Bandits for Exploration: Best Arm Identification and Discovery with Probabilistic Experts

Aurélien Garivier

Institut de Mathématiques de Toulouse

Imperial Probability Centre Multi-armed Bandits Session
October 31, 2014

Roadmap

The (stochastic) Multi-Armed Bandit Model

Environment K arms with parameters $\theta = (\theta_1, \dots, \theta_K)$ such that for any possible choice of arm $a_t \in \{1, \dots, K\}$ at time t, one receives the reward

$$X_t = X_{a_t,t}$$

where, for any $1 \leq a \leq K$ and $s \geq 1$, $X_{a,s} \sim \nu_a$, and the $(X_{a,s})_{a,s}$ are independent.

Reward distributions $\nu_a \in \mathcal{F}_a$ parametric family, or not. Examples: canonical exponential family, general bounded rewards

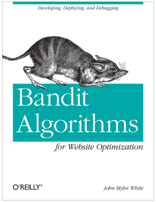
Example Bernoulli rewards: $\theta \in [0,1]^K$, $\nu_a = \mathcal{B}(\theta_a)$

Strategy The agent's actions follow a dynamical strategy $\pi = (\pi_1, \pi_2, \dots)$ such that

$$A_t = \pi_t(X_1, \dots, X_{t-1})$$

Real challenges

- Randomized clinical trials
  - original motivation since the 1930's
  - dynamic strategies can save resources
- Recommender systems:
  - advertisement
  - website optimization
  - news, blog posts, ...



- Computer experiments
  - large systems can be simulated in order to optimize some criterion over a set of parameters
  - but the simulation cost may be high, so that only few choices are possible for the parameters
- Games and planning (tree-structured options)

Performance Evaluation, Regret

Cumulated Reward $S_T = \sum_{t=1}^{T} X_t$

Our goal Choose π so as to maximize

$$\begin{aligned}\mathbb{E}\left[S_T\right] &= \sum_{t=1}^{T}\sum_{a=1}^{K}\mathbb{E}\Big[\mathbb{E}\left[X_t \mathbb{1}\{A_t = a\}|X_1,\dots,X_{t-1}\right]\Big] \\ &= \sum_{a=1}^{K}\mu_a \mathbb{E}\left[N_a^\pi(T)\right]\end{aligned}$$

where $N_a^\pi(T) = \sum_{t\le T}\mathbb{1}\{A_t = a\}$ is the number of draws of arm a up to time T, and $\mu_a = E(\nu_a)$.

Regret Minimization equivalent to minimizing

$$R_T = T\mu^* - \mathbb{E}\left[S_T\right] = \sum_{a:\mu_a<\mu^*}(\mu^* - \mu_a)\mathbb{E}\left[N_a^\pi(T)\right]$$

where $\mu^* \in \max\{\mu_a : 1 \le a \le K\}$

Asymptotically Optimal Strategies

- A strategy π is said to be consistent if, for any $(\nu_a)_a \in \mathcal{F}^K$,

$$\frac{1}{T}\mathbb{E}[S_T] \to \mu^*$$

- The strategy is efficient if for all $\theta \in [0,1]^K$ and all $\alpha > 0$,

$$R_T = o(T^\alpha)$$

- There are efficient strategies and we consider the best achievable asymptotic performance among efficient strategies

The Bound of Lai and Robbins

One-parameter reward distribution $\nu_a = \nu_{\theta_a}, \theta_a \in \Theta \subset \mathbb{R}$.

Theorem [Lai and Robbins, '85]

If π is an efficient strategy, then, for any $\theta \in \Theta^K$,

$$\liminf_{T\to\infty} \frac{R_T}{\log(T)} \geq \sum_{a:\mu_a<\mu^*} \frac{\mu^* - \mu_a}{\mathrm{KL}(\nu_a, \nu^*)}$$

where $\mathrm{KL}(\nu, \nu')$ denotes the Kullback-Leibler divergence

For example, in the Bernoulli case:

$$KL\big(\tilde{B}(p), \tilde{B}(q)\big) = d_{\text{BER}}(p, q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$$

The Bound of Burnetas and Katehakis

More general reward distributions $\nu_a \in \mathcal{F}_a$

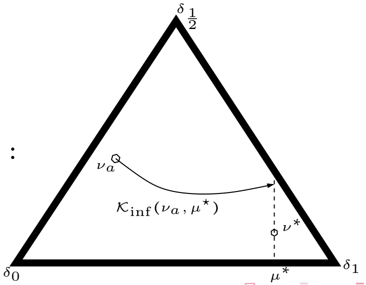
Theorem [Burnetas and Katehakis, '96]

If π is an efficient strategy, then, for any $\theta \in [0,1]^K$,

$$\liminf_{T\to\infty} \frac{R_T}{\log(T)} \geq \sum_{a:\mu_a<\mu^*} \frac{\mu^* - \mu_a}{K_{inf}(\nu_a, \mu^*)}$$

where

$$K_{inf}(\nu_a, \mu^*) = \inf \left\{ K(\nu_a, \nu') : \nu' \in \mathcal{F}_a, E(\nu') \geq \mu^* \right\}$$



Upper Confidence Bound Strategies

UCB [Lai&Robins '85; Agrawal '95; Auer&al '02]

- Construct an upper confidence bound for the expected reward of each arm:

$$\underbrace{\frac{S_a(t)}{N_a(t)}}_{\text{estimated reward}} + \underbrace{\sqrt{\frac{\log(t)}{2N_a(t)}}}_{\text{exploration bonus}}$$

- Choose the arm with the highest UCB

- It is an *index strategy* [Gittins '79]
- Its behavior is easily interpretable and intuitively appealing

Performance of UCB

For rewards in $[0, 1]$, the regret of UCB is upper-bounded as

$$E[R_T] = O(\log(T))$$

(finite-time regret bound) and

$$\limsup_{T\to\infty} \frac{\mathbb{E}[R_T]}{\log(T)} \leq \sum_{a:\mu_a<\mu^*} \frac{1}{2(\mu^* - \mu_a)}$$

Yet, in the case of Bernoulli variables, the rhs. is greater than suggested by the bound by Lai & Robbins

Many variants have been suggested to incorporate an estimate of the variance in the exploration bonus (e.g., [Audibert&al '07])

The KL-UCB algorithm

The KL-UCB Algorithm, Annals of Statistics 2013
joint work with O. Cappé, O-A. Maillard, R. Munos, G. Stoltz

Parameters: An operator $\Pi_{\mathcal{F}} : \mathfrak{M}_1(\mathcal{S}) \to \mathcal{F}$; a non-decreasing function $f : \mathbb{N} \to \mathbb{R}$

Initialization: Pull each arm of $\{1, \ldots, K\}$ once

for $t = K$ to $T - 1$ **do**

compute for each arm a the quantity

$$U_a(t) = \sup\left\{E(\nu) : \quad \nu \in \mathcal{F} \quad \text{and} \quad KL\Big(\Pi_{\mathcal{F}}\big(\hat{\nu}_a(t)\big), \nu\Big) \leq \frac{f(t)}{N_a(t)}\right\}$$

pick an arm $\quad A_{t+1} \in \underset{a \in \{1,\ldots,K\}}{\arg\max}\; U_a(t)$

end for

Exponential Family Rewards

- Assume that $\mathcal{F}_a = \mathcal{F} =$ *canonical exponential family*, i.e. such that the pdf of the rewards is given by

$$p_{\theta_a}(x) = \exp\big(x\theta_a - b(\theta_a) + c(x)\big), \quad 1 \le a \le K$$

for a parameter $\theta \in \mathbb{R}^K$, expectation $\mu_a = \dot{b}(\theta_a)$

-
$$U_a(t) = \sup\left\{\mu \in \overline{I} : \quad d\big(\hat{\mu}_a(t), \mu\big) \le \frac{f(t)}{N_a(t)}\right\}$$

- For instance,
  - for Bernoulli rewards:

$$d_{\text{BER}}(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$$

  - for exponential rewards $p_{\theta_a}(x) = \theta_a e^{-\theta_a x}$:

$$d_{\exp}(u,v) = u - v + u\log\frac{u}{v}$$

- The analysis is generic and yields a non-asymptotic regret bound optimal in the sense of Lai and Robbins.

Regret bound

Theorem: Assume that all arms belong to a canonical, regular, exponential family $\mathcal{F} = \{\nu_\theta : \theta \in \Theta\}$ of probability distributions indexed by its natural parameter space $\Theta \subseteq \mathbb{R}$. Then, with the choice $f(t) = \log(t) + 3\log\log(t)$ for $t \geq 3$, the number of draws of any suboptimal arm a is upper bounded for any horizon $T \geq 3$ as

$$\mathbb{E}[N_a(T)] \leq \frac{\log(T)}{d(\mu_a, \mu^\star)} + 2\sqrt{\frac{2\pi\sigma_{a,\star}^2 \left(d'(\mu_a, \mu^\star)\right)^2}{\left(d(\mu_a, \mu^\star)\right)^3}}\sqrt{\log(T) + 3\log(\log(T))}$$
$$+ \left(4e + \frac{3}{d(\mu_a, \mu^\star)}\right)\log(\log(T)) + 8\sigma_{a,\star}^2 \left(\frac{d'(\mu_a, \mu^\star)}{d(\mu_a, \mu^\star)}\right)^2 + 6\,,$$

where $\sigma_{a,\star}^2 = \max\left\{\mathrm{Var}(\nu_\theta) : \ \mu_a \leq E(\nu_\theta) \leq \mu^\star\right\}$ and where $d'(\cdot, \mu^\star)$ denotes the derivative of $d(\cdot, \mu^\star)$.

Results: Two-Arm Scenario

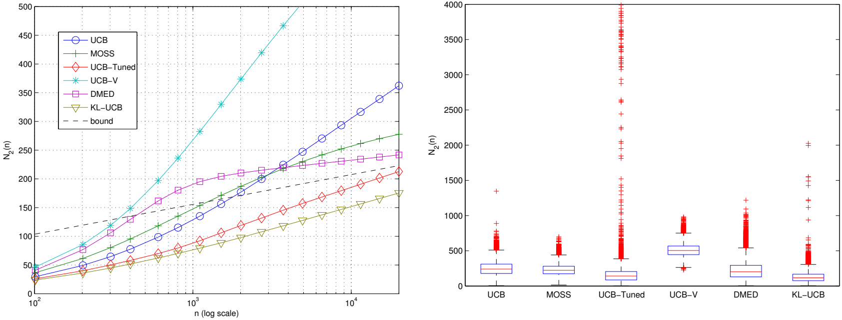


Figure: Performance of various algorithms when $\theta = (0.9, 0.8)$. Left: average number of draws of the sub-optimal arm as a function of time. Right: box-and-whiskers plot for the number of draws of the sub-optimal arm at time $T = 5,000$. Results based on $50,000$ independent replications

Non-parametric setting

- Rewards are only assumed to be bounded (say in $[0, 1]$)
- Need for an estimation procedure
  - with non-asymptotic guarantees
  - efficient in the sense of Stein / Bahadur

$\Longrightarrow$ Idea 1: use d_{BER} (Hoeffding)

$\Longrightarrow$ Idea 2: Empirical Likelihood [Owen '01]

- Bad idea: use Bernstein / Bennett

First idea: use d_{BER}

Idea: rescale to $[0, 1]$, and take the divergence d_{BER}.

$\longrightarrow$ because Bernoulli distributions maximize deviations among bounded variables with given expectation:

Lemma (Hoeffding '63)

Let X denote a random variable such that $0 \le X \le 1$ and denote by $\mu = \mathbb{E}[X]$ its mean. Then, for any $\lambda \in \mathbb{R}$,

$$E\left[\exp(\lambda X)\right] \le 1 - \mu + \mu \exp(\lambda) .$$

This fact is well-known for the variance, but also true for all exponential moments and thus for Cramer-type deviation bounds

Regret Bound for kl-UCB

Theorem

With the divergence d_{BER}, for all $T > 3$,

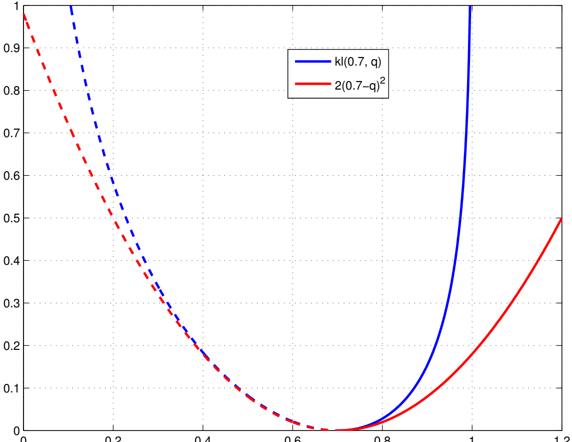
$$\mathbb{E}\left[N_a(T)\right] \leq \frac{\log(T)}{d_{\text{BER}}(\mu_a, \mu^\star)} + \frac{\sqrt{2\pi}\log\left(\frac{\mu^\star(1-\mu_a)}{\mu_a(1-\mu^\star)}\right)}{\left(d_{\text{BER}}(\mu_a, \mu^\star)\right)^{3/2}} \sqrt{\log(T) + 3\log\left(\log(T)\right)}$$
$$+ \left(4e + \frac{3}{d_{\text{BER}}(\mu_a, \mu^\star)}\right)\log\left(\log(T)\right) + \frac{2\left(\log\left(\frac{\mu^\star(1-\mu_a)}{\mu_a(1-\mu^\star)}\right)\right)^2}{\left(d_{\text{BER}}(\mu_a, \mu^\star)\right)^2} + 6\,.$$

- kl-UCB satisfies an improved logarithmic finite-time regret bound
- Besides, it is asymptotically optimal in the Bernoulli case

Comparison to UCB

KL-UCB addresses exactly the same problem as UCB, with the same generality, but it has always a smaller regret as can be seen from Pinsker's inequality

$$d_{\text{BER}}(\mu_1, \mu_2) \geq 2(\mu_1 - \mu_2)^2$$

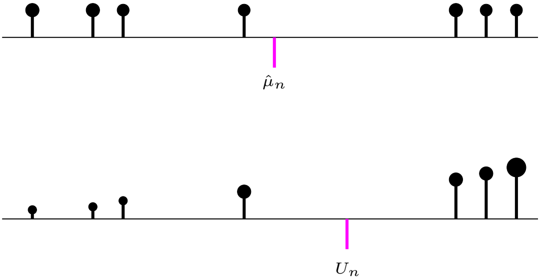


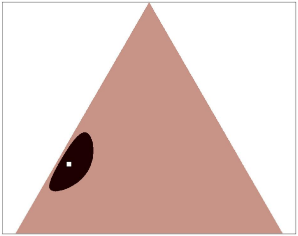
Idea 2: Empirical Likelihood

$$U(\hat{\nu}_n, \epsilon) = \sup\Big\{E(\nu') : \nu' \in \mathfrak{M}_1\big(\operatorname{Supp}(\hat{\nu}_n)\big) \text{ and } \operatorname{KL}(\hat{\nu}_n, \nu') \le \epsilon\Big\}$$

or, rather, *modified Empirical Likelihood*:

$$U(\hat{\nu}_n, \epsilon) = \sup\Big\{E(\nu') : \nu' \in \mathfrak{M}_1\big(\operatorname{Supp}(\hat{\nu}_n) \cup \{1\}\big) \text{ and } \operatorname{KL}(\hat{\nu}_n, \nu') \le \epsilon\Big\}$$





Coverage properties of the modified EL confidence bound

Proposition: Let $\nu_0 \in \mathfrak{M}_1([0,1])$ with $E(\nu_0) \in (0,1)$ and let $X_1, \ldots, X_n$ be independent random variables with common distribution $\nu_0 \in \mathfrak{M}_1\big([0,1]\big)$, not necessarily with finite support. Then, for all $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}\big\{U(\hat{\nu}_n, \epsilon) \le E(\nu_0)\big\} &\le \mathbb{P}\Big\{K_{inf}\big(\hat{\nu}_n,\, E(\nu_0)\big) \ge \epsilon\Big\}\\ &\le e(n+2)\exp(-n\epsilon)\;.\end{aligned}$$

Remark: For $\{0,1\}$–valued observations, it is readily seen that $U(\hat{\nu}_n, \epsilon)$ boils down to the upper-confidence bound above.
$\Longrightarrow$ This proposition is at least not always optimal: the presence of the factor n in front of the exponential $\exp(-n\epsilon)$ term is questionable.

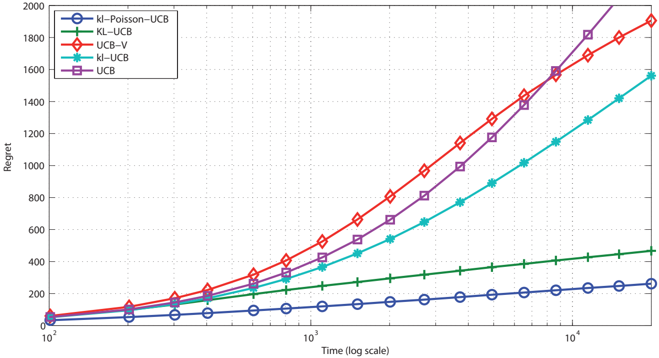
Regret bound

Theorem: Assume that $\mathcal{F}$ is the set of finitely supported probability distributions over $\mathcal{S} = [0,1]$, that $\mu_a > 0$ for all arms a and that $\mu^\star < 1$. There exists a constant $M(\nu_a, \mu^\star) > 0$ only depending on ν_a and $\mu^\star$ such that, with the choice $f(t) = \log(t) + \log\big(\log(t)\big)$ for $t \geq 2$, for all $T \geq 3$:

$$
\begin{aligned}
\mathbb{E}\big[N_a(T)\big] \;\leq\; & \frac{\log(T)}{K_{inf}\big(\nu_a,\mu^\star\big)} + \frac{36}{(\mu^\star)^4}\big(\log(T)\big)^{4/5}\log\big(\log(T)\big) \\
& + \left(\frac{72}{(\mu^\star)^4} + \frac{2\mu^\star}{(1-\mu^\star)\,K_{inf}\big(\nu_a,\mu^\star\big)^2}\right)\big(\log(T)\big)^{4/5} \\
& + \frac{(1-\mu^\star)^2\,M(\nu_a,\mu^\star)}{2(\mu^\star)^2}\big(\log(T)\big)^{2/5} \\
& + \frac{\log\big(\log(T)\big)}{K_{inf}\big(\nu_a,\mu^\star\big)} + \frac{2\mu^\star}{(1-\mu^\star)\,K_{inf}\big(\nu_a,\mu^\star\big)^2} + 4\,.
\end{aligned}
$$

Example: truncated Poisson rewards

- for each arm $1 \leq a \leq 6$ is associated with ν_a, a Poisson distribution with expectation $(2+a)/4$, truncated at 10.
- $N = 10,000$ Monte-Carlo replications on an horizon of $T = 20,000$ steps.



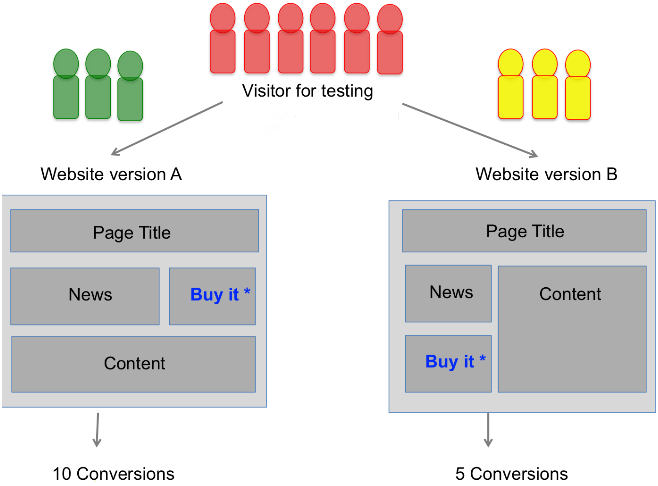
Take-home message on classical bandit algorithms

1. Use kl-UCB rather than UCB-1 or UCB-2
2. Use KL-UCB if speed is not a problem
3. todo: improve on the deviation bounds, address general non-parametric families of distributions
4. Alternative: Bayesian-flavored methods:
   - Bayes-UCB [Kaufmann, Cappé, G.]
   - Thompson sampling [Kaufmann & al.]

Roadmap

Motivation

On the Complexity of Best Arm Identification in Multi-Armed Bandit Models, ArXiv (COLT 2014)
joint work with O. Cappé and E. Kaufmann



Our goal

Improve performance:

- ➜ fixed number of test users $-$ $>$ smaller probability of error
- ➜ fixed probability of error $-$ $>$ fewer test users

Tools: sequential allocation and stopping

The model

A two-armed bandit model is

- a set $\nu = (\nu_1, \nu_2)$ of two probability distributions ('arms') with respective means μ_1 and μ_2
- $a^* = \text{argmax}_a\ \mu_a$ is the (unknown) best am

To find the best arm, an agent interacts with the bandit model with

- a *sampling rule* $(A_t)_{t\in\mathbb{N}}$ where $A_t \in \{1, 2\}$ is the arm chosen at time t (based on past observations) $- >$ a sample $Z_t \sim \nu_{A_t}$ is observed
- a *stopping rule* τ indicating when he stops sampling the arms
- a *recommendation rule* $\hat{a}_\tau \in \{1, 2\}$ indicating which arm he thinks is best (at the end of the interaction)

In classical A/B Testing, the sampling rule A_t is uniform on $\{1, 2\}$ and the stopping rule $\tau = t$ is fixed in advance.

Two possible goals

The agent's goal is to design a strategy $\mathcal{A} = ((A_t), \tau, \hat{a}_\tau)$ satisfying

Fixed-budget setting	**Fixed-confidence setting**
$\tau = t$	$\mathbb{P}_\nu(\hat{a}_\tau \neq a^*) \leq \delta$
$p_t(\nu) := \mathbb{P}_\nu(\hat{a}_t \neq a^*)$ as small as possible	$\mathbb{E}_\nu[\tau]$ as small as possible

An algorithm using uniform sampling is

Fixed-budget setting	**Fixed-confidence setting**
a classical test of $(\mu_1 > \mu_2)$ against $(\mu_1 < \mu_2)$ based on t samples	a sequential test of $(\mu_1 > \mu_2)$ against $(\mu_1 < \mu_2)$ with probability of error uniformly bounded by δ

[Siegmund 85]: sequential tests can save samples !

The complexities of best-arm identification

For a class $\mathcal{M}$ bandit models, algorithm $\mathcal{A} = ((A_t), \tau, \hat{a}_\tau)$ is...

Fixed-budget setting	**Fixed-confidence setting**
consistent on $\mathcal{M}$ if $\forall \nu \in \mathcal{M}, \ p_t(\nu) = \mathbb{P}_\nu(\hat{a}_t \neq a^*) \underset{t\to\infty}{\longrightarrow} 0$	δ-PAC on $\mathcal{M}$ if $\forall \nu \in \mathcal{M}, \ \mathbb{P}_\nu(\hat{a}_\tau \neq a^*) \leq \delta$

From the literature

$p_t(\nu) \simeq \exp\left(-\frac{t}{CH(\nu)}\right)$	$\mathbb{E}_\nu[\tau] \simeq C'H'(\nu)\log\frac{1}{\delta}$
[Audibert et al. 10],[Bubeck et al. 11] [Bubeck et al. 13],...	[Mannor Tsitsilis 04],[Even-Dar al. 06] [Kalanakrishnan et al.12],...

Two complexities

$\kappa_{\mathrm{B}}(\nu) = \inf_{\mathcal{A}\text{ cons.}} \left(\limsup_{t\to\infty} -\frac{1}{t}\log p_t(\nu)\right)^{-1}$	$\kappa_{\mathrm{C}}(\nu) = \inf_{\mathcal{A}\ \delta-\text{PAC}} \limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)}$
for a probability of error $\leq \delta$, budget $t \simeq \kappa_B(\nu)\log\frac{1}{\delta}$	for a probability of error $\leq \delta$ $\mathbb{E}_\nu[\tau] \simeq \kappa_C(\nu)\log\frac{1}{\delta}$

Changes of distribution

New formulation for a change of distribution

Let ν and ν' be two bandit models. Let N_1 (resp. N_2) denote the total number of draws of arm 1 (resp. arm 2) by algorithm $\mathcal{A}$). For any $A \in \mathcal{F}_\tau$ such that $0 < \mathbb{P}_\nu(A) < 1$

$$\mathbb{E}_\nu[N_1]\text{KL}(\nu_1, \nu_1') + \mathbb{E}_\nu[N_2]\text{KL}(\nu_2, \nu_2') \geq d_{\text{BER}}\big(\mathbb{P}_\nu(A), \mathbb{P}_{\nu'}(A)\big),$$

where $d_{\text{BER}}(x, y) = x\log(x/y) + (1 - x)\log\big((1 - x)/(1 - y)\big)$.

General lower bounds

Theorem 1

Let $\mathcal{M}$ be a class of two armed bandit models that are continuously parametrized by their means. Let $\nu = (\nu_1, \nu_2) \in \mathcal{M}$.

Fixed-budget setting	Fixed-confidence setting
any consistent algorithm satisfies $\limsup_{t\to\infty} -\frac{1}{t}\log p_t(\nu) \leq \mathrm{K}^*(\nu_1, \nu_2)$ with $\mathrm{K}^*(\nu_1,\nu_2) = \mathrm{KL}(\nu^*, \nu_1) = \mathrm{KL}(\nu^*, \nu_2)$	any δ-PAC algorithm satisfies $\mathbb{E}_\nu[\tau] \geq \frac{1}{\mathrm{K}_*(\nu_1,\nu_2)}\log\left(\frac{1}{2\delta}\right)$ with $\mathrm{K}_*(\nu_1,\nu_2) = \mathrm{KL}(\nu_1, \nu_*) = \mathrm{KL}(\nu_2, \nu_*)$
Thus, $\kappa_B(\nu) \geq \frac{1}{\mathrm{K}^*(\nu_1,\nu_2)}$	Thus, $\kappa_C(\nu) \geq \frac{1}{\mathrm{K}_*(\nu_1,\nu_2)}$

Fixed-budget setting

For fixed (known) values σ_1, σ_2, we consider Gaussian bandit models

$$\mathcal{M} = \left\{\nu = \left(\mathcal{N}\left(\mu_1, \sigma_1^2\right), \mathcal{N}\left(\mu_2, \sigma_2^2\right)\right) : (\mu_1, \mu_2) \in \mathbb{R}^2, \mu_1 \neq \mu_2\right\}$$

- Theorem 1:

$$\kappa_B(\nu) \geq \frac{2(\sigma_1 + \sigma_2)^2}{(\mu_1 - \mu_2)^2}$$

- A strategy allocating $t_1 = \left\lceil \frac{\sigma_1}{\sigma_1+\sigma_2} t \right\rceil$ samples to arm 1 and $t_2 = t - t_1$ samples to arm 1, and recommending the empirical best satisfies

$$\liminf_{t\to\infty} -\frac{1}{t} \log p_t(\nu) \geq \frac{(\mu_1 - \mu_2)^2}{2(\sigma_1 + \sigma_2)^2}$$

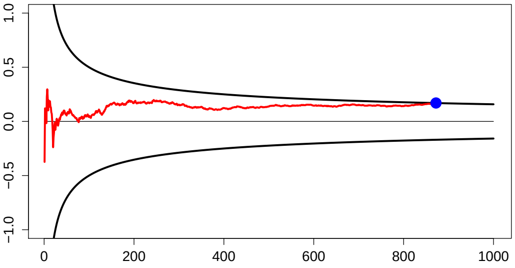
$$\boxed{\kappa_B(\nu) = \frac{2(\sigma_1 + \sigma_2)^2}{(\mu_1 - \mu_2)^2}}$$

Fixed-confidence setting: Algorithm

The α-Elimination algorithm with exploration rate $\beta(t,\delta)$

- chooses A_t in order to keep a proportion $N_1(t)/t \simeq \alpha$
- if $\hat{\mu}_a(t)$ is the empirical mean of rewards obtained from a up to time t, $\sigma_t^2(\alpha) = \sigma_1^2/\lceil \alpha t \rceil + \sigma_2^2/(t - \lceil \alpha t \rceil)$,

$$\tau = \inf\left\{t \in \mathbb{N} : |\hat{\mu}_1(t) - \hat{\mu}_2(t)| > \sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)}\right\}$$



- recommends the empirical best arm $\hat{a}_\tau = \text{argmax}_a \hat{\mu}_a(\tau)$

Fixed-confidence setting: Results

- From Theorem 1:

$$\mathbb{E}_\nu[\tau] \geq \frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2} \log\left(\frac{1}{2\delta}\right)$$

- $\frac{\sigma_1}{\sigma_1+\sigma_2}$-Elimination with $\beta(t,\delta) = \log\frac{t}{\delta} + 2\log\log(6t)$ is δ-PAC and

$$\forall \epsilon > 0, \quad \mathbb{E}_\nu[\tau] \leq (1+\epsilon)\frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2} \log\left(\frac{1}{2\delta}\right) + \underset{\delta\to 0}{o_\epsilon}\left(\log\frac{1}{\delta}\right)$$

$$\boxed{\kappa_C(\nu) = \frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}}$$

Gaussian distributions: Conclusions

For any two fixed values of σ_1 and σ_2,

$$\boxed{\kappa_B(\nu) = \kappa_C(\nu) = \frac{2(\sigma_1 + \sigma_2)^2}{(\mu_1 - \mu_2)^2}}$$

If the variances are equal, $\sigma_1 = \sigma_2 = \sigma$,

$$\kappa_B(\nu) = \kappa_C(\nu) = \frac{8\sigma^2}{(\mu_1 - \mu_2)^2}$$

- uniform sampling is optimal only when $\sigma_1 = \sigma_2$
- $1/2$-Elimination is δ-PAC for a smaller exploration rate $\beta(t, \delta) \simeq \log(\log(t)/\delta)$

Lower bounds for Bernoulli bandit models

$$\mathcal{M} = \{\nu = (\mathcal{B}(\mu_1), \mathcal{B}(\mu_2)) : (\mu_1, \mu_2) \in]0;1[^2, \mu_1 \neq \mu_2\},$$

shorthand: $\mathrm{K}(\mu, \mu') = \mathrm{KL}\left(\mathcal{B}(\mu), \mathcal{B}(\mu')\right)$.

Fixed-budget setting	**Fixed-confidence setting**
any consistent algorithm satisfies $\limsup_{t\to\infty} -\frac{1}{t}\log p_t(\nu) \leq \mathrm{K}^*(\mu_1, \mu_2)$ (Chernoff information)	any δ-PAC algorithm satisfies $\mathbb{E}_\nu[\tau] \geq \frac{1}{\mathrm{K}_*(\mu_1,\mu_2)}\log\left(\frac{1}{2\delta}\right)$

$$\boxed{\mathrm{K}^*(\mu_1, \mu_2) > \mathrm{K}_*(\mu_1, \mu_2)}$$

Algorithms using uniform sampling

	For any consistent...	For any δ-PAC...
... algorithm	$p_t(\nu) \gtrsim e^{-\mathrm{K}^*(\mu_1,\mu_2)t}$	$\frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)} \gtrsim \frac{1}{\mathrm{K}_*(\mu_1,\mu_2)}$
... algorithm using uniform sampling	$p_t(\nu) \gtrsim e^{-\frac{\mathrm{K}(\overline{\mu},\mu_1)+\mathrm{K}(\overline{\mu},\mu_2)}{2}t}$ with $\overline{\mu} = f(\mu_1,\mu_2)$	$\frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)} \gtrsim \frac{2}{\mathrm{K}(\mu_1,\underline{\mu})+\mathrm{K}(\mu_2,\underline{\mu})}$ with $\underline{\mu} = \frac{\mu_1+\mu_2}{2}$

Remark: Quantities in the same column appear to be close from one another

$\Rightarrow$ **Binary rewards: uniform sampling close to optimal**

Algorithms using uniform sampling

	For any consistent...	For any δ-PAC...
... algorithm	$p_t(\nu) \simeq e^{-\mathrm{K}^*(\mu_1,\mu_2)t}$	$\frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)} \gtrsim \frac{1}{\mathrm{K}_*(\mu_1,\mu_2)}$
... algorithm using uniform sampling	$p_t(\nu) \simeq e^{-\frac{\mathrm{K}(\overline{\mu},\mu_1)+\mathrm{K}(\overline{\mu},\mu_2)}{2}t}$ with $\overline{\mu} = f(\mu_1,\mu_2)$	$\frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)} \gtrsim \frac{2}{\mathrm{K}(\mu_1,\underline{\mu})+\mathrm{K}(\mu_2,\underline{\mu})}$ with $\underline{\mu} = \frac{\mu_1+\mu_2}{2}$

Remark: Quantities in the same column appear to be close from one another

$\Rightarrow$ **Binary rewards: uniform sampling close to optimal**

Fixed-budget setting

We show that

$$\boxed{\kappa_B(\nu) = \frac{1}{\mathrm{K}^*(\mu_1, \mu_2)}}$$

(matching algorithm not implementable in practice)

The algorithm using uniform sampling and recommending the empirical best arm is preferable (and very close to optimal)

Fixed-confidence setting

δ-PAC algorithms using uniform sampling satisfy

$$\frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)} \geq \frac{1}{I_*(\nu)} \quad \text{with} \quad I_*(\nu) = \frac{\mathrm{K}\left(\mu_1, \frac{\mu_1+\mu_2}{2}\right) + \mathrm{K}\left(\mu_2, \frac{\mu_1+\mu_2}{2}\right)}{2}.$$

The algorithm using uniform sampling and

$$\tau = \inf\left\{t \in 2\mathbb{N}^* : |\hat{\mu}_1(t) - \hat{\mu}_2(t)| > \log\frac{\log(t)+1}{\delta}\right\}$$

is δ-PAC but not optimal: $\frac{\mathbb{E}[\tau]}{\log(1/\delta)} \simeq \frac{2}{(\mu_1-\mu_2)^2} > \frac{1}{I_*(\nu)}$.

A better stopping rule NOT based on the difference of empirical means

$$\tau = \inf\left\{t \in 2\mathbb{N}^* : tI_*(\hat{\mu}_1(t), \hat{\mu}_2(t)) > \log\frac{\log(t)+1}{\delta}\right\}$$

Bernoulli distributions: Conclusion

Regarding the complexities:

- $\kappa_B(\nu) = \frac{1}{\mathrm{K}^*(\mu_1,\mu_2)}$
- $\kappa_C(\nu) \geq \frac{1}{\mathrm{K}_*(\mu_1,\mu_2)} > \frac{1}{\mathrm{K}^*(\mu_1,\mu_2)}$

Thus

$$\boxed{\kappa_C(\nu) > \kappa_B(\nu)}$$

Regarding the algorithms

- There is not much to gain by departing from uniform sampling
- In the fixed-confidence setting, a sequential test based on the difference of the empirical means is no longer optimal

Conclusion on Best Arm Identification

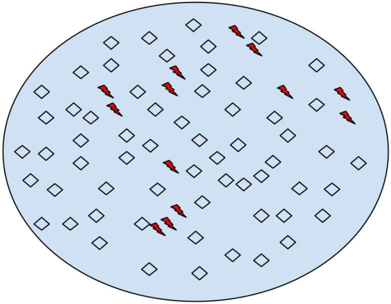
- the complexities $\kappa_B(\nu)$ and $\kappa_C(\nu)$ are not always equal (and feature some different informational quantities)
- for Bernoulli distributions and Gaussian with similar variances, strategies using uniform sampling are (almost) optimal
- strategies using random stopping do not necessarily lead to a saving in terms of the number of sample used
- Generalization to m best arms identification among K arms

Roadmap

The model

Optimal Discovery with Probabilistic Expert Advice: Finite Time Analysis and Macroscopic Optimality, JMLR 2013
joint work with S. Bubeck and D. Ernst

- Subset $A \subset \mathcal{X}$ of important items
- $|\mathcal{X}| \gg 1$, $|A| \ll |\mathcal{X}|$
- Access to $\mathcal{X}$ only by probabilistic experts $(P_i)_{1 \leq i \leq K}$: sequential independent draws

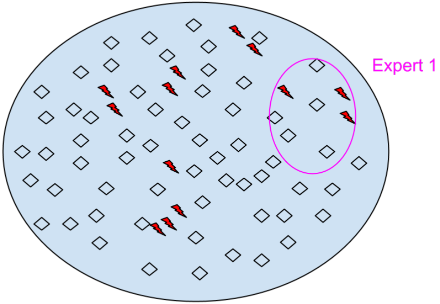


Goal: discover rapidly the elements of A

The model

Optimal Discovery with Probabilistic Expert Advice: Finite Time Analysis and Macroscopic Optimality, JMLR 2013
joint work with S. Bubeck and D. Ernst

- Subset $A \subset \mathcal{X}$ of important items
- $|\mathcal{X}| \gg 1$, $|A| \ll |\mathcal{X}|$
- Access to $\mathcal{X}$ only by probabilistic experts $(P_i)_{1 \le i \le K}$: sequential independent draws

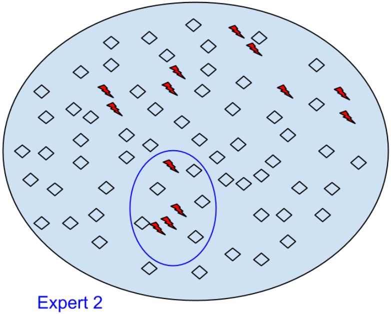


Goal: discover rapidly the elements of A

The model

Optimal Discovery with Probabilistic Expert Advice: Finite Time Analysis and Macroscopic Optimality, JMLR 2013
joint work with S. Bubeck and D. Ernst

- Subset $A \subset \mathcal{X}$ of important items
- $|\mathcal{X}| \gg 1$, $|A| \ll |\mathcal{X}|$
- Access to $\mathcal{X}$ only by probabilistic experts $(P_i)_{1 \leq i \leq K}$: sequential independent draws

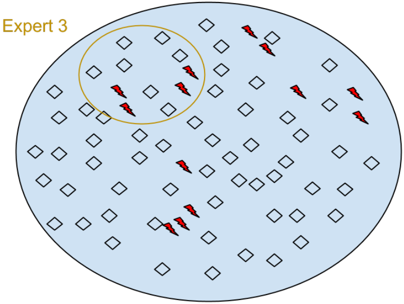


Goal: discover rapidly the elements of A

The model

Optimal Discovery with Probabilistic Expert Advice: Finite Time Analysis and Macroscopic Optimality, JMLR 2013
joint work with S. Bubeck and D. Ernst

- Subset $A \subset \mathcal{X}$ of important items
- $|\mathcal{X}| \gg 1$, $|A| \ll |\mathcal{X}|$
- Access to $\mathcal{X}$ only by probabilistic experts $(P_i)_{1 \le i \le K}$: sequential independent draws



Goal: discover rapidly the elements of A

Optimal Exploration with Probabilistic Expert Advice

Search space : $A \subset \Omega$ discrete set

Probabilistic experts : $P_i \in \mathfrak{M}_1(\Omega)$ for $i \in \{1, \dots, K\}$

Requests : at time t, calling expert I_t yields a realization of $X_t = X_{I_t,t}$ independent with law P_a

Goal : find as many distinct elements of A as possible with few requests :

$$F_n = \mathrm{Card}\left(A \cap \{X_1, \dots, X_n\}\right)$$

Goal

At each time step $t = 1, 2, \ldots$:

- pick an index $I_t = \pi_t\big(I_1, Y_1, \ldots, I_{s-1}, Y_{s-1}\big) \in \{1, \ldots, K\}$ according to past observations
- observe $Y_t = X_{I_t, n_{I_t,t}} \sim P_{I_t}$, where

$$n_{i,t} = \sum_{s \le t} \mathbb{1}\{I_s = i\}$$

Goal: design the strategy $\pi = (\pi_t)_t$ so as to maximize the number of important items found after t requests

$$F^\pi(t) = \Big|A \cap \big\{Y_1, \ldots, Y_t\big\}\Big|$$

Assumption: non-intersecting supports

$$A \cap \mathrm{supp}(P_i) \cap \mathrm{supp}(P_j) = \varnothing \text{ for } i \neq j$$

Is it a Bandit Problem ?

It looks like a bandit problem. . .

- sequential choices among K options
- want to maximize cumulative rewards
- exploration vs exploitation dilemma

. . . but it is **not a bandit problem** !

- rewards are not i.i.d.
- destructive rewards: no interest to observe twice the same important item
- all strategies eventually equivalent

The oracle strategy

Proposition: Under the non-intersecting support hypothesis, the greedy oracle strategy selecting the expert with highest 'missing mass'

$$I_t^* \in \arg\max_{1 \le i \le K} P_i \left(A \setminus \{Y_1, \dots, Y_t\} \right)$$

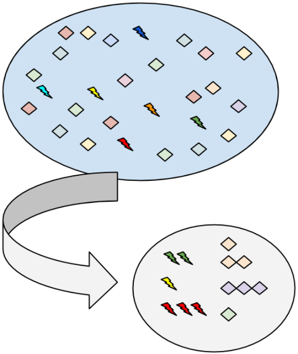
is optimal: for every possible strategy π, $\mathbb{E}\big[F^\pi(t)\big] \le \mathbb{E}\big[F^*(t)\big]$.

Remark: the proposition if false if the supports may intersect

$\Longrightarrow$ **estimate the "missing mass of important items"!**

Missing mass estimation

Let us first focus on one expert i: $P = P_i, X_n = X_{i,n}$



$X_1, \ldots, X_n$ independent draws of P

$$O_n(x) = \sum_{m=1}^{n} \mathbb{1}\{X_m = x\}$$

How to 'estimate' the total mass of the *unseen* important items

$$R_n = \sum_{x \in A} P(x) \mathbb{1}\{O_n(x) = 0\} \ ?$$

The Good-Turing Estimator

Idea: use the **hapaxes** = items seen only once (linguistic)

$$\hat{R}_n = \frac{U_n}{n}, \quad \text{where } U_n = \sum_{x \in A} \mathbb{1}\{O_n(x) = 1\}$$

Lemma [Good '53]: For *every* distribution P,

$$0 \leq \mathbb{E}\big[\hat{R}_n\big] - \mathbb{E}\big[R_n\big] \leq \frac{1}{n}$$

Proposition: With probability at least $1 - \delta$ for *every* P,

$$\hat{R}_n - \frac{1}{n} - (1 + \sqrt{2})\sqrt{\frac{\log(4/\delta)}{n}} \leq R_n \leq \hat{R}_n + (1 + \sqrt{2})\sqrt{\frac{\log(4/\delta)}{n}}$$

See [McAllester and Schapire '00, McAllester and Ortiz '03]:

- deviations of $\hat{R}_n$: McDiarmid's inequality
- deviations of R_n: negative association

The Good-UCB algorithm [Bubeck, Ernst & G.]

Optimistic algorithm based on Good-Turing's estimator :

$$I_{t+1} = \arg\max_{i \in \{1,\dots,K\}} \left\{ \frac{H_i(t)}{N_i(t)} + c\sqrt{\frac{\log(t)}{N_i(t)}} \right\}$$

- $N_i(t)$ = number of draws of P_i up to time t
- $H_i(t)$ = number of elements of A seen exactly once thanks to P_i
- c = tuning parameter

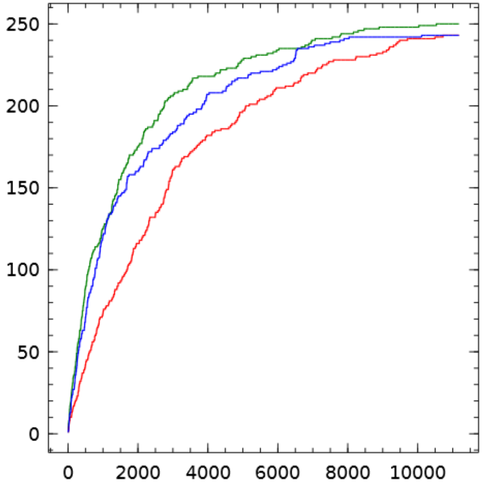
Classical analysis

Theorem: For any $t \geq 1$, under the non-intersecting support assumption, Good-UCB (with constant $C = (1+\sqrt{2})\sqrt{3}$) satisfies

$$\mathbb{E}\left[F^*(t) - F^{UCB}(t)\right] \leq 17\sqrt{Kt\log(t)} + 20\sqrt{Kt} + K + K\log(t/K)$$

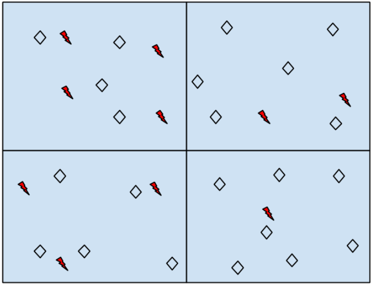
Remark: Usual result for bandit problem, but not-so-simple analysis

A Typical Run of Good-UCB



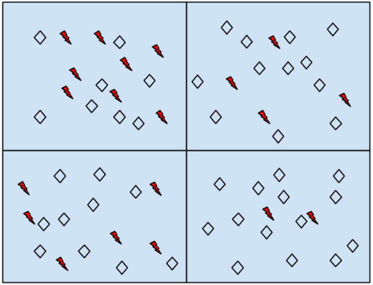
The macroscopic limit

- Restricted framework: $P_i = \mathcal{U}\{1, \ldots, N\}$
- $N \to \infty$
- $|A \cap \mathrm{supp}(P_i)|/N \to q_i \in (0,1)$, $q = \sum_i q_i$



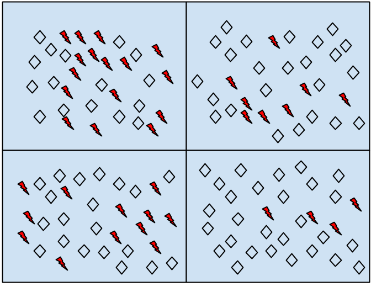
The macroscopic limit

- Restricted framework: $P_i = \mathcal{U}\{1, \ldots, N\}$
- $N \to \infty$
- $|A \cap \mathrm{supp}(P_i)|/N \to q_i \in (0,1)$, $q = \sum_i q_i$



The macroscopic limit

- Restricted framework: $P_i = \mathcal{U}\{1, \ldots, N\}$
- $N \to \infty$
- $|A \cap \mathrm{supp}(P_i)|/N \to q_i \in (0,1)$, $q = \sum_i q_i$



The Oracle behaviour

The limiting discovery process of the Oracle strategy is *deterministic*

Proposition: For every $\lambda \in (0, q_1)$, for every sequence $(\lambda^N)_N$ converging to λ as N goes to infinity, almost surely

$$\lim_{N \to \infty} \frac{T_*^N(\lambda^N)}{N} = \sum_i \left(\log \frac{q_i}{\lambda} \right)_+$$

Oracle vs. uniform sampling

Oracle: The proportion of important items not found after Nt draws tends to

$$q - F^*(t) = I(t)\underline{q}_{I(t)} \exp\left(-t/I(t)\right) \leq K\underline{q}_K \exp(-t/K)$$

with $\underline{q}_K = \left(\prod_{i=1}^K q_i\right)^{1/K}$ the geometric mean of the $(q_i)_i$.

Uniform: The proportion of important items not found after Nt draws tends to $K\bar{q}_K \exp(-t/K)$

$\Longrightarrow$ Asymptotic ratio of efficiency

$$\rho(q) = \frac{\bar{q}_K}{\underline{q}_K} = \frac{\frac{1}{K}\sum_{i=1}^k q_i}{\left(\prod_{i=1}^k q_i\right)^{1/K}} \geq 1$$

larger if the $(q_i)_i$ are unbalanced

Macroscopic optimality

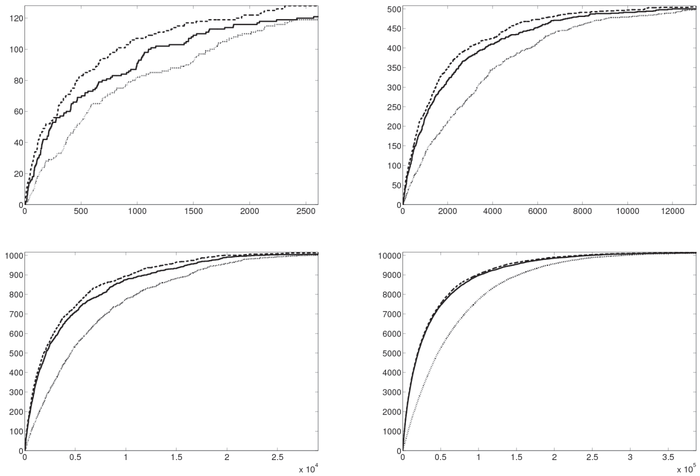
Theorem: Take $C = (1+\sqrt{2})\sqrt{c+2}$ with $c > 3/2$ in the Good-UCB algorithm.

- For every sequence $(\lambda^N)_N$ converging to λ as N goes to infinity, almost surely

$$\limsup_{N\to+\infty} \frac{T^N_{UCB}(\lambda^N)}{N} \leq \sum_i \left(\log \frac{q_i}{\lambda}\right)_+$$

- The proportion of items found after Nt steps F^{GUCB} converges *uniformly* to F^* as N goes to infinity

Simulation



Number of items found by Good-UCB (line), the oracle (bold dashed), and by uniform sampling (light dotted) as a function of time, for sample sizes $N = 128$, $N = 500$, $N = 1000$ et $N = 10000$, in an environment with 7 experts.