

Introduction to stochastic and adversarial multi-armed bandit problems, with some recent results of the French guys

Gilles Stoltz

CNRS – École normale supérieure – HEC Paris



- 1 Description of the stochastic and the adversarial settings
 - Multi-armed bandits
 - Stochastic setting
 - Adversarial setting
- 2 Randomized strategies for the adversarial setting
 - Upper bounds on the regret
 - Extension: Shortest path problems
 - Minimax order of magnitude of the regret
- 3 Deterministic strategies for the stochastic setting
 - Finite number of arms
 - Extension: Continuum of arms
- 4 Reward of a recommendation instead of a cumulative reward
 - Stochastic setting, modified goal
 - A different exploration–exploitation trade-off



A slot machine with K arms indexed by $\{1, \dots, K\}$ is available.

At each step of a **repeated game**, the forecaster pulls an arm $I_t \in \{1, \dots, K\}$ and gets some **bounded** reward Y_t (say, in $[0, 1]$) associated with it.

He only observes the reward Y_t corresponding to the arm he chose; he does **not observe** the reward he would have got had he chosen a **different** arm.

Thus, at round $t \geq 2$, he can only base his decision on the past observations Y_1, \dots, Y_{t-1} .

His aim is to **maximize** the **sum** of the obtained rewards (in expectation or with high probability).

Stochastic setting:

To each arm $k \in \{1, \dots, K\}$ corresponds a probability distribution ν_k over $[0, 1]$.

The distributions ν_1, \dots, ν_K do not change during the repeated game.

The reward Y_t obtained at round t by the choice of the arm I_t is drawn **independently at random** according to ν_{I_t} .

The sum of the rewards equals $Y_1 + \dots + Y_n$ at round n .

In this case, most results concern the expectation of the sum of the rewards.



Stochastic setting (continued):

K distributions ν_1, \dots, ν_K with respective expectations μ_1, \dots, μ_K are given.

We denote by $\mu^* = \max_{k=1, \dots, K} \mu_k$ the largest expectation.

At round t , the arm I_t is chosen and a reward $Y_t \sim \nu_{I_t}$ is drawn independently.

By the tower rule, $\mathbb{E}[Y_t] = \mathbb{E}[\mu_{I_t}]$.

Maximizing the expectation of the sum of rewards is equivalent to maximizing

$$\mathbb{E}[\mu_{I_1} + \dots + \mu_{I_n}]$$

and even, to **minimizing** the expectation of the **regret**

$$\mathbb{E}[R_n] \quad \text{where} \quad R_n = n \mu^* - (\mu_{I_1} + \dots + \mu_{I_n})$$

Adversarial setting:

The rewards are chosen by an **opponent player**.

For each round $t = 1, 2, \dots$,

- 1 the opponent chooses the rewards $z_{1,t}, \dots, z_{K,t}$ without revealing them;
- 2 the forecaster chooses simultaneously a probability distribution \mathbf{p}_t over $\{1, \dots, K\}$ and draws an arm I_t at random according to \mathbf{p}_t ;
- 3 the forecaster gets the reward $Y_t = z_{I_t,t}$;
- 4 the forecaster only observes Y_t while the opponent may observe \mathbf{p}_t and I_t .

The goal is still to ensure that the sum of rewards $z_{I_1,1} + \dots + z_{I_n,n}$ is large (with **high probability**).



Adversarial setting (continued):

The regret is defined again as the difference between the performance of the best arm and the one of the forecaster

$$R_n = \max_{k=1,\dots,K} \{z_{k,1} + \dots + z_{k,n}\} - (z_{I_1,1} + \dots + z_{I_n,n})$$

The heuristic is that if the **regret** is **small**, the **cumulative reward** of the forecaster should be **large**.

- 1 Description of the stochastic and the adversarial settings
 - Multi-armed bandits
 - Stochastic setting
 - Adversarial setting
- 2 Randomized strategies for the adversarial setting
 - Upper bounds on the regret
 - Extension: Shortest path problems
 - Minimax order of magnitude of the regret
- 3 Deterministic strategies for the stochastic setting
 - Finite number of arms
 - Extension: Continuum of arms
- 4 Reward of a recommendation instead of a cumulative reward
 - Stochastic setting, modified goal
 - A different exploration–exploitation trade-off



The key idea is to first **estimate** the unobserved rewards $z_{1,t}, \dots, z_{K,t}$.

The estimates are given, for all k , by

$$\tilde{z}_{k,t} = \frac{z_{I_t,t}}{p_{I_t,t}} \mathbb{I}_{\{I_t=k\}}$$

We denote by \mathbb{E}_t the **conditional expectation** at round t with respect to the information available to the forecaster and the opponent player at the beginning of round t .

(This fixes the values of \mathbf{p}_t and of the $z_{k,t}$, only the choice of I_t according to \mathbf{p}_t involves randomness.)

The key idea is to first **estimate** the unobserved rewards $z_{1,t}, \dots, z_{K,t}$.

The estimates are given, for all k , by

$$\tilde{z}_{k,t} = \frac{z_{I_t,t}}{p_{I_t,t}} \mathbb{I}_{\{I_t=k\}}$$

We denote by \mathbb{E}_t the **conditional expectation** at round t .

The estimates above are (conditionally) **unbiased**: since I_t is distributed as \mathbf{p}_t ,

$$\mathbb{E}_t \left[\tilde{z}_{k,t} \right] = \frac{z_{k,t}}{p_{k,t}} \mathbb{E}_t \left[\mathbb{I}_{\{I_t=k\}} \right] = \frac{z_{k,t}}{p_{k,t}} p_{k,t} = z_{k,t}$$

See **Auer, Cesa-Bianchi, Freund, and Schapire '02**.



Actually, for technical reasons, we consider in a first time the estimates

$$\hat{z}_{k,t} = 1 - \left(\frac{1 - z_{I_t,t}}{p_{I_t,t}} \right) \mathbb{I}_{\{I_t=k\}}$$

which are **still** conditionally **unbiased**.

Exponentially weighted average predictor (version 1)

With a parameter $\eta > 0$ to be tuned: \mathbf{p}_1 is uniform and for $t \geq 2$,

$$p_{k,t} = \frac{\exp\left(\eta \sum_{s=1}^{t-1} \hat{z}_{k,s}\right)}{\sum_{j=1}^K \exp\left(\eta \sum_{s=1}^{t-1} \hat{z}_{j,s}\right)}$$

Theorem

Let $\eta = \sqrt{(2 \ln K)/n}$.

Then for **all strategies** of the opponent player with rewards picked in $[0, 1]$, the exponentially weighted average predictor using the estimates $\hat{z}_{k,t}$ satisfies

$$\max_{k=1, \dots, K} \mathbb{E} \left[z_{k,1} + \dots + z_{k,n} \right] - \mathbb{E} \left[z_{I_1,1} + \dots + z_{I_1,n} \right] \leq \sqrt{2nK \ln K}$$

The proof is short and would take only two slides.

Adversarial opponent to full information
 The player chooses a strategy that depends on the full information of the opponent's strategy.

Upper bound on the regret
 For all strategies of the opponent player, the regret is bounded by $\sqrt{2nK \ln K}$.

Proof of the lemma

Recall that $\mu_{i,t} = w_{i,t-1}/W_{t-1}$, where $W_{t-1} = w_{1,t-1} + \dots + w_{K,t-1}$, $w_{i,t} = 1$, and for $t \geq 2$,
 $w_{i,t} = \exp(-\eta z_{i,t-1}) = \exp\left(-\eta \sum_{s=1}^{t-1} z_{i,s}\right)$

On the one hand,
 $\ln \frac{W_t}{W_{t-1}} \geq \ln \frac{\sum_{i=1}^K \mu_{i,t} w_{i,t}}{W_{t-1}} = -\eta \sum_{i=1}^K \mu_{i,t} z_{i,t} - \ln W_t$

On the other hand, using $e^x \leq 1 + x + x^2/2$ for $x \geq 0$ and $\ln(1+x) \leq x$, we have, for $t = 1, \dots, n$,
 $\ln \frac{W_t}{W_{t-1}} \leq \ln \frac{\sum_{i=1}^K \mu_{i,t} (1 + \eta z_{i,t} + \frac{\eta^2 z_{i,t}^2}{2})}{W_{t-1}} \leq \ln \sum_{i=1}^K \mu_{i,t} (1 + \eta z_{i,t} + \frac{\eta^2 z_{i,t}^2}{2})$

Summing the upper bounds over $t = 1, \dots, n$ and combining with the lower bound,
 $\sum_{t=1}^n \sum_{i=1}^K \mu_{i,t} z_{i,t} \leq \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{i=1}^K \mu_{i,t} z_{i,t}^2$

Gilles Stoltz

Adversarial opponent to full information
 The player chooses a strategy that depends on the full information of the opponent's strategy.

Upper bound on the regret
 For all strategies of the opponent player, the regret is bounded by $\sqrt{2nK \ln K}$.

Proof

We recall that
 $\tilde{\mu}_{i,t} = \frac{w_{i,t}}{W_t} = \frac{\exp(-\eta z_{i,t})}{\sum_{j=1}^K \exp(-\eta z_{j,t})}$

Using the lemma, we first have
 $\sum_{t=1}^n \sum_{i=1}^K \mu_{i,t} \tilde{\mu}_{i,t} - \min_{j=1, \dots, K} \sum_{t=1}^n \tilde{\mu}_{j,t} \leq \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{i=1}^K \mu_{i,t} z_{i,t}^2$

Taking the expectations of both sides and noting that
 $\sum_{i=1}^K \mu_{i,t} \tilde{\mu}_{i,t} = \mathbb{E}[\mu_{i,t}]$

and
 $\mathbb{E} \left[\sum_{i=1}^K \mu_{i,t} \tilde{\mu}_{i,t} \right] = \mathbb{E} \left[\frac{\mathbb{E}[w_{i,t}]}{W_t} \right] = \mathbb{E}[\mu_{i,t}] = \mathbb{E}[z_{i,t}]$

yields
 $\mathbb{E} \left[\sum_{t=1}^n \sum_{i=1}^K \mu_{i,t} z_{i,t} \right] - \min_{j=1, \dots, K} \mathbb{E} \left[\sum_{t=1}^n z_{j,t} \right] \leq \frac{\ln K}{\eta} + \frac{\eta}{2} \mathbb{E} \left[\sum_{t=1}^n \sum_{i=1}^K \mu_{i,t} z_{i,t}^2 \right]$

Gilles Stoltz



To get **high-probability** bounds, one needs to ensure, e.g., that all arms are pulled sufficiently often by fixing a common lower bound on the probabilities that they are played.

This is an instance of the **exploration – exploitation** trade-off.

Exponentially weighted average predictor (version 2: Exp3)

With parameters $\eta, \gamma > 0$ to be tuned: \mathbf{p}_1 is uniform and for $t \geq 2$,

$$p_{k,t} = (1 - \gamma) \frac{\exp\left(\eta \sum_{s=1}^{t-1} \hat{z}_{k,s}\right)}{\sum_{j=1}^K \exp\left(\eta \sum_{s=1}^{t-1} \hat{z}_{j,s}\right)} + \frac{\gamma}{K}$$

This forecaster ensures that with probability at least $1 - \delta$ (for properly chosen η and γ) and for **all strategies** of the opponent,

$$\max_{k=1,\dots,K} \{z_{k,1} + \dots + z_{k,n}\} - (z_{l_1,1} + \dots + z_{l_n,n}) \leq \square n^{2/3} \ln \frac{1}{\delta}$$



Issue: Average behavior fine but too large random deviations

By taking into account the **mixing** with the **uniform** distribution, one gets an additional γn term in the regret bound.

In addition, various quantities need to be dealt with concentration techniques, e.g., by **Bernstein's** inequality for martingales,

$$\sum_{t=1}^n \widehat{z}_{k,t} \leq \sum_{t=1}^n z_{k,t} + \sqrt{\sum_{t=1}^n \text{Var}_t \widehat{z}_{k,t} \ln \frac{1}{\delta}} + \dots$$

where the conditional variances satisfy

$$\text{Var}_t \widehat{z}_{k,t} \leq \mathbb{E}_t \left[\widehat{z}_{k,t}^2 \right] \leq 2 \left(1 + \left(\frac{1 - z_{k,t}}{p_{k,t}} \right)^2 \mathbb{E}_t \left[\mathbb{I}_{\{I_t=k\}} \right] \right) \leq \frac{\square}{\gamma}$$

since the remaining $1/p_{k,t}$ is of the order of $1/\gamma$.

In total, a term γn has to be **balanced** with a $\sqrt{n/\gamma}$ term, via the choice $\gamma \sim n^{-1/3}$.



We use again the estimates $\tilde{z}_{k,t} = \frac{z_{l_t,t}}{p_{l_t,t}} \mathbb{I}_{\{l_t=k\}}$ but bias them.

Exponentially weighted average predictor (version 3: Exp3.P)

With parameters $\eta, \gamma, \beta > 0$ to be tuned: \mathbf{p}_1 is uniform and for $t \geq 2$,

$$p_{k,t} = (1 - \gamma) \frac{\exp\left(\eta \sum_{s=1}^{t-1} \tilde{z}_{k,s} + \frac{\beta}{p_{k,s}}\right)}{\sum_{j=1}^K \exp\left(\eta \sum_{s=1}^{t-1} \tilde{z}_{j,s} + \frac{\beta}{p_{j,s}}\right)} + \frac{\gamma}{K}$$

The bias terms are used to compensate the main deviation terms when applying Bernstein's inequality.

Shortest path:

A directed graph indicates the possible paths from A to B . The paths share some edges and one only observes the time needed on the edges composing the chosen path.

The regret can be made less than something of the order of $\sqrt{nK \ln K}$ by the previous techniques, but since the number K of paths can be **exponential** in the number E of edges,

- a direct implementation is impossible,
- the bound can be bad since it involves a \sqrt{K} factor.

György, Linder, Lugosi, Ottucsák' 08 deal with both issues and show an efficient forecaster (with **polynomial complexity**), whose regret bound scales essentially with $\sqrt{nE \ln E}$.

Minimax orders of magnitude:

We consider here, for the sake of simplicity, the **expected regret** $\mathbb{E}[R_n]$, where we recall that

$$R_n = \max_{k=1,\dots,K} \{z_{k,1} + \dots + z_{k,n}\} - (z_{I_1,1} + \dots + z_{I_n,n})$$

We consider the following **worst-case** quantities,

$$\sup_{\mathcal{I}} \mathbb{E}[R_n] \leq \sup_{\mathcal{M}_1([0,1]^{Kn})} \mathbb{E}[R_n] = \sup_{z_{k,t}} \mathbb{E}[R_n] \leq \sup_{\tau} \mathbb{E}[R_n]$$

where the suprema are taken

- over all stochastic settings \mathcal{I} ,
- over all joint distributions $\mathcal{M}_1([0,1]^{Kn})$ over $[0,1]^{Kn}$,
- over all individual sequences $z_{k,t}$ of $[0,1]^{Kn}$,
- over all possible strategies τ of the opponent player picking rewards in $[0,1]$.

An easy adaptation of the proofs leading to previous results shows that Exp3.P satisfies

$$\sup_{\tau} \mathbb{E}[R_n] \leq \square \sqrt{nK \ln K}$$

where the supremum is taken over all possible **strategies** τ of the **opponent** player picking rewards in $[0, 1]$.

Auer, Cesa-Bianchi, Freund, and Schapire '02 also proved that

$$\sup_{\mathcal{I}} \mathbb{E}[R_n] \geq \square \sqrt{nK}$$

where the supremum is taken over the set \mathcal{I} of all **stochastic settings** (i.e., product distributions over $[0, 1]^{Kn}$ parameterized by distributions ν_1, \dots, ν_K over $[0, 1]$).

The question is to close the gap and determine if/where the $\sqrt{\ln K}$ is needed.



A recent advance is given by **Audibert and Bubeck '09**.

They exhibit a policy such that

$$\sup_{\mathcal{M}_1([0,1]^{Kn})} \mathbb{E}[R_n] = \sup_{z_{k,t}} \mathbb{E}[R_n] \leq \square \sqrt{nK}$$

This policy is obtained by a careful generalization of Exp3.P resorting to **general reweightings** (other than exponential ones).

This shows that for stochastic settings, individual sequences, and joint distributions, the minimax orders of magnitude in n and K are **\sqrt{nK}** .

They however leave the gap open for the case

$$\sup_{\tau} \mathbb{E}[R_n]$$

where the supremum is taken over all **strategies** of the **opponent**.



- 1 Description of the stochastic and the adversarial settings
 - Multi-armed bandits
 - Stochastic setting
 - Adversarial setting
- 2 Randomized strategies for the adversarial setting
 - Upper bounds on the regret
 - Extension: Shortest path problems
 - Minimax order of magnitude of the regret
- 3 **Deterministic strategies for the stochastic setting**
 - Finite number of arms
 - Extension: Continuum of arms
- 4 Reward of a recommendation instead of a cumulative reward
 - Stochastic setting, modified goal
 - A different exploration–exploitation trade-off

Stochastic setting, reminder:

K distributions ν_1, \dots, ν_K with respective expectations μ_1, \dots, μ_K are given.

We denote by $\mu^* = \max_{k=1, \dots, K} \mu_k$ the largest expectation.

At round t , the arm I_t is chosen and a reward $Y_t \sim \nu_{I_t}$ is drawn independently.

We aim at **minimizing** the expectation of the **regret**

$$\mathbb{E}[R_n] \quad \text{where} \quad R_n = n\mu^* - (\mu_{I_1} + \dots + \mu_{I_n})$$

A deterministic strategy:

We denote by

$$T_k(n) = \sum_{t=1}^n \mathbb{I}_{\{I_t=k\}} \quad \text{and} \quad \hat{\mu}_{k,n} = \frac{1}{T_k(n)} \sum_{t:I_t=k} Y_t$$

the **number of times** a given arm k was played up to round n and the **average reward** it obtained at these rounds.

Upper confidence bound [UCB1] forecaster

At rounds $t = 1, \dots, K$, play $I_t = t$.

At rounds $t \geq K + 1$, play

$$I_t \in \operatorname{argmax}_{k=1, \dots, K} \hat{\mu}_{k,t-1} + \sqrt{\frac{2 \ln(t-1)}{T_k(t-1)}}$$

A deterministic strategy, continued:

Theorem

For all stochastic environments, the regret of UCB1 is bounded, in a *distribution-dependent* sense, as

$$\mathbb{E}[R_n] \leq \left(8 \sum_{i: \mu_i < \mu^*} \frac{1}{\mu^* - \mu_i} \right) \ln n + 5K .$$

Comments and remarks

- Upper **confidence bounds** are constructed using Hoeffding-type inequalities
- We have another instance of the **exploration-exploitation** trade-off (use of the empirical averages versus additional confidence terms that only depend on the number of times a given arm was pulled)

A deterministic strategy, continued:

Lai and Robbins '85 showed that any “good” forecaster is bound to suffer a regret larger than $C \ln n$, where C is a distribution-dependent constant (whose value they discuss more precisely).

A simple adaptation of the proof of the distribution-dependent bound for UCB1 leads to a **distribution-free** bound, which is optimal up to a $\sqrt{\ln n}$ factor in view of the previous results.

Theorem

*The regret of UCB1 is bounded, in a **distribution-free** sense against all stochastic environments supported by $[0, 1]$, as*

$$\sup_{\mathcal{I}} \mathbb{E}[R_n] \leq \square \sqrt{nK \ln n} .$$

Continuum of arms:

In somewhat more complicated stochastic scenarios, there are uncountably many arms x indexed by some topological space \mathcal{X} .

To each arm x corresponds a distribution ν_x with expectation μ_x .

A minimal requirement is that $x \in \mathcal{X} \mapsto \mu_x$ is **continuous**, but depending on the context, stronger regularity assumptions are made.

The goal is still to have a $o(n)$ bound on the regret.

Early references: Agrawal '95, Kleinberg '04

Hierarchical algorithms based on **UCB1**: Kocsis and Szepesvari '06;
Gelly, Wang, Munos, Teytaud' 06; Coquelin and Munos '07;
Bubeck, Munos, Stoltz, Szepesvari '09



- 1 Description of the stochastic and the adversarial settings
 - Multi-armed bandits
 - Stochastic setting
 - Adversarial setting
- 2 Randomized strategies for the adversarial setting
 - Upper bounds on the regret
 - Extension: Shortest path problems
 - Minimax order of magnitude of the regret
- 3 Deterministic strategies for the stochastic setting
 - Finite number of arms
 - Extension: Continuum of arms
- 4 Reward of a recommendation instead of a cumulative reward
 - Stochastic setting, modified goal
 - A different exploration–exploitation trade-off



New goal:

Before, we were interested in getting large **cumulative rewards** $Y_1 + \dots + Y_n$ and had to perform exploration and exploitation at the same time.

Which are the links with the identification of a **good** arm? (On purpose I did not write the **best** arm.)

The quality of an arm j is given, say, by $\Delta_j = \mu^* - \mu_j$, the difference between the expectations of the rewards it yields and the ones generated by the best arm.



New repeated game:

Parameters: K probability distributions for the rewards of the arms, ν_1, \dots, ν_K , with expectations μ_1, \dots, μ_K

For each round $t = 1, 2, \dots$,

- the forecaster chooses $I_t \in \{1, \dots, K\}$ according to some probability distribution \mathbf{p}_t ;
- the environment draws the reward Y_t for this action, independently from ν_{I_t} ;
- the forecaster outputs a recommendation $J_t \in \{1, \dots, K\}$, possibly at random according to some distribution \mathbf{q}_t ;
- if the environment sends a stopping signal, then the game takes an end; otherwise, the next round starts.

New assessment of the quality of a strategy:

The **simple regret** at round n is defined as

$$r_n = \mu^* - \mu_{J_n} = \Delta_{J_n}$$

and we will mostly be interested in $\mathbb{E}[r_n]$.

Note that this criterion is **smoother** than the probability of finding the best arm,

$$\mathbb{P}\left\{J_n \in \operatorname{argmax}_{k=1,\dots,K} \mu_k\right\}$$

and that the links with the expectation of the cumulative regret

$$\mathbb{E}[R_n] = n\mu^* - \mathbb{E}[\mu_{I_1} + \dots + \mu_{I_n}]$$

are not that obvious at first sight.



Simple strategies:

Note first that a strategy in this setting is given by a pair of two sub-strategies,

- an **allocation strategy** that determines the \mathbf{p}_t and thus the pulled arms I_t ;
- a **recommendation strategy** that determines the \mathbf{q}_t and thus the recommended arms J_t .

Here are two simple recommendation strategies,

- the **empirical distribution of plays** chooses

$$\mathbf{q}_n = \frac{1}{n} \sum_{t=1}^n \delta_{I_t} ;$$

- the **most played arm** is given by

$$J_n \in \operatorname{argmax}_{k=1,\dots,K} T_k(n) = \operatorname{argmax}_{k=1,\dots,K} \sum_{t=1}^n \mathbb{I}_{\{I_t=k\}}$$

A small cumulative regret entails a small simple regret:

Lemma

For *all allocations strategies* with cumulative regret denoted by R_n , the simple regret r_n of the recommendation strategy based on them is bounded by,

- for the empirical distributions of plays, $\mathbb{E}[r_n] \leq \frac{\mathbb{E}[R_n]}{n}$
- for the most played arm, $\mathbb{E}[r_n] \leq K \frac{\mathbb{E}[R_n]}{n}$

It is however difficult to say something universal about the recommendation formed by the arm with the **current best average** reward.

However, too small a cumulative regret entails a large simple regret:

Theorem (Bubeck, Munos, Stoltz '09)

For *any pair of allocation and recommendation strategies* and any function $\varepsilon : \{1, 2, \dots\} \rightarrow \mathbb{R}$ such that

for all (Bernoulli) distributions ν_1, \dots, ν_K on the rewards, there exists a constant $C \geq 0$ with $\mathbb{E}[R_n] \leq C \varepsilon(n)$,

the following holds true:

for all sets of $K \geq 3$ distinct Bernoulli distributions on the rewards, with parameters different from 1, there exists a constant $D \geq 0$ and an ordering ν_1, \dots, ν_K of the considered distributions such that

$$\mathbb{E}[r_n] \geq \frac{1}{2} \min\{\Delta_k : \Delta_k > 0\} e^{-D\varepsilon(n)}$$

Comments and consequences:

For instance, when using **UCB1** as an **allocation** strategy, $\epsilon(n) = \ln n$ and the rate of decrease of the simple regrets towards 0 is **at fastest polynomial**, no matter how clever is the recommendation strategy.

On the other hand, it can be shown that with a **uniform allocation** strategy (pull all arms sequentially), the rate is **exponential**, as suggested by the theorem (since R_n is linear in n).

However, simulation studies showed that these results had essentially an asymptotic meaning (when **n is large**).

In Bubeck, Munos, Stoltz '09, we provide bounds when using UCB1 as an allocation strategy in the case of **moderate values of n** . They are better than the ones that can be proved for the uniform allocation.

