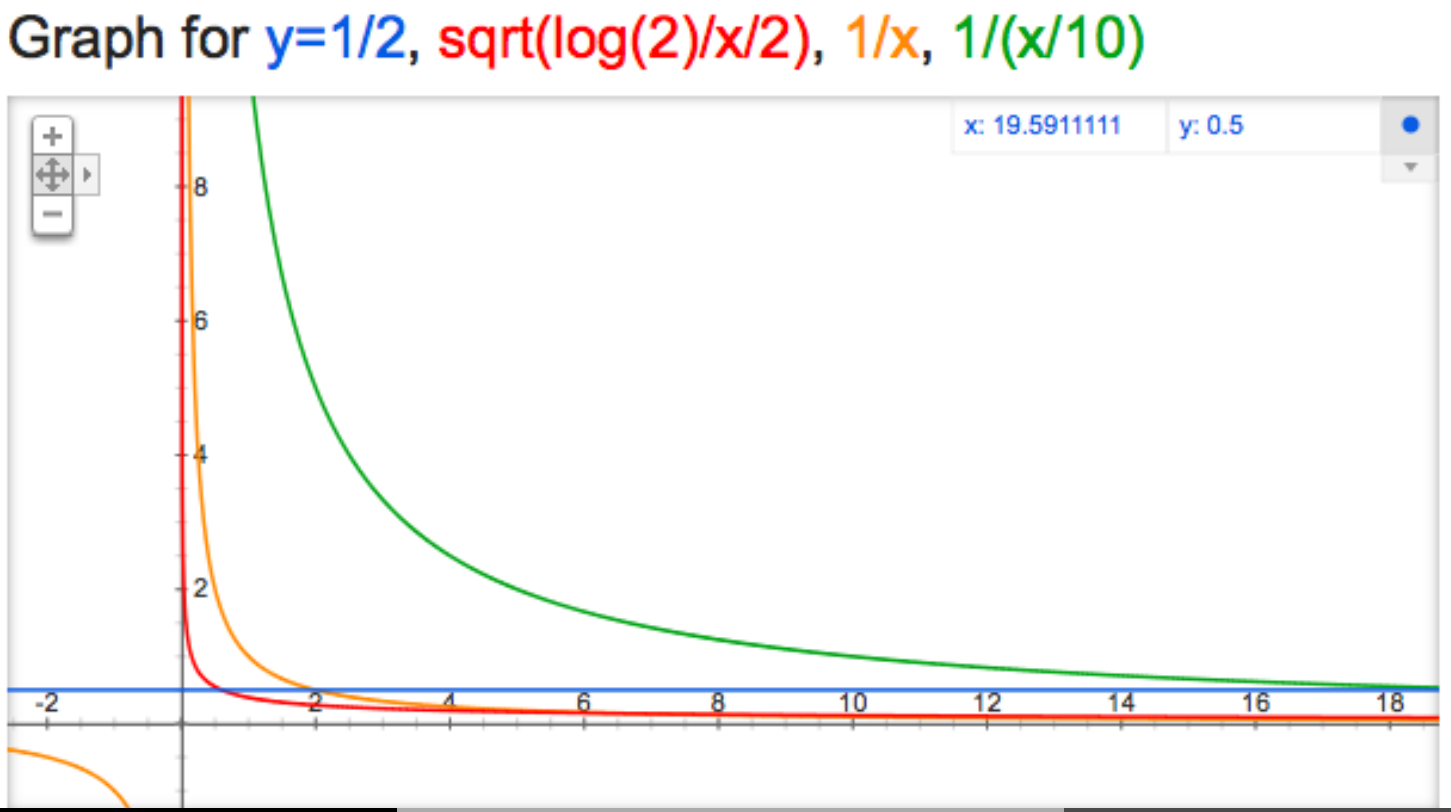
- Initially, we explore each arm with same probability : $1/K$
- The exploration rate then decrease adaptively.
- Suppose that we know $\hat{\Delta}_t(a) = \Delta_t(a)$, the exploration rate of arm a then goes from the blue line, to the red one, and eventually to orange line.



conti

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

- Suppose that we know $\hat{\Delta}_t(a) = \Delta_t(a)$.
- The arm with larger gap change to the third stage ($\frac{\ln t}{t\Delta_t(a)^2}$) earlier than the arm with smaller gap.

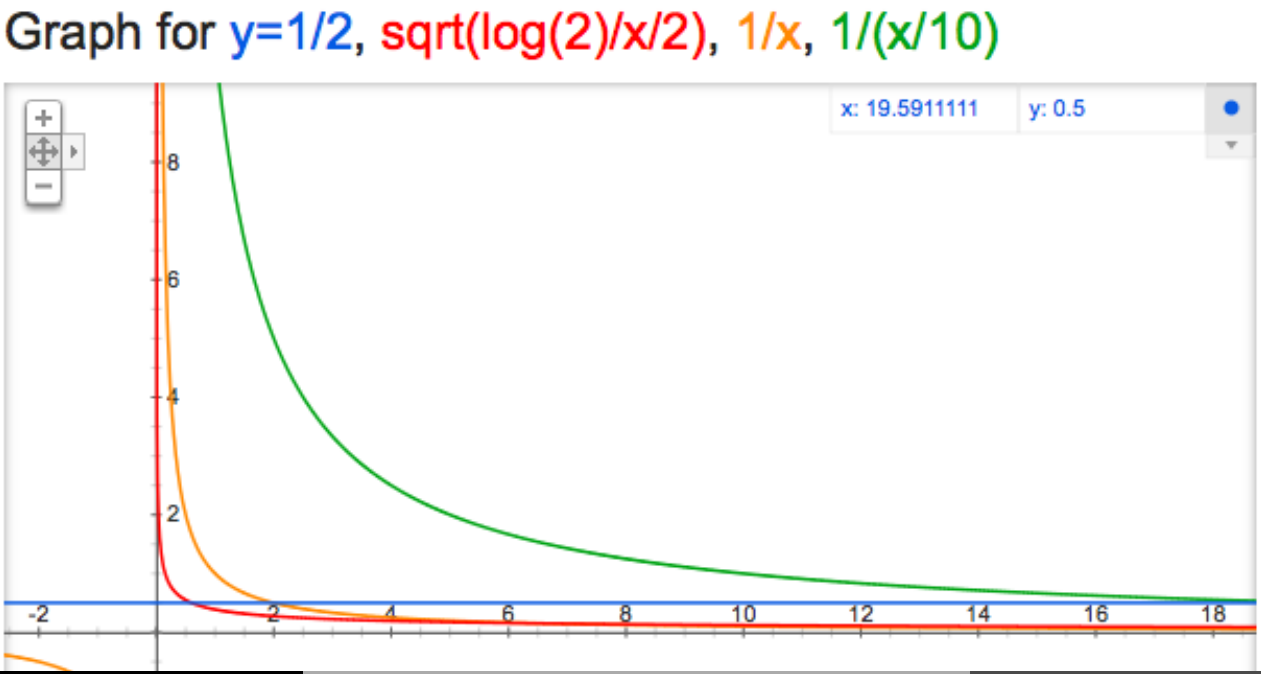


conti

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

- Now we consider the estimated gap $\hat{\Delta}_t(a)$.
- Once the gap of a arm can be appropriately measured, we know when to change to the third stage.

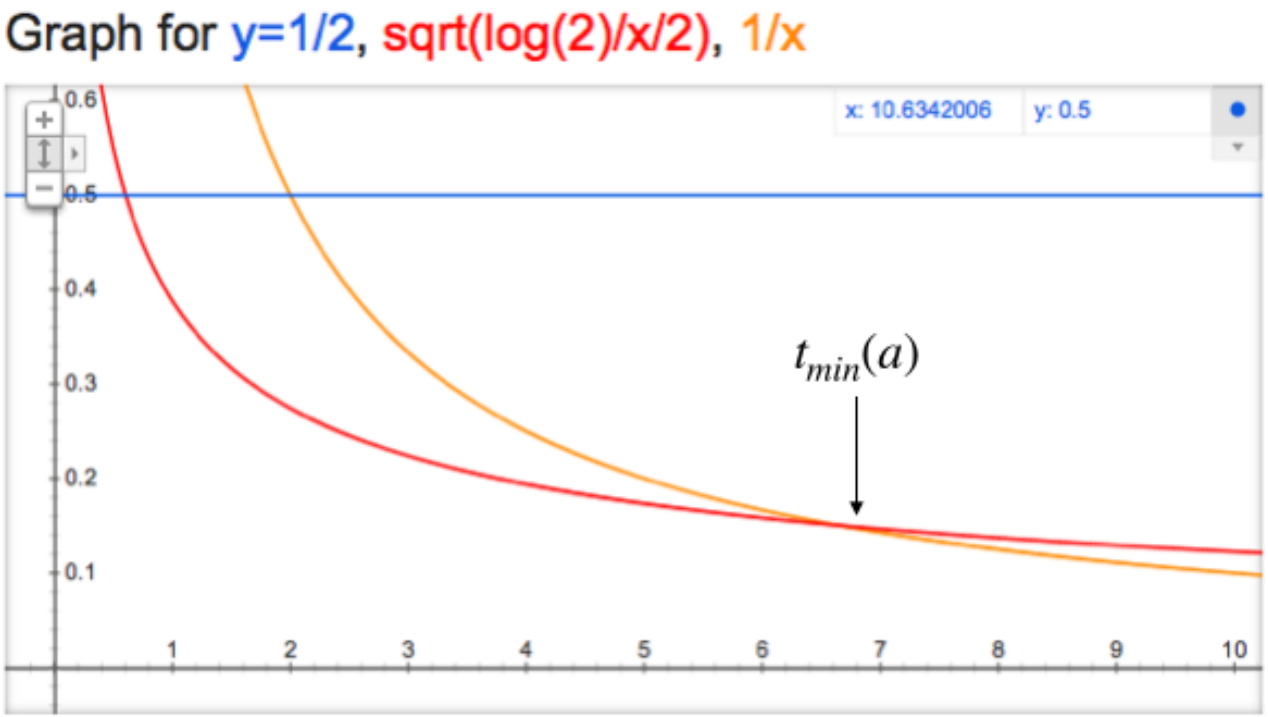
$$\hat{\Delta}_t(a) \text{ goes from 0 to } \Delta_t(a)$$



conti

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

- $\hat{\Delta}_t(a)$ goes from 0 to $\Delta_t(a)$
- Since for all arms, they have same probability during the blue line stage and red line stage. The minimum time to identify the gap of each arm is $t_{min}(a)$.



Algorithm

Algorithm 2 EXP3++

for $t = 1, 2, \cdots$ **do**

$\forall a$, estimate upper bound ($UCB_t(a)$) and lower bound ($LCB_t(a)$)

$\forall a,\ \hat{\Delta}_t(a) = max\{0, LCB_t(a) - \min_{a'} UCB_t(a')\}$

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

$\forall a,\ \rho_t(a) = \exp\{-\eta_t \tilde{L}_{t-1}(a)\}/\mathcal{Z}_t$

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

Play action A_t according to $\tilde{\rho}_t$

Observe and suffer the loss $\ell_t(A_t)$

$\forall a,\ \tilde{\ell}_t(a) = \frac{\ell_t(a)}{\tilde{\rho}_t(a)} \mathbb{1}\{A_t = a\}$

$\forall a,\ \tilde{L}_t(a) = \tilde{L}_{t-1}(a) + \tilde{\ell}_t(a)$

end for

Analysis - regret

Pseudo regret :

$$R(t) = \sum_a \mathbb{E}[N_t(a)] \cdot \Delta(a)$$

Since the number of iterations for each arm to be played depends on the distribution $\tilde{\rho}_t$,

$$\begin{aligned}
\mathbb{E}[N_t(a)] &= \sum_{s=1}^{t} \mathbb{E}[\tilde{\rho}_s] \\
&= \sum_{s=1}^{t} \mathbb{E}[(1 - \sum_{a'} \epsilon_s(a'))\rho_t(a) + \epsilon_s(a)] \\
&\leq \sum_{s=1}^{t} \mathbb{E}[\rho_s(a)] + \sum_{s=1}^{t} \mathbb{E}[\epsilon_s(a)] \qquad (3)
\end{aligned}$$

$\epsilon_s(a)$

$$\epsilon_s(a) = \min\{\frac{1}{K}, \sqrt{\frac{\ln K}{sK}}, \frac{\ln s}{s\hat{\Delta}_s(a)^2}\}$$

To obtain the upper bound of $\epsilon_s(a)$ after t_{min}, we have to properly control the lower bound of $\hat{\Delta}_s(a)$.

$\rho_s(a)$

By definition of $\rho_s(a)$, we have

$$\rho_s(a) = \exp\{-\eta_s \tilde{L}_{s-1}(a)\}/\mathcal{Z}_s \leq \exp\{-\eta_s \tilde{\Delta}_s(a)\},$$

where $\tilde{\Delta}_s(a) := \tilde{L}_{s-1}(a) - \tilde{L}_{s-1}(a^*)$. To obtain an upper bound for $\rho_s(a)$, we need a lower bound for $\tilde{\Delta}_s(a)$.

Lemma 1

For $t \geq t_{min}(a)$,

$$\Pr\left[\tilde{\Delta}_t(a) \leq \frac{1}{2} t\Delta(a)\right] \leq \delta$$

Proof of Lemma 1

Theorem 9 (Bernstein's inequality for martingales) *Let $X_1, \ldots, X_n$ be a martingale difference sequence with respect to filtration $\mathcal{F}_1, \ldots, \mathcal{F}_n$, where each X_j is bounded from above, and let $S_i = \sum_{j=1}^{i} X_j$ be the associated martingale. Let $\nu_n = \sum_{j=1}^{n} \mathbb{E}\left[(X_j)^2 \middle| \mathcal{F}_{j-1}\right]$ and $c_n = \max_{1 \le j \le n} \{X_j\}$. Then for any $\delta > 0$:*

$$\mathbb{P}\left\{\left(S_n \ge \sqrt{2\nu \ln \frac{1}{\delta}} + \frac{c \ln \frac{1}{\delta}}{3}\right) \wedge (\nu_n \le \nu) \wedge (c_n \le c)\right\} \le \delta.$$

$$\begin{aligned}
&\Pr\left[\tilde{\Delta}_t(a) \le \frac{1}{2} t\Delta(a)\right] \\
&= \Pr\left[t\Delta(a) - \tilde{\Delta}_t(a) \ge \frac{1}{2} t\Delta(a)\right] \\
&\le \Pr\left[c_t \ge c\right] + \Pr\left[\nu_t \ge \nu\right] \\
&+ \Pr\left[\left(t\Delta(a) - \tilde{\Delta}_t(a) \ge \frac{1}{2} t\Delta(a)\right) \wedge \Pr\left[c_t \le c\right] \wedge \left(\nu_t \le \nu\right)\right] \qquad (4)
\end{aligned}$$

conti

Here $X_i = \Delta(a) - (\tilde{\ell}_i(a) - \tilde{\ell}_i(a^*))$. Therefore, we have to control

1. $c_t = max_{i \le t}\{X_i\}$
2. $\nu_t = \sum_{i=1}^{t} \mathbb{E}[(X_i)^2]$

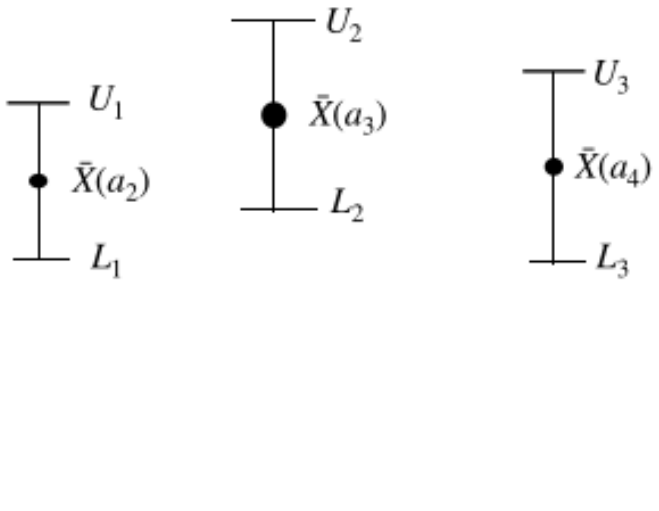
$$X_i \le 1 + \tilde{\ell}_i(a^*) \le 1 + \frac{1}{\epsilon_i(a^*)} \tag{5}$$

$$\begin{aligned} \mathbb{E}[(X_i)^2] &\le \mathbb{E}[(\tilde{\ell}_i(a) - \tilde{\ell}_i(a^*))^2] \\ &= \mathbb{E}[\tilde{\ell}_i(a)^2 + \tilde{\ell}_i(a^*)^2] \end{aligned}$$

$$\mathbb{E}[\tilde{\ell}_i(a)^2] = \tilde{\rho}_i(a) \left(\frac{\ell_i(a)}{\tilde{\rho}_i(a)}\right)^2 \le \frac{1}{\tilde{\rho}_i(a)} \le \frac{1}{\epsilon_i(a)} \tag{6}$$

control $\hat{\Delta}_s(a)$

The remaining thing is to give upper and lower bound for $\epsilon_t(a)$ when $s \geq t_{min}$. That is, to give upper and lower bound for $\hat{\Delta}_t(a)$.



1. $\Pr\left[\hat{\Delta}_t(a) \geq \Delta_t(a)\right] \leq \frac{1}{t^2}$
2. $\Pr\left[\hat{\Delta}_t(a) \leq \frac{1}{2}\Delta_t(a)\right] \leq ?$

conti

If $\hat{\Delta}_t(a) \leq \frac{1}{2}\Delta_t(a)$ happens, it means that $N_t(a)$ and $N_t(a^*)$ is not large enough. We claim that this would happen with just low probability.

Note that $\epsilon_t(a) = \min\{\frac{1}{K}, \sqrt{\frac{\ln K}{tK}}, \frac{\ln t}{t\hat{\Delta}_t(a)^2}\}$. Also, since we know $\hat{\Delta}_t(a) \leq \Delta_t(a)$ with high probability, we have $\epsilon_t(a) \geq \frac{\ln t}{t\Delta(a)^2}$ with high probability.

However, how could an arm with large exploration rate but being played just for few times ?

Therefore, we have

$$\forall a, t \geq t_{min}(a) \quad \Pr\left[N_t(a) \leq \frac{\ln t}{\Delta(a)^2}\right] \leq \frac{1}{t\Delta(a)^2}$$

With some further effort, we get (2) in the last page with low probability.

summary

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

- Initially, we explore each arm with same probability : $1/K$. The exploration rate then decrease adaptively.
- With enough exploration rate ($\epsilon_t(a) \geq \frac{\ln t}{t\Delta(a)^2}$ with high probability), every arm has been played for enough times, which leads to proper estimation of the gap $\hat{\Delta}_t(a)$ between the suboptimal arms and the best one.
- Once the arm is identified to be suboptimal, the exploration rate of it then decreases as $\frac{\ln t}{t\hat{\Delta}_t(a)^2}$
- Also, since the best arm has $\hat{\Delta}(a) \approx \Delta(a) = 0$, it eventually have the highest exploration rate.

Results of EXP3++ Algorithm

Theorem 1 (2014)

The regret of the EXP3++ in the adversarial regime for any t satisfies :

$$R(t) \leq 4\sqrt{Kt \log K}$$

Theorem 2(2017)

The expected regret of EXP3++ in the stochastic regime satisfies :

$$R(t) = \mathcal{O}\left(\sum_{a:\Delta(a)>0} \frac{(\ln t)^2}{\Delta(a)}\right) + \tilde{\mathcal{O}}\left(\sum_{a:\Delta(a)>0} \frac{K}{\Delta(a)^3}\right)$$

The latest result

"An Optimal Algorithm for Stochastic and Adversarial Bandits" – by Julian Zimmert, Yevgeny Seldin 2018

- They achieves the optimal (up to constants) finite time regret in both adversarial and stochastic multi-armed bandits. (pseudo-regret in the stochastic case and expected regret in the adversarial case)
- Auer and Chiang (2016) have shown the impossibility of achieving the optimal high-probability adversarial regret bounds simultaneously with the optimal stochastic pseudo-regret bounds.