SPADRO kick-off meeting: Empirical Likelihood for Optimistic Algorithms in Dynamic Resource Allocation

Sébastien Bubeck, Olivier Cappé, Damien Ernst, Aurélien Garivier, Sarah Filippi, Emilie Kaufmann, Odalric-Ambrym Maillard, Eric Moulines, Rémi Munos, Gilles Stoltz

Institut de Mathématique de Toulouse, Université Paul Sabatier

April 9th, 2014

Clinical Trials

Idealized situation of clinical trials :

- 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

$\Rightarrow$ 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 $1 \leq a \leq K$ 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

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

Performance Evaluation, Regret

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

Our goal Choose π so as to maximize

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

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

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

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 uniformly efficient if for all $\theta \in [0,1]^K$ and all $\alpha > 0$,

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

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

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

The Bound of Burnetas and Katehakis

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

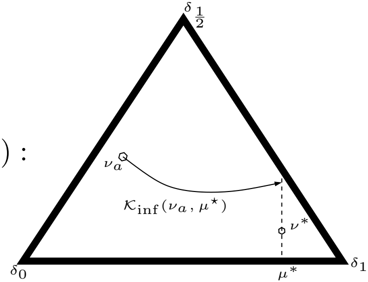
Theorem [Burnetas and Katehakis, '96]

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

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



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]

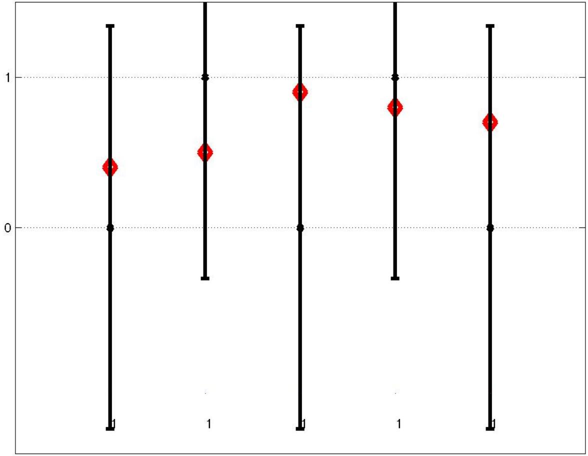
- Construct an upper confidence bound for the expected reward of each arm :

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

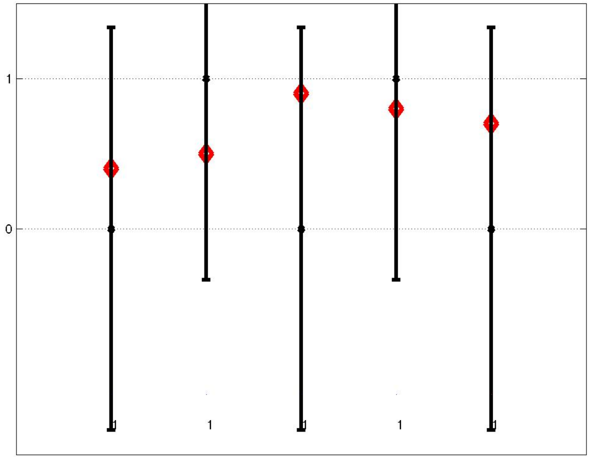
- Choose the arm with the highest UCB

- 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])

The KL-UCB algorithm

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

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

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

end for

Parametric setting : Exponential Families

- 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.

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

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

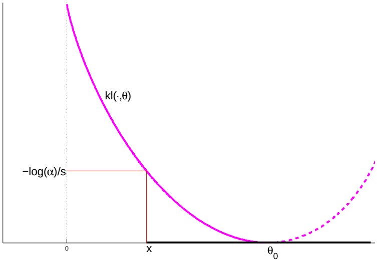
pick an arm $\quad A_{t+1} \in \underset{a \in \{1,\ldots,K\}}{\arg\max}\ U_a(t)$

end for

The kl Upper Confidence Bound in Picture

If $Z_1, \ldots, Z_s \overset{iid}{\sim} \mathcal{B}(\theta_0)$, $x < \theta_0$ and if $\hat{p}_s = (Z_1 + \cdots + Z_s)/s$, then

$$\mathbb{P}_{\theta_0}\left(\hat{p}_s \leq x\right) \leq \exp\left(-s\,\mathrm{kl}(x, \theta_0)\right)$$



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

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

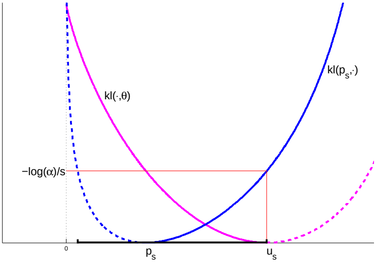
$\Longrightarrow$ upper confidence bound for p at risk α :

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

The kl Upper Confidence Bound in Picture

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

$$\mathbb{P}_{\theta_0}\left(\hat{p}_s \leq x\right) \leq \exp\left(-s\,\mathrm{kl}(x,\theta_0)\right)$$



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

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

$\Longrightarrow$ upper confidence bound for p at risk α :

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

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 $\epsilon > 1$,

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

Thus,

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

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

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

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 $d(\,\cdot\,, \mu^\star)$.

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]

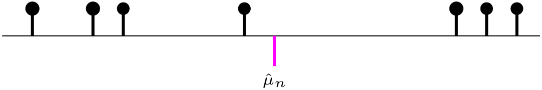
- Bad idea : use Bernstein / Bennett

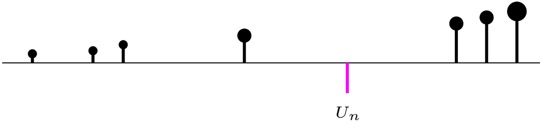
Empirical Likelihood

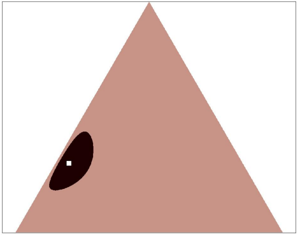
$$U(\hat{\nu}_n, \epsilon) = \sup\Big\{ E(\nu') : \nu' \in \mathfrak{M}_1\big(\mathrm{Supp}(\hat{\nu}_n)\big) \text{ and } \mathrm{KL}(\hat{\nu}_n,\, \nu') \leq \epsilon \Big\}$$

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 $\epsilon > 0$,

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

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.

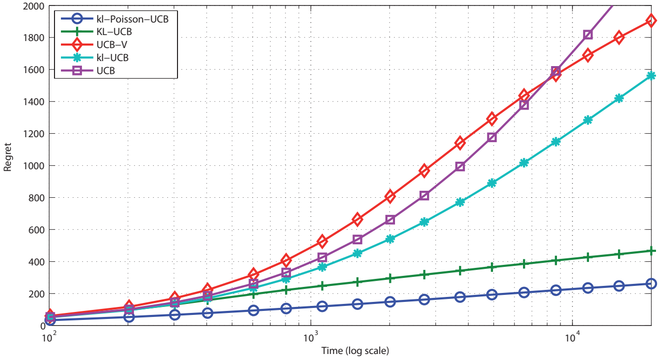
Regret bound

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.