Dynamic resource allocation: Bandit problems and extensions

Aurélien Garivier

Institut de Mathématiques de Toulouse

MAD Seminar, Université Toulouse 1
October 3rd, 2014

Roadmap

1. The Bandit Model
2. Lower Bound for the Regret
3. Optimistic Algorithms
4. An Optimistic Algorithm based on Kullback-Leibler Divergence
5. Parametric setting: the kl-UCB Algorithm
6. Non-parametric setting and Empirical Likelihood
7. Extensions

Dynamic resource allocation

Imagine you are a doctor:

- patients visit you *one after another* for a given disease
- you prescribe one of the (say) *5 treatments* available
- the treatments are *not equally efficient*
- you do not know which one is the best, you *observe the effect* of the prescribed treatment on each patient

⇒ What do you do?

- You must choose each prescription using only the *previous observations*
- Your goal is not to estimate each treatment's efficiency precisely, but to *heal as many patients as possible*

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 $1 \le a \le K$ and $s \ge 1$, $X_{a,s} \sim \nu_a$, 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

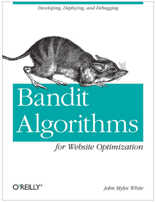
Example Bernoulli rewards: $\theta \in [0,1]^K$, $\nu_a = \mathcal{B}(\theta_a)$

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, ...


Developing, Deploying, and Debugging

Bandit Algorithms

for Website Optimization

O'REILLY

John Myles White

- 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

$$\mathbb{E}\left[S_T\right] = \sum_{t=1}^{T}\sum_{a=1}^{K} \mathbb{E}\left[\mathbb{E}\left[X_t \mathbb{1}\{A_t = a\} | X_1, \ldots, X_{t-1}\right]\right]$$

$$= \sum_{a=1}^{K} \mu_a \mathbb{E}\left[N_a^\pi(T)\right]$$

where $N_a^\pi(T) = \sum_{t \leq T} \mathbb{1}\{A_t = a\}$ 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 $\mu^* \in \max\{\mu_a : 1 \leq a \leq K\}$

Roadmap

1. The Bandit Model
2. Lower Bound for the Regret
3. Optimistic Algorithms
4. An Optimistic Algorithm based on Kullback-Leibler Divergence
5. Parametric setting: the kl-UCB Algorithm
6. Non-parametric setting and Empirical Likelihood
7. Extensions

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 $\nu_a \in \mathcal{F}_a$

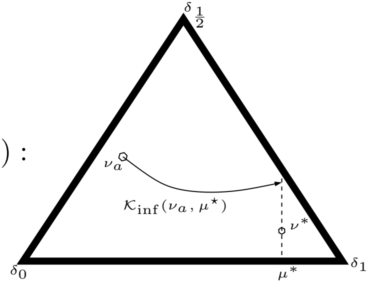
Theorem [Burnetas and Katehakis, '96]

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

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



Intuition

- First assume that μ^* is known and that T is fixed
- How many draws n_a of ν_a are necessary to know that $\mu_a < \mu^*$ with probability at least $1 - 1/T$?
- Test: $H_0 : \mu_a = \mu^*$ against $H_1 : \nu = \nu_a$
- Stein's Lemma: if the first type error $\alpha_{n_a} \leq 1/T$, then

$$\beta_{n_a} \gtrsim \exp\big(-n_a K_{inf}(\nu_a, \mu^*)\big)$$

$\Longrightarrow$ it can be smaller than $1/T$ if

$$n_a \geq \frac{\log(T)}{K_{inf}(\nu_a, \mu^*)}$$

- How to do as well without knowing μ^* and T in advance? Not asymptotically?

Roadmap

Optimism in the Face of Uncertainty

Optimism in an heuristic principle popularized by [Lai&Robins '85; Agrawal '95] which consists in letting the agent

> play as if the environment was the most favorable
> among all environments that are sufficiently likely
> given the observations accumulated so far

Surprisingly, this simple heuristic principle can be instantiated into algorithms that are robust, efficient and easy to implement in many scenarios pertaining to reinforcement learning

Upper Confidence Bound Strategies

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

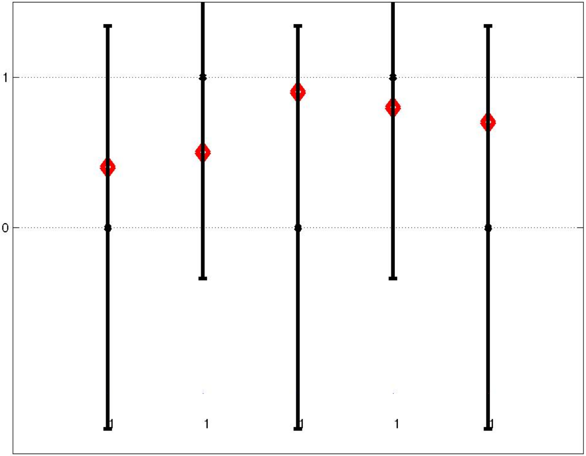
- Construct an upper confidence bound for the expected reward of each arm:

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

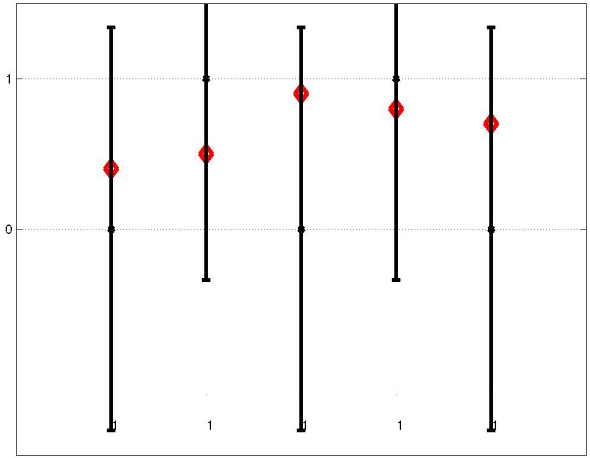
- Choose the arm with the highest UCB

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

UCB in Action



UCB in Action



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])

Roadmap

The KL-UCB algorithm [Cappé,G.&al '13]

```
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 $\{1, \dots, K\}$ once
for $t = K$ to $T - 1$ do
    compute for each arm $a$ the quantity
        $U_a(t) = \sup\left\{ E(\nu) : \quad \nu \in \mathcal{F} \quad \text{and} \quad KL\Big(\Pi_{\mathcal{F}}(\hat{\nu}_a(t)), \nu\Big) \leq \frac{f(t)}{N_a(t)} \right\}$
    pick an arm $\quad A_{t+1} \in \underset{a \in \{1,\dots,K\}}{\arg\max} \; U_a(t)$
end for
```

Roadmap

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 $\mu_a = \dot{b}(\theta_a)$
- The KL-UCB si simply:
$$U_a(t) = \sup\left\{\mu \in \overline{I} : \quad d\big(\hat{\mu}_a(t), \mu\big) \le \frac{f(t)}{N_a(t)}\right\}$$
- 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 $p_{\theta_a}(x) = \theta_a \mathrm{e}^{-\theta_a x}$:
$$d_{\exp}(u, v) = u - v + u \log\frac{u}{v}$$
- The analysis is generic and yields a non-asymptotic regret bound optimal in the sense of Lai and Robbins.

Parametric version: the kl-UCB algorithm

Parameters: $\mathcal{F}$ parameterized by the expectation $\mu \in I \subset \mathbb{R}$ with divergence d, a non-decreasing function $f : \mathbb{N} \to \mathbb{R}$

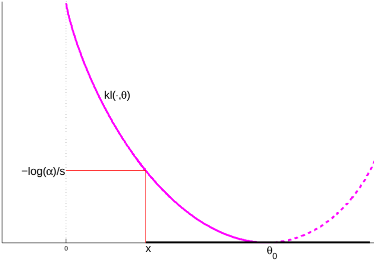
Initialization: Pull each arm of $\{1, \ldots, K\}$ once

```
for $t = K$ to $T - 1$ do
    compute for each arm $a$ the quantity
        $U_a(t) = \sup\left\{\mu \in \overline{I} : \quad d\big(\hat{\mu}_a(t), \mu\big) \leq \frac{f(t)}{N_a(t)}\right\}$
    pick an arm    $A_{t+1} \in \operatorname*{arg\,max}_{a \in \{1,\ldots,K\}} U_a(t)$
end for
```

The kl Upper Confidence Bound in Picture

If $Z_1, \dots, Z_s \overset{iid}{\sim} \tilde{B}(\theta_0)$, $x < \theta_0$ and if $\hat{p}_s = (Z_1 + \dots + Z_s)/s$, then by Chernoff's inequality

$$\mathbb{P}_{\theta_0}\left(\hat{p}_s \leq x\right) \leq \exp\left(-s d_{\text{BER}}(x, \theta_0)\right).$$



In other words, if $\alpha = \exp\left(-s d_{\text{BER}}(x, \theta_0)\right)$:

$$\mathbb{P}_{\theta_0}\left(\hat{p}_s \leq x\right) = \mathbb{P}_{\theta_0}\left(d_{\text{BER}}(\hat{p}_s, \theta_0) \leq -\frac{\log(\alpha)}{s},\ \hat{p}_s < \theta_0\right) \leq \alpha$$

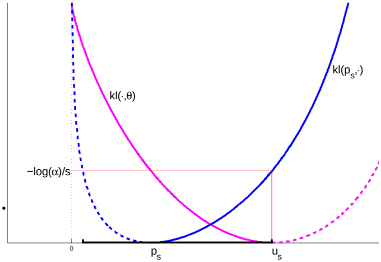
$\Longrightarrow$ Upper Confidence Bound for p at risk α :

$$u_s = \sup\left\{\theta > \hat{p}_s : d_{\text{BER}}(\hat{p}_s, \theta) \leq -\frac{\log(\alpha)}{s}\right\}.$$

The kl Upper Confidence Bound in Picture

If $Z_1, \dots, Z_s \overset{iid}{\sim} \tilde{B}(\theta_0)$, $x < \theta_0$ and if $\hat{p}_s = (Z_1 + \dots + Z_s)/s$, then by Chernoff's inequality

$$\mathbb{P}_{\theta_0}\left(\hat{p}_s \leq x\right) \leq \exp\left(-s d_{\text{BER}}(x, \theta_0)\right).$$



In other words, if $\alpha = \exp\left(-s d_{\text{BER}}(x, \theta_0)\right)$:

$$\mathbb{P}_{\theta_0}\left(\hat{p}_s \leq x\right) = \mathbb{P}_{\theta_0}\left(d_{\text{BER}}(\hat{p}_s, \theta_0) \leq -\frac{\log(\alpha)}{s},\ \hat{p}_s < \theta_0\right) \leq \alpha$$

$\Longrightarrow$ Upper Confidence Bound for p at risk α :

$$u_s = \sup\left\{\theta > \hat{p}_s : d_{\text{BER}}(\hat{p}_s, \theta) \leq -\frac{\log(\alpha)}{s}\right\}.$$

Key Tool: Deviation Inequality for Self-Normalized Sums

- Problem: random number of summands
- Solution: peeling trick (as in the proof of the LIL)

Theorem For all $\epsilon > 1$,

$$\mathbb{P}\Big(\mu_a > \hat{\mu}_a(t) \quad \text{and} \quad N_a(t)\, d\big(\hat{\mu}_a(t),\, \mu_a\big) \geq \epsilon\Big) \leq e\big\lceil \epsilon \log(t) \big\rceil\, e^{-\epsilon}\,.$$

Thus,

$$P\big(U_a(t) < \mu_a\big) \leq e\big\lceil f(t) \log(t) \big\rceil\, e^{-f(t)}$$

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

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

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

Results: Two-Arm Scenario

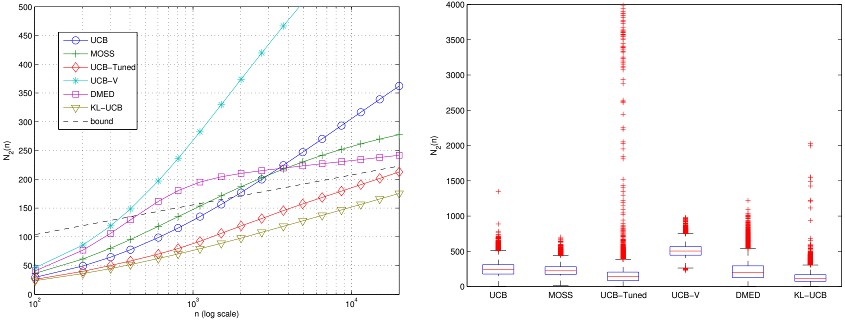


Figure: Performance of various algorithms when $\theta = (0.9, 0.8)$. Left: average number of draws of the sub-optimal arm as a function of time. Right: box-and-whiskers plot for the number of draws of the sub-optimal arm at time $T = 5,000$. Results based on $50,000$ independent replications

Results: Ten-Arm Scenario with Low Rewards

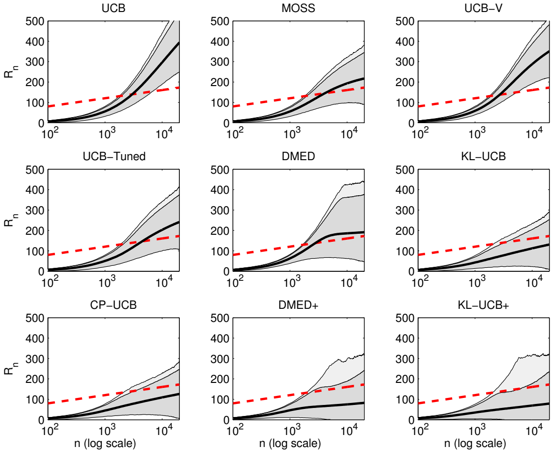


Figure: Average regret as a function of time when $\theta = (0.1, 0.05, 0.05, 0.05, 0.02, 0.02, 0.02, 0.01, 0.01, 0.01)$. Red line: Lai & Robbins lower bound; thick line: average regret; shaded regions: central 99% region an upper 99.95% quantile

Roadmap

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 $0 \leq X \leq 1$ and denote by $\mu = \mathbb{E}[X]$ its mean. Then, for any $\lambda \in \mathbb{R}$,

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

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 $T > 3$,

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

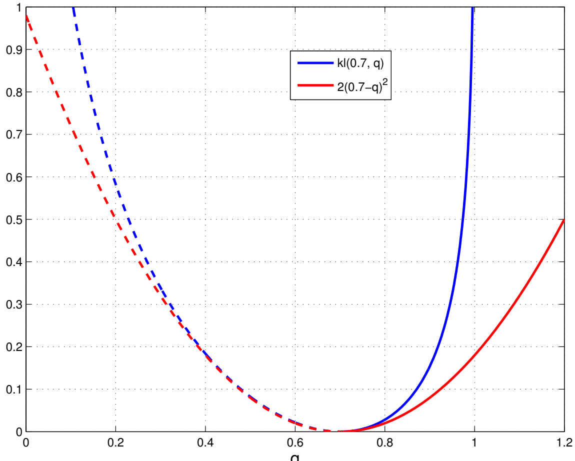
$$+ \left(4e + \frac{3}{d_{\text{BER}}(\mu_a, \mu^\star)}\right)\log\big(\log(T)\big) + \frac{2\left(\log\left(\frac{\mu^\star(1-\mu_a)}{\mu_a(1-\mu^\star)}\right)\right)^2}{\big(d_{\text{BER}}(\mu_a, \mu^\star)\big)^2} + 6\,.$$

- 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

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

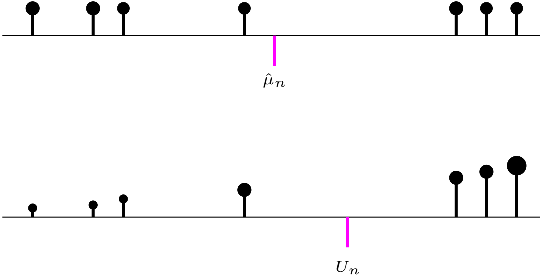


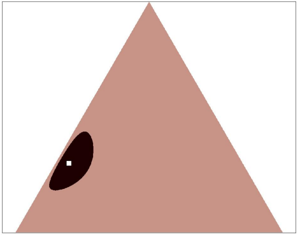
Idea 2: Empirical Likelihood

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

or, rather, *modified Empirical Likelihood*:

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





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.

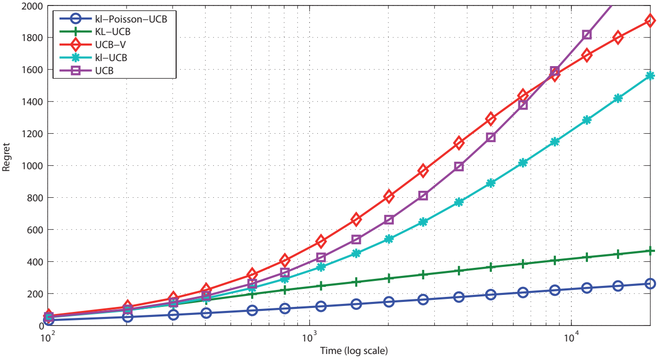
Regret bound

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 $T \geq 3$:

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

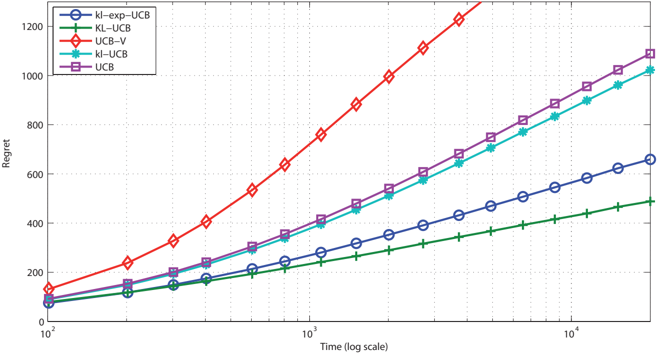
Example: truncated Poisson rewards

- 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.



Example: truncated Exponential rewards

- exponential rewards with respective parameters $1/5$, $1/4$, $1/3$, $1/2$ and 1, truncated at $x_{\max} = 10$;
- kl-UCB uses the divergence $d(x, y) = x/y - 1 - \log(x/y)$ prescribed for genuine exponential distributions, but it ignores the fact that the rewards are truncated.



Take-home message on bandit algorithms

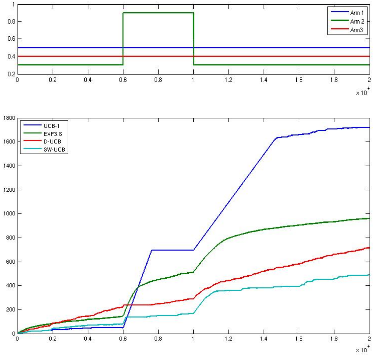
1. Use kl-UCB rather than UCB-1 or UCB-2
2. Use KL-UCB if speed is not a problem
3. todo: improve on the deviation bounds, address general non-parametric families of distributions
4. Alternative: Bayesian-flavored methods:
    - Bayes-UCB [Kaufmann, Cappé, G.]
    - Thompson sampling [Kaufmann & al.]

Roadmap

1. The Bandit Model
2. Lower Bound for the Regret
3. Optimistic Algorithms
4. An Optimistic Algorithm based on Kullback-Leibler Divergence
5. Parametric setting: the kl-UCB Algorithm
6. Non-parametric setting and Empirical Likelihood
7. Extensions

Non-stationary Bandits [G. Moulines '11]

- Changepoint : reward distributions change *abruptly*
- Goal : *follow the best arm*
- Application : scanning tunnelling microscope



- Variants D-UCB et SW-UCB including a progressive *discount* of the past
- Bounds $O(\sqrt{n \log n})$ are proved, which is (almost) optimal

(Generalized) Linear Bandits [Filippi, Cappé, G. & Szepesvári '10]

- Bandit with contextual information:

  $$\mathbb{E}[X_t|A_t] = \mu(m'_{A_t}\theta_*)$$

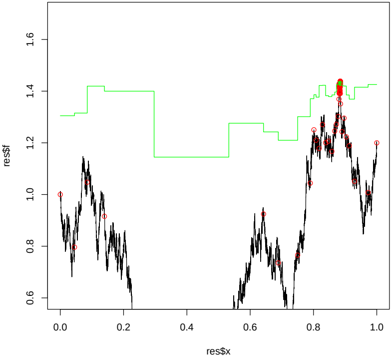
  where $\theta_* \in \mathbb{R}^d$ is an unkown parameter and $\mu : \mathbb{R} \to \mathbb{R}$ is a link function
- Example : binary rewards

  $$\mu(x) = \frac{\exp(x)}{1 + \exp(x)}$$

- Application : targeted web ads
- GLM-UCB : regret bound depending on dimension d and not on the number of arms

Stochastic Optimization

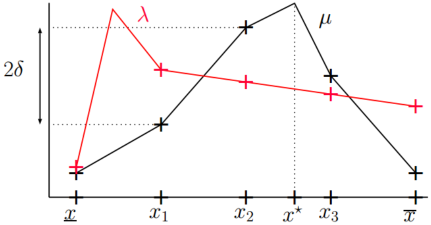
- Goal : Find the maximum of a function $f : C \subset \mathbb{R}^d \to \mathbb{R}$ (possibly) observed in noise
- Application : DAS



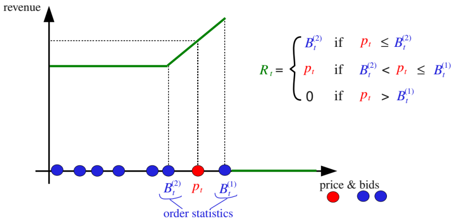
- Model : f is the realization of a Gaussian Process (or has a small norm in some RKHS)
- GP-UCB : evaluate f at the point $x \in C$ where the confidence interval for $f(x)$ has the highest upper-bound

Recent Advances in Continuous bandits

Unimodal bandits without smoothness: trisection algorithms, and better [Combes, Proutière '14]
Application to internet network traffic optimization



Reserve Price Optimization in Second-price Options [Cesa-Bianchi, Gentile, Mansour '13]
Application to advertisement systems



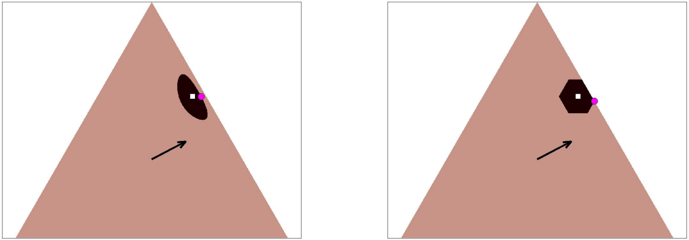
$$R_t = \begin{cases} B_t^{(2)} & \text{if} \quad p_t \leq B_t^{(2)} \\ p_t & \text{if} \quad B_t^{(2)} < p_t \leq B_t^{(1)} \\ 0 & \text{if} \quad p_t > B_t^{(1)} \end{cases}$$

Markov Decision Processes (MDP) [Filippi, Cappé & G. '10]

The system is in state S_t which evolves as a Markov Chain:

$$S_{t+1} \sim P(\cdot; S_t, A_t) \text{ et } R_t = r(S_t, A_t) + \varepsilon_t$$

Optimistic algorithm: search the best transition matrix in a neighborhood of the ML estimate



The use of Kullback-Leibler neighborhoods leads to better performance and has desirable propoerties.

Optimal Exploration with Probabilistic Expert Advice

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

Probabilistic experts : $P_a \in \mathfrak{M}_1(\Omega)$ for $a \in \mathcal{A}$

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

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

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

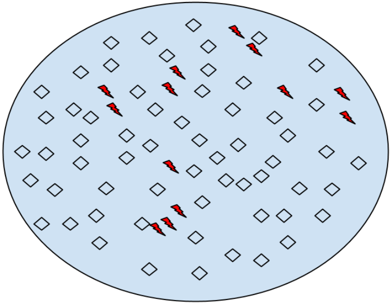
$\neq$ bandit : finding the same element twice is no use !

Oracle : selects the expert with highest 'missing mass'

$$A^*_{t+1} = \arg\max_{a \in \mathcal{A}} P_a\left(B \backslash \{X_1, \dots, X_t\}\right)$$

The model

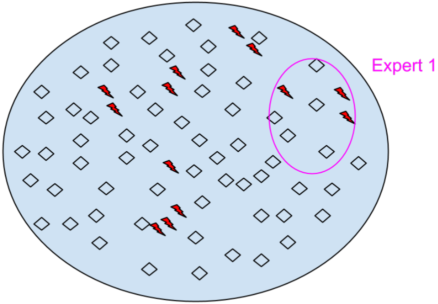
- Subset $A \subset \mathcal{X}$ of important items
- $|\mathcal{X}| \gg 1$, $|A| \ll |\mathcal{X}|$
- Access to $\mathcal{X}$ only by probabilistic experts $(P_i)_{1 \leq i \leq K}$: sequential independent draws



Goal: discover rapidly the elements of A

The model

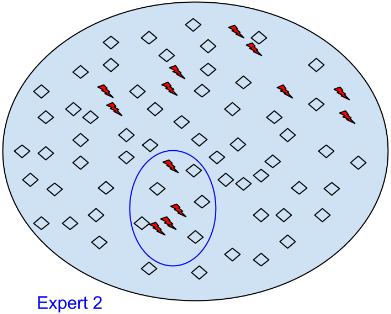
- Subset $A \subset \mathcal{X}$ of important items
- $|\mathcal{X}| \gg 1$, $|A| \ll |\mathcal{X}|$
- Access to $\mathcal{X}$ only by probabilistic experts $(P_i)_{1 \leq i \leq K}$: sequential independent draws



Goal: discover rapidly the elements of A

The model

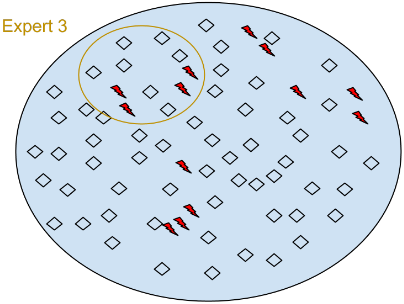
- Subset $A \subset \mathcal{X}$ of important items
- $|\mathcal{X}| \gg 1$, $|A| \ll |\mathcal{X}|$
- Access to $\mathcal{X}$ only by probabilistic experts $(P_i)_{1\leq i\leq K}$: sequential independent draws



Goal: discover rapidly the elements of A

The model

- 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

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

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

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

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

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

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

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"!**

Estimating the missing mass

Notation :

- $X_t \overset{iid}{\sim} P \in \mathfrak{M}_1(\Omega)$, $O_n(\omega) = \sum_{t=1}^{n} \mathbb{1}\{X_t = \omega\}$
- $Z_n(x) = \mathbb{1}\{O_n(\omega) = 0\}$
- $H_n(\omega) = \mathbb{1}\{O_n(\omega) = 1\}, \quad H_n = \sum_{\omega \in B} H_n(\omega)$

Problem : estimate the missing mass

$$R_n = \sum_{\omega \in B} P(\omega) Z_n(\omega)$$

Good-Turing : 'estimator' $\hat{R}_n = H_n/n$ st. $\mathbb{E}[\hat{R}_n - R_n] \in [0, 1/n]$.

Concentration : by McDiarmid's inequality, with probability $\geq 1 - \delta$

$$\left| \hat{R}_n - E[\hat{R}_n] \right| \leq \sqrt{\frac{(2/n + p_{\max})^2 \; n \; \log(2/\delta)}{2}}$$

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

Optimistic algorithm based on Good-Turing's estimator :

$$A_{t+1} = \arg\max_{a \in \mathcal{A}} \left\{ \frac{H_a(t)}{N_a(t)} + c\sqrt{\frac{\log(t)}{N_a(t)}} \right\}$$

- $N_a(t) =$ number of draws of P_a up to time t
- $H_a(t) =$ number of elements of B seen exactly once thanks to P_a
- $c =$ tuning parameter

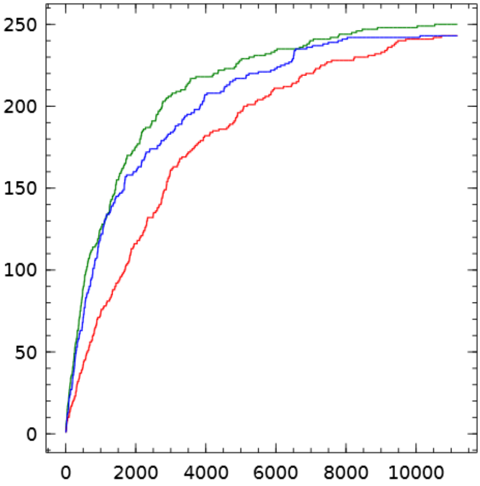
Classical analysis

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

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

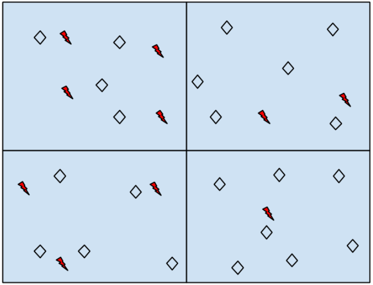
Remark: Usual result for bandit problem, but not-so-simple analysis

Good-UCB en action



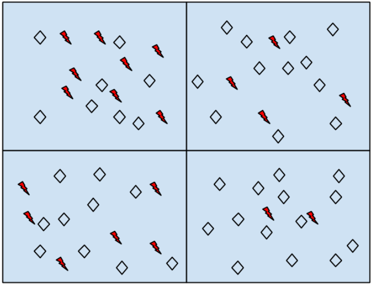
The macroscopic limit

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



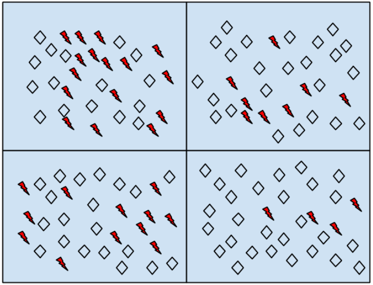
The macroscopic limit

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



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

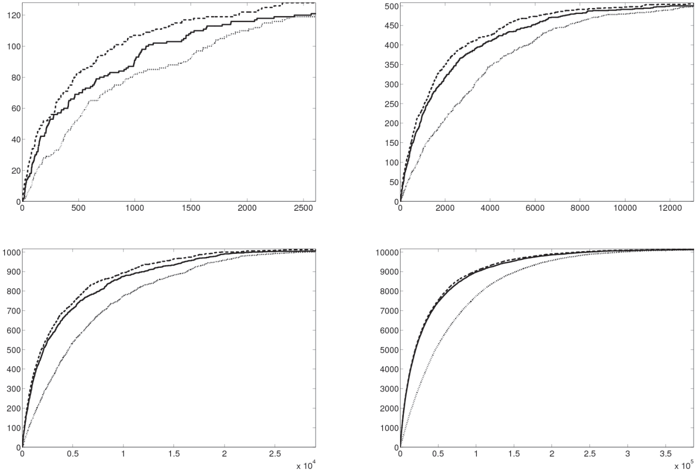
Theorem: Take $C = (1+\sqrt{2})\sqrt{c+2}$ with $c > 3/2$ in the Good-UCB algorithm.

- 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.

Bibliography

1. **[Abbasi-Yadkori&al '11]** Yasin Abbasi-Yadkori, Dávid Pál, Csaba Szepesvári: Online Least Squares Estimation with Self-Normalized Processes: An Application to Bandit Problems CoRR abs/1102.2670: (2011)
2. **[Agrawal '95]** R. Agrawal. Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. Advances in Applied Probability, 27(4) :1054-1078, 1995.
3. **[Audibert&al '09]** J-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. Theoretical Computer Science, 410(19), 2009
4. **[Auer&al '02]** P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47(2) :235-256, 2002.
5. **[Bubeck, Ernst&G. '11]** S. Bubeck, D. Ernst, A. Garivier, Optimal discovery with probabilistic expert advice: finite time analysis and macroscopic optimality Journal of Machine Learning Research vol. 14 Feb. 2013 p.601-623
6. **[De La Pena&al '04]** V.H. De La Pena, M.J. Klass, and T.L. Lai. Self-normalized processes : exponential inequalities, moment bounds and iterated logarithm laws. Annals of Probability, 32(3) :1902-1933, 2004.
7. **[Filippi, Cappé&Garivier '10]** S. Filippi, O. Cappé, and A. Garivier. Optimism in reinforcement learning and Kullback-Leibler divergence. In Allerton Conf. on Communication, Control, and Computing, Monticello, US, 2010.
8. **[Filippi, Cappé, G.& Szepesvari '10]** S. Filippi, O. Cappé, A. Garivier, and C. Szepesvari. Parametric bandits : The generalized linear case. In Neural Information Processing Systems (NIPS), 2010.
9. **[G.&Cappé '11]** A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In 23rd Conf. Learning Theory (COLT), Budapest, Hungary, 2011.
10. **[Cappé,G.&al '13]** O. Cappé, A. Garivier, O-A. Maillard, R. Munos, G. Stoltz, Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation, Annals of Statistics, accepted
11. **[G.&Leonardi '11]** A. Garivier and F. Leonardi. Context tree selection : A unifying view. Stochastic Processes and their Applications, 121(11) :2488-2506, Nov. 2011.
12. **[G.&Moulines '11]** A. Garivier and E. Moulines. On upper-confidence bound policies for non-stationary bandit problems. In Algorithmic Learning Theory (ALT), volume 6925 of Lecture Notes in Computer Science, 2011.
13. **[Lai&Robins '85]** T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 6(1) :4-22, 1985.