Latent Bandits

Odalric-Ambrym Maillard, Shie Mannor

Latent Bandits

Odalric-Ambrym Maillard[1,3], Shie Mannor[2,3]

[1] The Technion, Faculty of Electrical Engineering 32000 Haifa, ISRAEL
[2] odalric-ambrym.maillard@ens-cachan.org
[3] shie@ee.technion.ac.il

Résumé : We consider a multi-armed bandit problem where the reward distributions are indexed by two sets –one for arms, one for type– and can be partitioned into a small number of clusters according to the type. First, we consider the setting where all reward distributions are known and all types have the same underlying cluster, the type's identity is, however, unknown. Second, we study the case where types may come from different classes, which is significantly more challenging. Finally, we tackle the case where the reward distributions are completely unknown. In each setting, we introduce specific algorithms and derive non-trivial regret performance. Numerical experiments show that, in the most challenging agnostic case, the proposed algorithm achieves excellent performance in difficult scenarios.

Mots-clés : Multi-armed Bandits, Latent variables, Regret analysis.

1 Introduction

In a recommender system Li *et al.* (2010, 2011); Adomavicius & Tuzhilin (2005), an agent must display an ad to each incoming client, and a context vector summarizes the observed properties of a client, such as its navigation history or its geographic localization. In a cognitive radio Avner *et al.* (2012); Filippi *et al.* (2008), an agent must select a communication channel, based on its current known location and network conditions, while avoiding collision with other sources (such as radar, WiFI, etc). Both examples can be analyzed within the contextual-multi-armed bandit framework Langford & Zhang (2007); Lu *et al.* (2010), where the contexts summarize the information available to the learner. However, the context alone may not be sufficient to solve these problems optimally : In recommender systems, information such as gender or salary, is typically missing (due to privacy). In cognitive radios, information that a source (or an existing user) is close or far is unknown. In both cases, important information about the reward structure is *not observed*. Such would enable to classify similar situations and possibly output much better predictions.

We study in this paper the underlying problem that we call the *latent multi-armed bandit* problem (we do not consider the contextual part of the problem, that is handled by previous work). More formally, let $\{\nu_{a,b}\}_{a\in\mathcal{A},b\in\mathcal{B}}$ be a set of real-valued probability distributions, that is indexed by two finite sets $\mathcal{A}$ of items (actions) and $\mathcal{B}$ of types. For clarity, and to highlight the role of latent information, we assume that both sets are finite. Extension to continuous parametric settings such as linear contextual-bandit Abbasi-Yadkori *et al.* (2011); Dani *et al.* (2008) is straightforward. We denote $\mu_{a,b} \in \mathbb{R}$ the mean of $\nu_{a,b}$ and assume $\nu_{a,b}$ to be R-sub-Gaussian (with known R), that is

$$\forall \lambda \in \mathbb{R} \quad \log \mathbb{E}_{\nu_{a,b}} \exp\left(\lambda(X-\mu_{a,b})\right) \leqslant R^2\lambda^2/2\,. \tag{1}$$

At each step $n \in \mathbb{N}$, Nature selects some $b_n \in \mathcal{B}$ according to some unknown stochastic process Υ. Then b_n is revealed, and we must select some $a_n \in \mathcal{A}$. Finally, a reward X_n is sampled from ν_{a_n,b_n} and observed. Our goal is to find for all N a sequence of actions $a_{1:N} = \{a_n\}_{1\leqslant n\leqslant N}$ with maximal cumulated reward. The optimal sequence is given by $\{\star_{b_n}\}_{n\in\mathbb{N}}$ where $\star_b \in \operatorname{argmax}_{a\in\mathcal{A}} \mathbb{E}_{X\sim\nu_{a,b}}[X]$. The *expected regret* of an algorithm $\mathfrak{A}$ that produces a sequence of actions $a_{1:N}$ is then simply defined by

$$\mathfrak{R}_N^{\mathfrak{A}} = \sum_{n=1}^{N} \mathbb{E}_{X_n\sim\nu_{\star_{b_n},b_n}}\Big[X_n\Big] - \sum_{n=1}^{N} \mathbb{E}_{X_n\sim\nu_{a_n,b_n}}\Big[X_n\Big]\,.$$

We model the latent information by assuming that $\mathcal{B}$ is partitioned into C clusters $\mathcal{C} = \{\mathcal{B}_c\}_{c=1,\dots,C}$ such that the distributions $\{\nu_{a,b}\}_{a\in\mathcal{A}}$ are the same for each $b \in \mathcal{B}_c$. This common distribution is denoted $\nu_{a,c}$ and

called a *cluster distribution*. We denote the optimal action in $\mathcal{B}_c$ by $\star_c$, and introduce the optimality gaps $\Delta_{a,c} = \mu_{\star_c,c} - \mu_{a,c}$. Both the partition and the number of clusters are unknown.

In the recommender system example, $\mathcal{B}$ would be the set of Ids of users having a same context, partitioned for instance into $C = 4$ groups according to whether the user is a Male/Female and has High/Low income. In the cognitive radio scenario, $\mathcal{B}$ could represent hours of the day, partitioned into $C = 2^3$ parts according to three local radios being active or not [1].

Previous work In Agrawal *et al.* (1989) and more recently in Salomon & Audibert (2011) the case when all cluster distributions are known and all users b come from the same unknown cluster $\mathbf{c}$ is considered. In this already non-trivial setting, Agrawal *et al.* (1989) provided an asymptotic lower bound that significantly differs from the standard lower bound known for the multi-armed bandit problem Lai & Robbins (1985); Burnetas & Katehakis (1996), thus showing that the problem is intrinsically different from a bandit problem. They also analyze a near-optimal (yet costly) algorithm for that problem. In Salomon & Audibert (2011), a simpler algorithm is introduced and analyzed with less tight guarantee. We contribute to that setting in Section 2 with a tighter regret bound for a simple algorithm. We then consider two challenging extensions. In Section 3 users may come from different (instead of one) clusters, and in Section 4 nothing is known about the environment. These new settings could be loosely related to Slivkins (2011) and Hazan & Megiddo (2007), as well as to the recent work Gheshlaghi azar *et al.* (2013).

Contribution In Section 2, we review the important case when the cluster distributions $\{\nu_{a,c}\}_{a\in\mathcal{A},c\in\mathcal{C}}$ are known, and all users come from the same cluster $\mathbf{c}$. We provide intuition about the setting, introduce a new algorithm called **`Single-K-UCB`** that is computationally less demanding than that of Agrawal *et al.* (1989), and prove an explicit finite-time bound (Theorem 4) on its regret, improving on Salomon & Audibert (2011). In Section 3, we analyze the significantly harder and largely unaddressed setting when the cluster distributions are still known, but the users may now come from all clusters. We provide a lower bound (Theorem 5) showing that when the number of clusters is too large with respect to the time horizon, sub-linear regret is not attainable. We introduce an algorithm called **`Multiple-K-UCB`** and prove a non-trivial regret bound (Theorem 6) that makes explicit the effect of the distribution of users Υ on the regret. In Section 4, we target the challenging setting when nothing is known (neither Υ, the cluster distributions, nor even the number of clusters). We provide regret bounds for benchmark **`UCB`**-like algorithms (Theorem 7), and a new algorithm called **`A-UCB`**. Despite the very general setting and poor available information, we are able to prove a weak result (Proposition 1), that enables us to deduce a regret guarantee under mild conditions on the structure of arms (Lemma 1,2). Numerical simulations show in Section 4.2 that the introduced algorithm achieves excellent performance in a number of hard situations. All proofs are provided in the extended version Maillard & Mannor (2013).

Notations. At round n, we denote the number of observations for the pair (a, b) by $N_{a,b}(n) = \sum_{t=1}^{n} \mathbb{I}\{a_t = a, b_t = b\}$ and use $\widehat{\nu}_{a,b}(n)$ and $\widehat{\mu}_{a,b}(n)$ to denote the empirical distribution and empirical mean built from the same observations, respectively. We also introduce $N_b(n) = \sum_{a\in\mathcal{A}} N_{a,b}(n)$. For observations associated to the pair a, b, we denote $U_{a,b}(n)$ a high probability upper bound on the mean $\mu_{a,b}$, and $L_{a,b}(n)$ a high probability lower bound. Unless specified, in the sequel we choose the following $U_{a,b}(n)$ coming from concentration inequality for R-sub-Gaussian variables (see (1)), and define $L_{a,b}(n)$ symmetrically :

$$\begin{aligned} U_{a,b}(n) &= \widehat{\mu}_{a,b}(n) + R\sqrt{\frac{2\log(N_b(n)^3)}{N_{a,b}(n)}} \\ L_{a,b}(n) &= \widehat{\mu}_{a,b}(n) - R\sqrt{\frac{2\log(N_b(n)^3)}{N_{a,b}(n)}}\,. \end{aligned}$$

One could instead use Hoeffding's inequality if the distributions have bounded support, empirical Bernstein's inequality to take the variance into account, self-normalized concentration inequality such as in Garivier & Moulines (2008); Abbasi-Yadkori *et al.* (2011), or even tighter upper bounds based on Kullback-Leibler divergence as explained in Cappé *et al.* (2013). These would lead to slightly improved constants in the regret bounds, at the price of clarity. Thus we focus here on bounds based on the mean only. Let the confidence set be $S_{a,b}(n) = [L_{a,b}(n), U_{a,b}(n)]$ and its size (the gap) be $G_{a,b}(n) = U_{a,b}(n) - L_{a,b}(n)$. To avoid some technical considerations, we assume that $S_{a,b}(n)$ is centered around $\widehat{\mu}_{a,b}(n)$.

1. We assume that radios are active at the same time everyday.

2 Known cluster distributions with single cluster arrivals.

In this section, we consider the case when all the distributions $\{\nu_{a,c}\}_{a\in\mathcal{A},c\in\mathcal{C}}$ are known and arrivals $\{b_n\}_{n\geqslant 1}$ belong to the same *unknown* cluster $\mathbf{c}\in\mathcal{C}$. The difference from a standard multi-armed bandit problem is that the set of possible distributions is finite and known. We can have for instance three arms, two clusters and Bernoulli distributions of respective parameter $0.8, 0.2, 0.9$ for one cluster, and Bernoulli distributions of parameter $0.8, 0.1, 0.5$ for the second one. This modifies the achievable guarantees :

Theorem 1 (Agrawal *et al.* (1989))
Let $\mathbf{c}\in\mathcal{C}$ be the true class (that is supp$(\Upsilon)\subset\mathcal{B}_{\mathbf{c}}$), and $\mathcal{A}_-=\mathcal{A}\setminus\{\star_{\mathbf{c}}\}$ be the set of sub-optimal arms. Then, a lower performance bound is

$$\liminf_{N\to\infty}\frac{\mathfrak{R}_N}{\log(N)}\geqslant\min_{\omega_{\mathbf{c}}\in\mathcal{P}(\mathcal{A}_-)}\max_{c'\in C(\mathbf{c})}\frac{\sum_{a\in\mathcal{A}_-}\omega_{\mathbf{c},a}\Delta_{a,\mathbf{c}}}{\sum_{a\in\mathcal{A}_-}\omega_{\mathbf{c},a}KL(\nu_{a,\mathbf{c}}||\nu_{a,c'})},$$

where $C(\mathbf{c})=\left\{c'\in\mathcal{C}:\nu_{\star_{\mathbf{c}},c'}=\nu_{\star_{\mathbf{c}},\mathbf{c}}\text{ and }\star_{\mathbf{c}}\neq\star'_c\right\}$.

Theorem 2 (Agrawal *et al.* (1989))
For each $c\in\mathcal{C}$, let $\omega_c^\star$ that achieves the minimum in the lower bound of Theorem 1. The algorithm proposed by Agrawal et al. *(1989) makes use of $\{\omega_c^\star\}_{c\in\mathcal{C}}$ and achieves*

$$\mathfrak{R}_N\leqslant\left(\max_{c'\in\mathcal{C}(\mathbf{c})}\frac{\sum_{a\in\mathcal{A}_-}\omega^\star_{\mathbf{c},a}\Delta_{a,\mathbf{c}}}{\sum_{a\in\mathcal{A}_-}\omega^\star_{a,\mathbf{c}}KL(\nu_{a,\mathbf{c}}||\nu_{a,c'})}+o(1)\right)\log(N).$$

Although theoretically appealing, it may be in general expensive to compute the quantities $\{\omega_c^\star\}_{c\in\mathcal{C}}$, which makes the algorithm less practical. On the other hand, Salomon & Audibert (2011) introduced the **GCL** algorithm, seemingly without being aware of the work of Agrawal *et al.* (1989) and got the following non-asymptotic result :

Theorem 3 (Salomon & Audibert (2011))
Assume that for all $c,c'\in\mathcal{C}$, for all $a\in\mathcal{A}$, then either $\nu_{a,c}\neq\nu_{a,c'}$ or (either $\star_c\neq a$ or $\star'_c=a$), or $\exists a'\neq a:\mathbb{P}_{\nu_{a',c}}(\frac{d\nu_{a',c}}{d\nu_{a',c'}}(X)>0)=0$. Then if $\mathbf{c}\in\mathcal{C}$ with unique best arm is the true environment, then for all $\beta>0$ it holds for some constants C,C' that

$$\forall n\forall a\neq\star_{\mathbf{c}}\,\mathbb{P}\left(\sum_{b\in\mathcal{B}_{\mathbf{c}}}N_{a,b}(n)\geqslant C\frac{\log(n)}{\Delta^2_{a,\mathbf{c}}}\right)\leqslant C'n^{-\beta}.$$

GCL is fairly easy to implement, however the way this bound is stated makes it hard to understand, all the more so that the constants are not explicit. Also the dependency with $\Delta^2_{a,\mathbf{c}}$ seems sub-optimal.

For completeness, we now introduce an efficient algorithm directly inspired from Agrawal's work. The price for the reduced complexity is that we lose the asymptotic optimality. We start with some intuition about our setting.

High level intuition For clarity, we focus on means only (instead of distributions). Let $\mathcal{C}_{n-1}=\left\{c\in\mathcal{C},\forall a\in\mathcal{A}:\mu_{a,c}\in S_{a,\mathcal{B}}(n-1)\right\}$ be the set of admissible classes at round $n-1$, where the confidence set $S_{a,\mathcal{B}}(n-1)$ is built using observations for the pairs $\{(a,b)\}_{b\in\mathcal{B}}$. Note that by concentration of measure, with high probability the true class $\mathbf{c}$ is admissible and thus $\mathcal{C}_{n-1}$ is not empty. Let then $\tilde{c}\in\mathcal{C}_{n-1}$ be an admissible class. It makes sense to pull its optimal arm $\star_{\tilde{c}}=\text{argmax}_{a\in\mathcal{A}}\,\mu_{a,\tilde{c}}$ (that is known). Now several situations may occur :

a) For another class $c'\in\mathcal{C}$, if $|\mu_{\star_{\tilde{c}},c'}-\mu_{\star_{\tilde{c}},\tilde{c}}|>G_{a,\mathcal{B}}(n-1)$, then c' cannot be admissible. Now if when c' is admissible then $\star_{\tilde{c}}=\star_{c'}$, it means that choosing to play $\star_{\tilde{c}}$ for $\tilde{c}\in\mathcal{C}_{n-1}$ is safe (that is $\star_{\tilde{c}}=\star_{\mathbf{c}}$ happens with high probability).

b) If $\exists c' \in \mathcal{C}$ such that both $|\mu_{\star_{\tilde{c}},c'} - \mu_{\star_{\tilde{c}},\tilde{c}}| \leqslant G_{a,\mathcal{B}}(n-1)$ and $\star_{\tilde{c}} \neq \star_{c'}$, there are many admissible classes that lead to different actions to play. The situation is tricky since playing arm $\star_{\tilde{c}}$ does not separate $\tilde{c}$ from c' (it may be that $\nu_{\star_{\tilde{c}},\tilde{c}} = \nu_{\star_{\tilde{c}},c'}$), and may moreover be sub-optimal since we may have $\star_{\tilde{c}} \neq \star_{\mathbf{c}}$.

Algorithm Agrawal *et al.* (1989) uses a fancy procedure to handle case b). Here, we note that if we choose the class $\tilde{c}$ (and thus action $\star_{\tilde{c}}$) with maximal best mean, this ensures that $\mu_{\star_{\mathbf{c}},\mathbf{c}} - \mu_{\star_{\tilde{c}},\mathbf{c}} \leqslant \mu_{\star_{\tilde{c}},\tilde{c}} - \mu_{\star_{\tilde{c}},\mathbf{c}}$ and thus a controlled error. This observation leads to the **`Single-K-UCB`** algorithm, whose pseudo-code is provided in Algorithm 1. Straightforwardly, if $\mathcal{C}_{n-1}$ is empty, it reduces to playing round-robin, in case a), $\mathcal{A}^\star_{n-1}$ is a singleton, and in case b), we have a controlled error.

Algorithm 1 The **`Single-K-UCB`** algorithm.

Require: The cluster distributions $\{\nu_{a,c}\}_{a\in\mathcal{A},c\in\mathcal{C}}$.

1: **for** $n = 1...N$ **do**

2: Receive $b_n \sim \Upsilon$.

3: Define the set of admissible classes $\mathcal{C}_{n-1} = \Big\{c \in \mathcal{C} : \forall a \in \mathcal{A}\, \mu_{a,c} \in S_{a,\mathcal{B}}(n-1)\Big\}$.

4: Define the set of "elite" admissible arms $\mathcal{A}^\star_{n-1} = \{a \in \mathcal{A}; \exists c \in \mathcal{C}_{n-1}\ \star_c = a\}$.

5: Choose the next arm (breaks ties with round-robin)

$$a_n = \underset{a=\star_c, c\in C_{n-1}}{\operatorname{argmax}} \mu_{\star_c,c}\,. \tag{2}$$

6: **end for**

Regret bound Such an algorithm enjoys the following regret performance :

Theorem 4 (Regret bound for single cluster arrivals)
The regret of **`Single-K-UCB`** *satisfies*

$$\mathfrak{R}_N^{\texttt{Single-K-UCB}} \leqslant \sum_{a\in\mathcal{A}^\star} \frac{24R^2\Delta_{a,\mathbf{c}}\log(N)}{\Delta^{+\,2}_{a,\mathbf{c}}} + \Delta_{a,\mathbf{c}}\Big(1+\frac{\pi^2}{3}\Big)\,,$$

where $\mathcal{A}^\star = \Big\{a \in \mathcal{A} : \exists c \in \mathcal{C}\ \text{s.t.}\ \star_c = a\Big\}$ *and*

$$\Delta^+_{a,\mathbf{c}} = \inf_{c'\in\mathcal{C}} \Big\{\mu_{a,c'} - \mu_{a,\mathbf{c}} : \star_{c'} = a \cap \mu_{\star_{c'},c'} \geqslant \mu_{\star_{\mathbf{c}},\mathbf{c}}\Big\}\,.$$

The notation $\Delta^+_{a,\mathbf{c}}$ comes from the fact that $\Delta^+_{a,\mathbf{c}} \geqslant \Delta_{a,\mathbf{c}}$. Note the link between this bound and that of Theorem 2 (also $\Delta^+_{a,\mathbf{c}}$ and $\mathcal{C}(\mathbf{c})$). Of course the bound of Theorem 2 can be better and this seems to be the price for the simplicity of **`Single-K-UCB`**. On the other hand, since Theorem 4 scales with $\Delta^+_{a,\mathbf{c}}$ (which can be arbitrarily larger than $\Delta_{a,\mathbf{c}}$; see Figure 1), it improves on the result of Theorem 3, and moreover provides explicit constants. Finally, it is straightforward to improve the constants using tighter confidence bounds as discussed in the introduction.

3 Known cluster distributions with multiple cluster arrivals.

We now turn to the more challenging case when the distributions $\{\nu_{a,c}\}_{a\in\mathcal{A},c\in\mathcal{C}}$ are still known to the learner, but when the users may come from different clusters, and the learner does not know what class c corresponds to some input $b \in \mathcal{B}$. In this setting, the lower bound from Theorem 1 can be strengthen. Indeed, without further assumptions, it may be the case that if the number of clusters C is too large with respect to the time horizon N, we don't have time to learn and we can not ensure to have sub-linear regret :

Theorem 5 (Regret lower-bound for multiple cluster arrivals)
Let Υ *be the uniform distribution over* $\mathcal{B}$ *and consider that the distributions are partitioned exactly into* $C > A$ *groups of equal size. Then, it holds*

$$\inf_{\text{algo}} \sup_{\nu_{a,c}} \mathfrak{R}_N \geqslant \frac{1}{20}\min\{\sqrt{NAC}, N\}\,.$$

This shows that for the scaling $C = \Omega(N)$ the problem becomes hopeless, since for any bandit algorithm there exists a set of distributions $\{\nu_{a,b}\}_{a\in\mathcal{A},b\in\mathcal{B}}$ such that the regret is linear in N.

Despite this difficulty, it is possible to slightly modify **`Single-K-UCB`** for that setting, which leads to algorithm 2 that enjoys the following regret performance.

Algorithm 2 The **`Multiple-K-UCB`** algorithm.

Require: The cluster distributions $\{\nu_{a,c}\}_{a\in\mathcal{A},c\in\mathcal{C}}$.

1: **for** $n = 1...N$ **do**

2: Receive $b = b_n \sim \Upsilon$.

3: Define the set of admissible classes $\mathcal{C}_{n-1}(b) = \Big\{c \in \mathcal{C}, \forall a \in \mathcal{A} : \mu_{a,c} \in S_{a,b}(n-1)\Big\}$.

4: Define the set of "elite" admissible arms $\mathcal{A}^\star_{n-1} = \{a \in \mathcal{A}; \exists c \in \mathcal{C}_{n-1}(b_n)\ \star_c = a\}$.

5: Choose the most optimistic "elite" arm

$$a_n = \underset{a=\star_c,\, c\in\mathcal{C}_{n-1}(b_n)}{\operatorname{argmax}} \mu_{\star_c,c}\,.$$

6: **end for**

Theorem 6 (Regret for multiple cluster arrivals)
The regret of **`Multiple-K-UCB`** *satisfies*

$$\mathfrak{R}_N^{\texttt{Multiple-K-UCB}} \leqslant \sum_{b\in\mathcal{B}} \sum_{a\in\mathcal{A}^\star} \min\left\{ \frac{24R^2\Delta_{a,c_b}\log(N\Upsilon(b))}{\Delta^{+\,2}_{a,c_b}} + O\big(\Upsilon(b)^{-1}\big), \Delta_{a,c_b} N\Upsilon(b) \right\},$$

where $c_b \in \mathcal{C}$ denotes the class corresponding to $b \in \mathcal{B}$.

In order to see the benefit of knowing the distributions $\{\nu_{a,c}\}_{a\in\mathcal{A},c\in\mathcal{C}}$, a natural benchmark algorithm is the one that simply plays independent copies of **`UCB`** on each $b \in \mathcal{B}$ (see Auer (2003)), without using the knowledge of the cluster distributions. We call this algorithm **`UCB`** on $\mathcal{B}$; see Algorithm 3. Importantly, due to the inequality $\Delta^+_{a,c_b} \geqslant \Delta_{a,c_b}$ and because only elite arms $a \in \mathcal{A}^\star$ are pulled, the regret of **`Multiple-K-UCB`** is never worse than that of **`UCB`** on $\mathcal{B}$ (Theorem 7) ; it can potentially be much smaller.

Algorithm 3 The **`UCB`** on $\mathcal{B}$ algorithm

1: **for** $n = 1...N$ **do**

2: Receive $b_t \sim \Upsilon$.

3: Compute the empirical means $\widehat{\mu}_{a,b}(n-1)$.

4: Choose the next arm (breaks ties arbitrary)

$$a_n = \underset{a\in\mathcal{A}}{\operatorname{argmax}}\, U_{a,b_n}(n-1)\,. \tag{3}$$

5: **end for**

Illustration In order to highlight the role played by $\Delta^+_{a,c}$, Figure 1 depicts the upper-bounds from Theorem 6 and and from Theorem 7, for one randomly generated problem (we do not compare the regret, but the bounds, to emphasize the theoretical gap). For clarity, we reported the values of $\Delta^+_{a,c}$ as well as of the optimality gaps $\Delta_{a,c}$ for each arm and each class. Here three arms that may be pulled by **`UCB`** on $\mathcal{B}$ are never pulled by **`Multiple-K-UCB`**. Note that the improvement can sometimes be huge : for instance when all $\star_c$ are equal, then $\Delta^+_{a,c} = \infty$ for all sub-optimal arm and the bound from Theorem 6 equals zero.

4 The agnostic case.

In Sections 2 and 3, using the knowledge of the cluster distributions, we derived regret bounds that may significantly improve on their equivalent agnostic version. We now detail an improvement that is even more effective and applicable both in case cluster distributions are known or not.

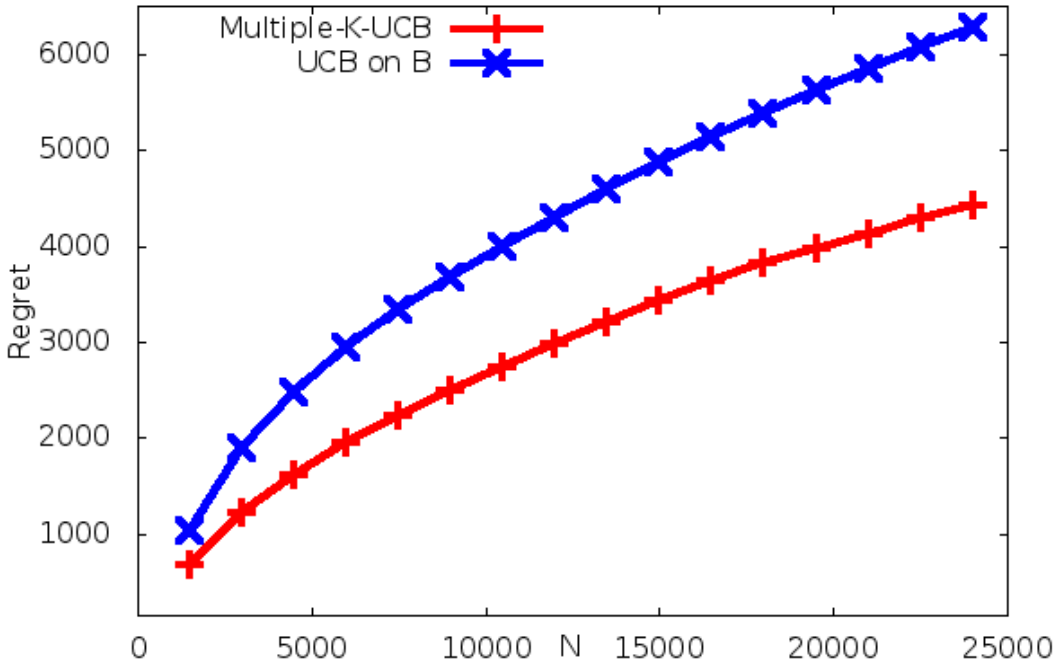

FIGURE 1 – Theoretical regret bounds for **`Multiple-K-UCB`** (Theorem 6) and **`UCB`** on $\mathcal{B}$ (Theorem 7) for one problem characterized by $|\mathcal{A}| = 3, |\mathcal{B}| = 50, |\mathcal{C}| = 4$ and

	1	2	3	4
$\mu_{a,c}$: 1	0.527	0.209	**0.713**	**0.762**
2	**0.717**	0.193	0.575	0.230
3	0.669	**0.751**	0.120	0.485
$\Delta^+_{a,c}$: 1	0.235	0.553	0.0	0.0
2	0.0	$+\infty$	0.142	$+\infty$
3	0.082	0.0	0.631	$+\infty$
$\Delta_{a,c}$: 1	0.190	0.542	0.0	0.0
2	0.0	0.558	0.138	0.533
3	0.0475	0.0	0.593	0.277

We first note that using estimates from each distributions $\nu_{a,b}$ *separately* in order to decide the best action for the cluster $c(b) = c$ seems sub-optimal since the number of samples $N_{a,b}(n)$ available for the couple (a, b) is typically small, while we could possibly gain much more by using all observations in each $\mathcal{B}_c$ (This is basically what happens in Section 2). Indeed, if two distributions $\nu_{a,b}$ and $\nu_{a,b'}$ are the same, then grouping the corresponding observations provides a faster convergence speed. In general, grouping subsets of $\{\nu_{a,b}\}_{b\in\mathcal{B}}$ may lead to a dramatic speed-up if we group similar distributions, and may create a bias if they significantly differ. Thus, there is a trade-off between getting *fast* versus *accurate* convergence, and it is a priori not clear whether we can get a provable improvement.

Benchmark We now introduce an oracle that knows the identity of the clusters perfectly. The simplest one is an algorithm that runs a version of **UCB** separately on each group $\mathcal{B}_c$ (and not each b). We call this benchmark **UCB** on $\mathcal{C}$. Note that although it knows the clusters this is not the best oracle : In some cases, it may be better to further group some clusters together. This algorithm is easy to analyze. To understand the kind of improvement we are targeting, the following theorem compares the regret of **UCB** on $\mathcal{B}$, to that of the oracle **UCB** on $\mathcal{C}$.

Theorem 7 (Baseline and Oracle regret for multiple cluster arrivals)
The expected regret at time N of the algorithm **UCB** *on $\mathcal{B}$ is upper bounded by*

$$\mathfrak{R}_N^{\text{UCB on } \mathcal{B}} \leqslant \sum_{b\in\mathcal{B}}\sum_{a\in\mathcal{A}} \min\Big\{\frac{24R^2\log(N\Upsilon(b))}{\Delta_{a,b}} + O\Big(\Upsilon(b)^{-1}\Big), \Delta_{a,b}N\Upsilon(b)\Big\},$$

where $\Delta_{a,b} = \mu_{\pi^\star(b),b} - \mu_{a,b}$ is the optimality gap of arm a for environment b. Similarly, the expected regret at time N of **UCB** *on $\mathcal{C}$ is upper bounded by*

$$\mathfrak{R}_N^{\text{UCB on } \mathcal{C}} \leqslant \sum_{c=1}^{C}\sum_{a\in\mathcal{A}} \min\Big\{\frac{24R^2\log(N\Upsilon(\mathcal{B}_c))}{\Delta_{a,c}} + O\Big(\Upsilon(\mathcal{B}_c)^{-1}\Big), \Delta_{a,c}N\Upsilon(\mathcal{B}_c)\Big\},$$

where $\Delta_{a,c}$ is the common value of the $\Delta_{a,b}$ for $b \in \mathcal{B}_c$.

As a result, the regret of **UCB** on $\mathcal{C}$ can be significantly smaller than the one of **UCB** on $\mathcal{B}$. Indeed, only looking at the term in factor of $\log(N)$, we get an improvement going from $\sum_{b\in\mathcal{B}}\sum_{a\in\mathcal{A}}\Delta_{a,b}^{-1}$ to $\sum_{c=1}^{C}\sum_{a\in\mathcal{A}}\Delta_{a,c}^{-1}$. This can be substantial, since typically C is much smaller than B. Note of course that the partition $\mathcal{C}$ is unknown in practice. Also, we emphasize that the lower bound of Theorem 5 also holds for that setting.

Grouping distributions We now detail the improvement we are going to consider. Let $B \subset \mathcal{B}$. We define, similarly to $\widehat{\mu}_{a,b}(n)$, $L_{a,b}(n)$ and $U_{a,b}(n)$ the empirical group estimate $\widehat{\nu}_{a,B}(n)$ with associated group mean $\mu_{a,B}(n)$, confidence intervals $U_{a,B}(n)$, $L_{a,B}(n)$ and confidence set $S_{a,B}(n)$, where

$$\begin{aligned}\widehat{\nu}_{a,B}(n) &= \frac{\sum_{b'\in\mathcal{B}}\widehat{\nu}_{a,b'}(n)N_{a,b'}(n)\mathbb{I}\{b'\in B\}}{\sum_{b'\in\mathcal{B}}N_{a,b'}(n)\mathbb{I}\{b'\in B\}},\\ \mu_{a,B}(n) &= \frac{\sum_{b'\in\mathcal{B}}\mu_{a,b'}N_{a,b'}(n)\mathbb{I}\{b'\in B\}}{\sum_{b'\in\mathcal{B}}N_{a,b'}(n)\mathbb{I}\{b'\in B\}}.\end{aligned}$$

Note that in the special case when $B = \mathcal{B}_c$, then $\mu_{a,\mathcal{B}_c}(n) = \mu_{a,c}$. This may not hold in general for other sets B in case the $\{\mu_{a,b'}\}_{b'\in B}$ are distinct from $\mu_{a,c}$. Thus, grouping the observations generally creates a bias. However, the speed of convergence of the group depends on $N_{a,B}(n) = \sum_{b'\in\mathcal{B}}N_{a,b'}(n)\mathbb{I}\{b'\in B\}$, which is typically much faster than that of a single point b (that depends on $N_{a,b}(n)$). Thus the confidence interval $S_{a,B}(n) = [L_{a,B}(n), U_{a,B}(n)]$ is potentially much smaller than $S_{a,b}(n)$. Finally, note that, by construction, we have $\mu_{a,B}(n) \in S_{a,B}(n)$ with high probability, but that for some $b \in B$ there is no reason that $\mu_{a,b} \in S_{a,B}(n)$ due to the introduced bias.

In order to leverage the structural bias, we restrict possible groups B, using two observations. First, if $\mu_{a,b} = \mu_{a,b'}$, then we must have $S_{a,b}(n) \cap S_{a,b'}(n) \neq \emptyset$ with high probability. More generally, a set B such that $\mu_{a,b} = \mu_{a,b'}$ for all $b, b' \in B$ must satisfy that for all $B' \subset B$ and all $B'' \subset B$, with high probability, $S_{a,B''}(n) \cap S_{a,B'}(n) \neq \emptyset$. Second, we define, for an adaptive $\varepsilon = \varepsilon_{a,b,b',n}$, the enlarged confidence bounds

$$\begin{aligned}U_{a,b}(n; 1+\varepsilon) &= \widehat{\mu}_{a,b}(n) + (1+\varepsilon)\big(U_{a,b}(n) - \widehat{\mu}_{a,b}(n)\big),\\ L_{a,b}(n; 1+\varepsilon) &= \widehat{\mu}_{a,b}(n) - (1+\varepsilon)\big(\widehat{\mu}_{a,b}(n) - L_{a,b}(n)\big),\end{aligned}$$

and then $S_{a,b}(n;1+\varepsilon)=[L_{a,b}(n;1+\varepsilon),U_{a,b}(n;1+\varepsilon)]$. This enables us to get the property that if $\mu_{a,b}=\mu_{a,b'}$ and [2] $G_{a,b'}(n)\leqslant\frac{\varepsilon}{2}G_{a,b}(n)$, we must have $S_{a,b'}(n)\subset S_{a,b}(n;1+\varepsilon)$ with high probability.

Note that we here focus only on mean-based procedures for clarity, but it is of course possible to use empirical distributions $\widehat{\nu}_{a,b}(n)$ to remove points sb' that have an obvious mismatch in Kullback-Leibler divergence. We do not discuss such improvements to avoid distracting the reader from the main message.

All in all, we define two sets of sets : First $\mathfrak{B}_b(n)$ for *compatible* sets, and then $\mathfrak{B}_b^+(n)$ for *maximally compatible* (or "elite") sets, that have maximal group speed of convergence and a controlled bias :

$$\mathfrak{B}_b(n)\stackrel{\text{def}}{=}\Big\{B\subset\mathcal{B}:\forall a\in\mathcal{A}\forall b',b''\in B\; S_{a,b'}(n)\subset S_{a,b''}(n;1+\varepsilon)$$
$$\cap\, b\in B\cap\forall B',B''\subset B,\; S_{a,B''}(n)\cap S_{a,B'}(n)\neq\emptyset\Big\},$$

$$\mathfrak{B}_b^+(n)\stackrel{\text{def}}{=}\underset{B\in\mathfrak{B}_b(n)}{\text{Argmax}}\, B\quad\textit{(for the relation }\subset\textit{)}\,. \tag{4}$$

(Note that Argmax returns a set, contrary to argmax.) These sets are simply groups of points that are compatible with the properties that we expect from confidence intervals. They are thus the natural candidates for an algorithm that tries to aggregate observations from several users together.

4.1 The Agnostic `UCB` for clustered-bandits.

We are now ready to introduce **`A-UCB`**, whose pseudo-code is provided as Algorithm 4.

Proving strong regret bounds in this agnostic setting is difficult without further assumptions, since the true class may change at each single time step. For that reason, we proceed in two steps : Proposition 1 controls the number of pulls of sub-optimal arms under some events, that we then handle in specific cases in Lemma 1 and 2. Note that **`A-UCB`** uses a parameter γ that enables to control the enlargement coefficient ε and that we discuss below.

Algorithm 4 The **`A-UCB`** algorithm

Require: Parameter γ.

1: **for** $n=1...N$ **do**
2: Receive $b_n\sim\Upsilon$,
3: Compute $\widehat{\mu}_{a,b}(n-1)$, then $U_{a,b}(n-1),L_{a,b}(n-1),S_{a,b}(n-1)$ and $G_{a,b}(n-1)$.
4: Define the quantity $\varepsilon=\varepsilon_{b_n,b',n-1}$ by

$$\max\left\{\sqrt{\frac{2\gamma\log(N_{b'}(n-1))}{\log(N_{b_n}(n-1))}}-1,0\right\}.$$

5: Compute the set $\mathfrak{B}_{b_n}^+(n-1)$ of maximally compatible aggregation sets via (4).
6: Pull an elite arm that is the most optimistic

$$a_n\in\underset{a\in\mathcal{A}}{\text{argmax}}\max_{B\in\mathfrak{B}_{b_n}^+(n-1)}U_{a,B}(n-1) \tag{5}$$

7: **end for**

Proposition 1 (Control of the number of pulls of sub-optimal arms)
Let $\Omega_n=\left\{\mathcal{B}_{\mathbf{c}_n}\in\mathfrak{B}_{b_n}(n-1)\right\}$ *be the event that the true class* $\mathbf{c}_n$ *is admissible at round* n, *and* $\mathcal{E}_n^\alpha=\left\{G_{\star_{\mathbf{c}_n},\mathcal{B}_{\mathbf{c}_n}}(n-1)<\alpha\Delta_{a_n,\mathbf{c}_n}\right\}$ *the event that the confidence interval of the optimal arm of cluster* $\mathcal{B}_{\mathbf{c}_n}$ *is*

2. This is because we restrict to confidence interval centered around $\widehat{\mu}_{a,b}(n)$; in general we would need $G_{a,b'}(n)\leqslant\varepsilon\min\{U_{a,b}(n)-\widehat{\mu}_{a,b}(n),\widehat{\mu}_{a,b}(n)-L_{a,b}(n)\}$.

small enough, for small $\alpha \in (0,1)$. Then,[3] *for a suboptimal a_n, under $\Omega_n \cap \mathcal{E}_n^\alpha$ and for all $\eta \in (\alpha, 1]$,*

$$\text{either } N_{a_n,b_n}(n-1) < \Big(1+\frac{\varepsilon}{2}\Big)^2 \frac{24R^2 \log(N_{b_n}(n-1))}{(\eta-\alpha)^2 \Delta^2_{a_n,\mathbf{c}_n}},$$
$$\text{or } N_{a_n,\mathcal{B}_{\mathbf{c}_n}}(n-1) < \frac{24R^2 \log(N_{\mathcal{B}_{\mathbf{c}_n}}(n-1))}{(1-\eta)^2 \Delta^2_{a_n,\mathbf{c}_n}}.$$

That is, the total number of pulls, for either the current user b_n or its class $\mathbf{c}_n$, of a chosen sub-optimal arm is controlled. As a result, the regret is small under $\Omega_n \cap \mathcal{E}_n^\alpha$.

In particular for small ε, α and $\eta \to 1$, Proposition 1 shows that under $\Omega_n \cap \mathcal{E}_n^\alpha$ the regret of **A-UCB** is essentially in between that of **UCB** on $\mathcal{B}$ and **UCB** on $\mathcal{C}$: up to constants, it is never worse than **UCB** on $\mathcal{B}$, which is the naive baseline, and can be significantly better by competing occasionally with the oracle **UCB** on $\mathcal{C}$. This is highlighted precisely on Figure 5, where **A-UCB** behaves like **UCB** on $\mathcal{B}$, in the beginning, and then progressively behaves like **UCB** on $\mathcal{C}$. It now remains to show that $\Omega_n \cap \mathcal{E}_n^\alpha$ happens *with high probability* in order to deduce a non-trivial regret bound.

Illustration Ω_n is the event that the true class $\mathbf{c}_n$ is admissible at round n,. Now the event $\mathcal{E}_n^\alpha$ essentially says that $N_{\star_{c_n},\mathcal{B}_{c_n}}(n-1) > O(\log(n))$, that is, since $N_{\star_c,\mathcal{B}_c}(n) = \sum_{b\in\mathcal{B}_c} N_{\star_c,b}(n)$, it is enough that one $N_{\star_{c_n},b}(n)$ be as large to ensure that $\mathcal{E}_n^\alpha$ happens. For illustration, let us turn to the case of Bernoulli distributions ($R = 1/2$) with $C = 4$ equally probable classes of equal size $B = 50$. Individual upper bound confidence bounds $U_{a,b}(25000)$ are non trivial (i.e. less than 1) if (a,b) is seen at least 15 times. Now if each pair $(\star_c, b)$ for $b \in \mathcal{B}_c$ is visited at least 15 times (out of the $\simeq 125$ available time steps for each $b \in \mathcal{B}_c$) then $G_{\star_c,\mathcal{B}_c}(25000) < 0.27$, and for 50 visits, the bound reduces to 0.145. Similarly, for $B = 250$ we get abound 0.12 with 15 visits of the optimal action, which is enough to ensure that $\mathcal{E}_n^\alpha$ happens in non-trivial situations. Of course these numbers can be significantly reduced by using better confidence bounds (see Abbasi-Yadkori *et al.* (2011)). Let us now provide conditions under which both $\mathcal{E}_n^\alpha$ and Ω_n happen.

Adaptive enlargement Let us first deal with the event Ω_n that $\mathcal{B}_{c_n}$ is admissible. To that end, we resort to an adaptive enlargement ε. Indeed a constant ε (such that $\varepsilon = 1$) does not always ensure that $\mathcal{B}_{c_n}$ is admissible with high probability, but only that a *subset* of $\mathcal{B}_{c_n}$ is admissible at round n. In order to better understand the set of points that are gathered in $S_{a,b}(n; 1+\varepsilon)$ and that are admissible, let us introduce the following problem-dependent quantity, that only depends on the law of arrivals Υ :

Definition 1 (Internal balance of arrivals in a cluster)
The γ-balance of $\mathcal{B}$ with respect to the cluster c, for a point $b \in \mathcal{B}_c$ is defined by

$$\mathcal{B}_c(b;\gamma) = \Big\{ b' \in \mathcal{B}_c : \Upsilon(b) \leqslant \gamma \Upsilon(b') \Big\}.$$

Together with this quantity, it is natural to introduce the *distortion factor* of group $\mathcal{B}_c$, defined by

$$\gamma_c = \frac{\max_{b\in\mathcal{B}_c} \Upsilon(b)}{\min_{b\in\mathcal{B}_c} \Upsilon(b)}.$$

These quantities enable us to quantify the effective number of points that are grouped together with $b \in \mathcal{B}$. This directly defines the speed-up the algorithm can achieve for this environment. Importantly, note that if $\gamma \geqslant \gamma_c$, then it holds that $\mathcal{B}_c(b;\gamma) = \mathcal{B}_c$ for all $b \in \mathcal{B}_c$. **A-UCB** uses an adaptive ε that ensures that if γ is essentially greater than γ_c, then $\mathcal{B}_c(b;\gamma)$ and thus $\mathcal{B}_c$ is admissible with high probability (but one should choose a small γ since the regret is increasing with γ) ; more precisely

Lemma 1 (Probability that the true class is admissible)
*In **A-UCB**, if γ is chosen such that $\gamma \geqslant \gamma_c + O\big(n^{-1/2}\big)$, then it holds that*

$$\mathbb{P}(\Omega_n) \geqslant 1 - O\Big(n^{-2} A \sum_{b\in\mathcal{B}} \Upsilon(b)^{-2}\Big) - 2|\mathcal{B}|n^{-2}.$$

3. In section C.2 of the extended version Maillard & Mannor (2013), we show a slightly stronger result, though more difficult to interpret.

Such a $O(n^{-2})$ control is standard in regret proofs.

Ensuring the optimal arm is pulled enough We now turn to $\mathcal{E}_n^\alpha$. In full generality, there is no reason that **A-UCB** makes $\mathcal{E}_n^\alpha$ happen. The following lemma however ensures that under a mild condition on the structure of the problem, this actually holds with high probability. A simple regret bound follows trivially.

Lemma 2 (Probability of small-enough confidence intervals under mismatch assumption)
Let us assume that Υ is the uniform distribution, that all clusters have the same size B_0, and that the cluster distributions satisfy $\forall c, c' \in \mathcal{C}\, \forall a \in \mathcal{A}$

$$\text{either } \mu_{\star_c,c} - \mu_{\star_c,c'} < \Delta_{a,c}/2 \text{ or } \mu_{\star_c,c} - \mu_{\star_c,c'} > \frac{3}{2}\Delta_{a,c}\,.$$

(That is, a mismatch between two classes is either clear or harmless.) In such a case, if **A-UCB** *is run with $\gamma \sim \gamma_c = 1$, then $\mathbb{P}(\mathcal{E}_n^\alpha) \geqslant 1 - O(n^{-2})$ holds for $\alpha = 1/2$.*

Combining Proposition 1 together with Lemma 1 and Lemma 2, we deduce that, in some specific situations we are able to control with high probability the number of pulls of a sub-optimal arm, and as a result, the regret of the considered strategy. We currently do not know how to extend the analysis to handle the most general case. Note that the mismatch assumption in Lemma 2 is easy to check and holds in several situations (but we believe it is not necessary in order to get a controlled regret).

4.2 Numerical experiments

In this section, we study the behavior of the algorithm **A-UCB** on some experiments.

Algorithms We use the vanilla version of **UCB** (that aggregates all contexts), **UCB** on $\mathcal{B}$ that is the naive application of **UCB** separately on each context, and the oracle **UCB** on $\mathcal{C}$. We implemented a simplified version of **A-UCB** where we do not compute the maximally compatible sets exactly (which is NP-hard in general), but average the means of the compatible sets instead. This slightly worsen the numerical constants in our results, even though characterizing entirely the effect of this relaxation in terms of regret and numerical efficiency goes beyond the scope of this paper.

Experiments We consider experiments with Bernoulli distributions : this is intuitively the hardest case, since one can only rely on the means to separate distributions ; it also appears in several applications. For each experiment, we show the number of actions $|\mathcal{A}|$, of users $|\mathcal{B}|$, of classes $|\mathcal{C}|$, and the parameters $\{\mu_{a,c}\}_{a\in\mathcal{A},c\in\mathcal{C}}$ when there are not too many. We plot the regret of all algorithms on the same figure : A thick line is used for the mean regret and dashed lines for quantiles at levels $0.25, 0.5, 0.75, 0.95$ and 0.99. In all experiments, the parameters $\{\Upsilon(b)\}_{b\in\mathcal{B}}$ are defined by $\Upsilon(b) = w_b / \sum_{b\in\mathcal{B}} w_b$, where the weights w_b are drawn uniformly randomly in $[0.1, 0.9]$. Thus for each class, the distortion factor γ_c is less than 9, and we set the parameter γ of **A-UCB** to the value $\gamma = 9$. For one experiment with given fixed parameters, the algorithms are run over several trials (500) for a large time horizon $N = 25000$. We do not report the values of $\{\Upsilon(b)\}_{b\in\mathcal{B}}$ since this is generally uninformative.

Figure 2 presents an expected situation, where both the naive **UCB** and **UCB** on $\mathcal{B}$ perform poorly with respect to the oracle, whereas **A-UCB** performs very well. Note that here the best arm is different in the different classes, with corresponding value that is always very high and well separated from other arms.

Figure 3 presents a tricky situation : **UCB** on $\mathcal{B}$ performs poorly, while both **A-UCB** compete with the oracle, and all are defeated by **UCB**, which is not surprising since here one arm is the best in all contexts.

Figure 4 presents a variant when the set of actions $\mathcal{A}$ is large. As expected the performance of all algorithms degrade, but **A-UCB** is still competitive with respect to the oracle and benchmark algorithms.

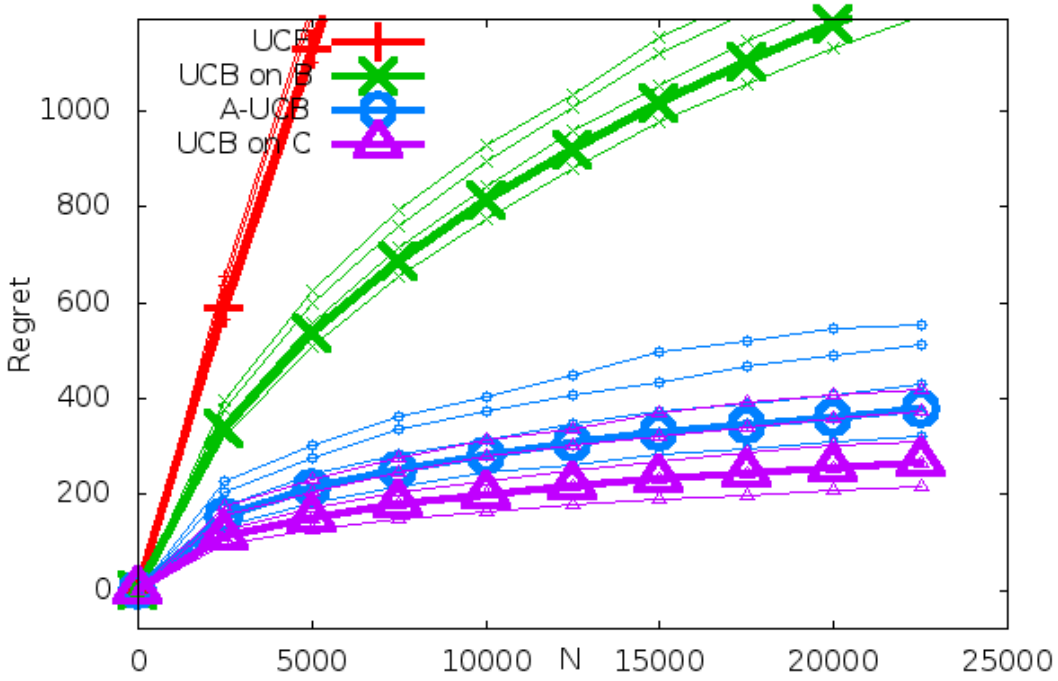

FIGURE 2 – Regret of several algorithms in the following scenario with $|\mathcal{A}| = 3, |\mathcal{B}| = 50, |\mathcal{C}| = 4$ and

$\mu_{a,c}$	1	2	3	4
1	0.527	0.209	**0.713**	**0.762**
2	**0.717**	0.193	0.575	0.230
3	0.669	**0.751**	0.120	0.485

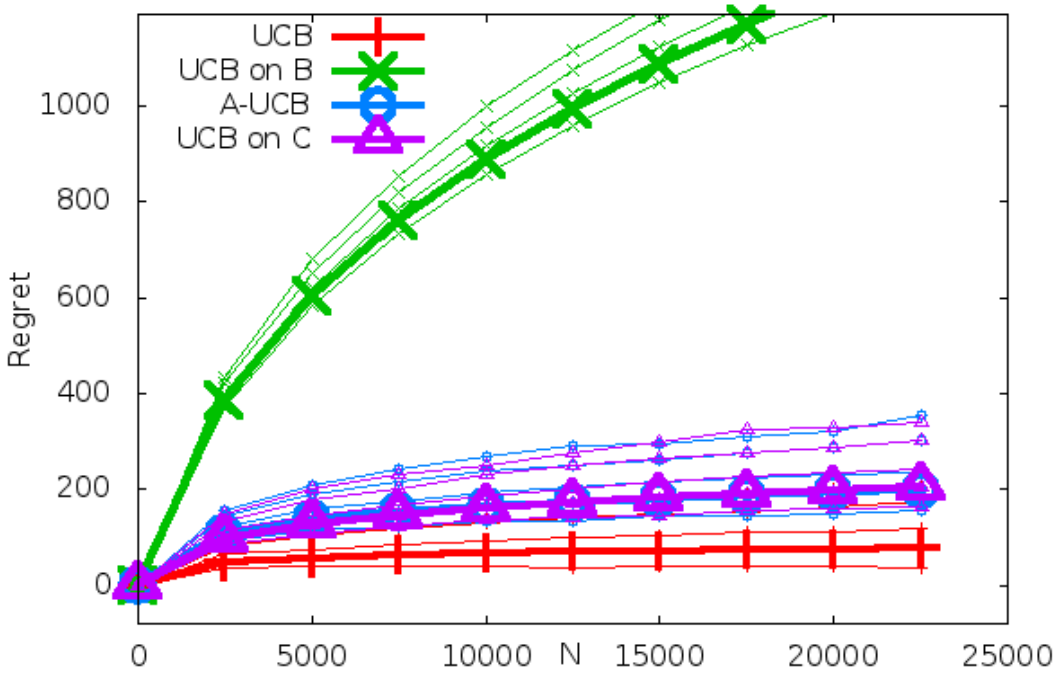

FIGURE 3 – Regret of several algorithms in the following scenario with $|\mathcal{A}| = 3, |\mathcal{B}| = 50, |\mathcal{C}| = 4$ and

$\mu_{a,c}$	1	2	3	4
1	0.370	0.750	0.609	0.207
2	0.150	0.290	0.475	0.464
3	**0.671**	**0.897**	**0.781**	**0.9**

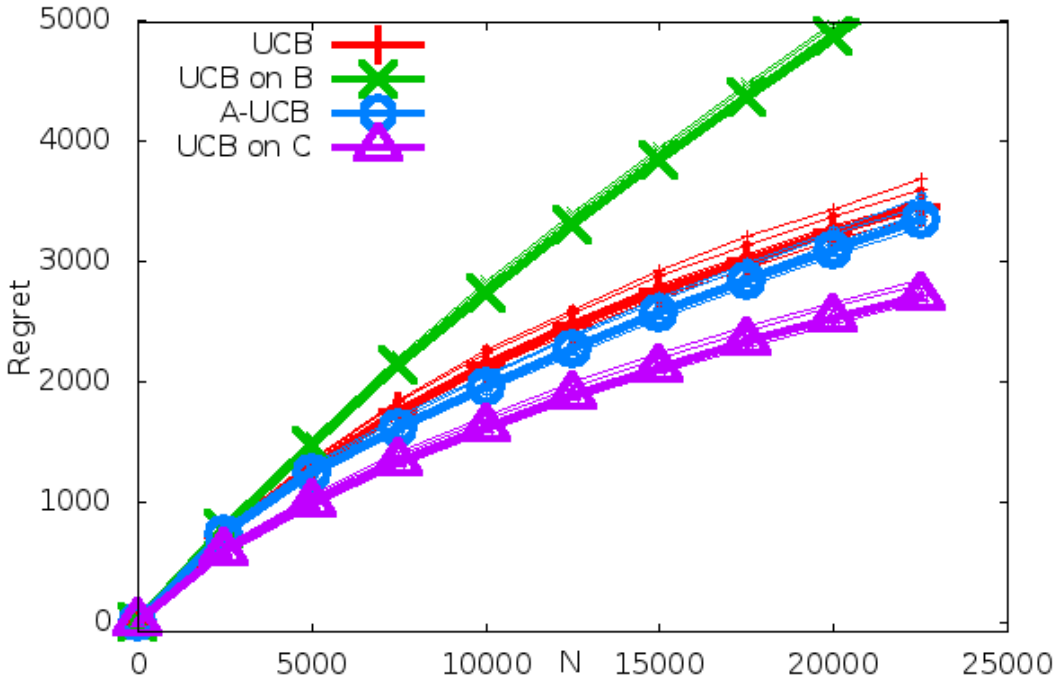

FIGURE 4 – Regret of several algorithms in some randomly generated situation with $|\mathcal{A}| = 50, |\mathcal{B}| = 50, |\mathcal{C}| = 4$.

Figure 5 presents a variant when the set of users $\mathcal{B}$ is large. Note that in this experiment, one only gets to see each b about 50 times, this setting is thus challenging. It can be seen that **A-UCB** still works fairly decently in this case. In accordance with Proposition 1, let us also remark that here **A-UCB** behaves initially like **UCB** on $\mathcal{B}$, and progressively behaves like **UCB** on $\mathcal{C}$ (though with a shifted regret due to the initial phase).

Finally figure 6 presents a variant when the number of clusters $\mathcal{C}$ is large. **A-UCB** still competes with the oracle here.

In all these experiments, we observe that **A-UCB** consistently competes with **UCB** on $\mathcal{C}$, while **UCB** and **UCB** on $\mathcal{B}$ sometimes obtain poor regret. This indicates that the proposed strategy is essentially able to capture the right information and does not under nor over-group the inputs b. This is promising.

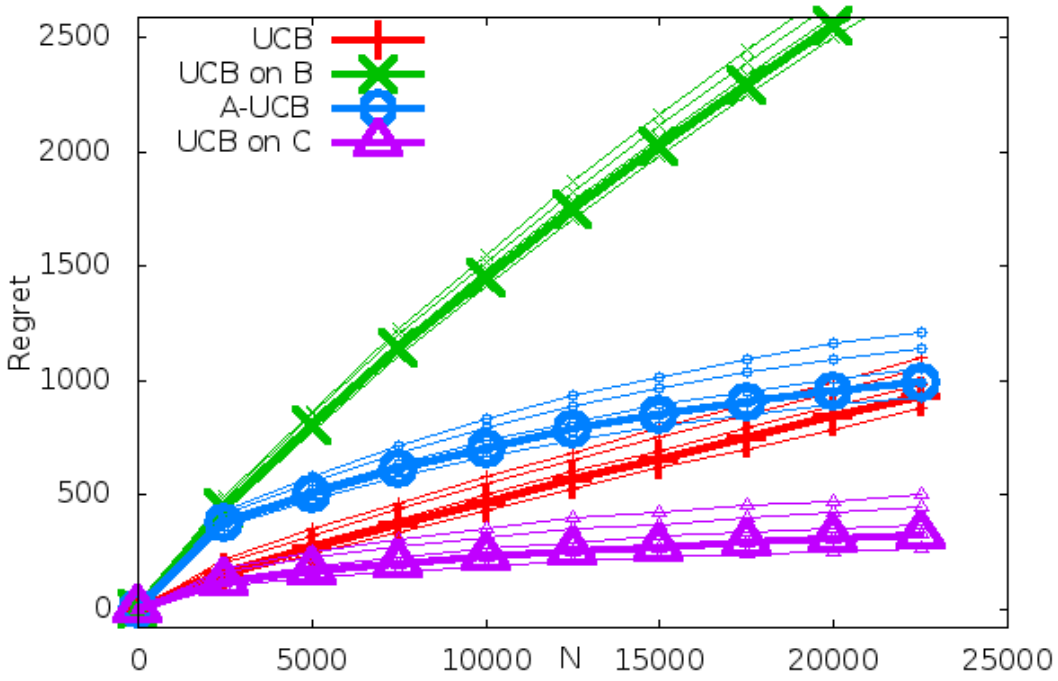

FIGURE 5 – Regret of several algorithms in the following scenario with $|\mathcal{A}| = 3, |\mathcal{B}| = 500, |\mathcal{C}| = 4$ and

$\mu_{a,c}$	1	2	3	4
1	0.1	0.621	0.1	**0.362**
2	**0.544**	**0.697**	**0.554**	0.181
3	0.512	0.409	0.234	0.1

5 Discussion

We introduced a novel setting for sequential decision making problem where there are some latent variables, such as in recommender systems, cognitive radio networks and others. We provided several contributions in a general framework in order to precisely address the issues raised by the latent structure. As a result, our contribution can be straightforwardly applied for instance to the linear-bandit setting (see Abbasi-Yadkori *et al.* (2011); Dani *et al.* (2008)), where the number of actions is replaced with the dimension of a feature space, and confidence intervals with confidence ellipsoids, and potentially many others.

Let us remark that we assumed in this work that the reward distributions are *clustered*, that is each $\nu_{a,b}$ is *one* of the $\{\nu_{a,c}\}_c$. A natural extension is to consider the case when each $\nu_{a,b}$ is a mixture of the $\{\nu_{a,c}\}_c$, with an underlying low-rank structure. This is left for future research.

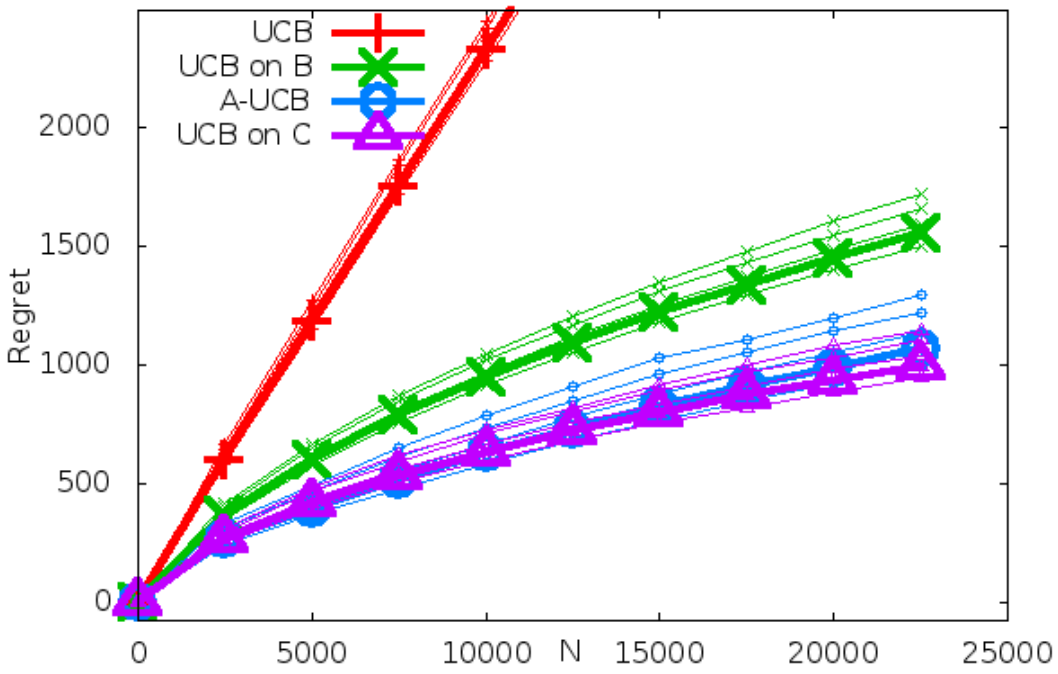

FIGURE 6 – Regret of several algorithms in some randomly generated situation with $|\mathcal{A}| = 3, |\mathcal{B}| = 100, |\mathcal{C}| = 50$.

In the non-trivial setting of Section 2, we showed that a simple procedure improves on Salomon & Audibert (2011) on the theoretical side and on Agrawal *et al.* (1989) on the computational side. We then introduced the more challenging setting of Section 3, that has not been addressed previously, and extended our procedure to that setting. We provided a lower-bound explaining why the setting is challenging and then a non trivial regret bound that makes appear explicitly the role of the distribution Υ of arrivals.

We finally tackled the agnostic setting, when not even the number of clusters is known. We introduced an algorithm that demonstrates excellent performance on a number of difficult situations, and provided a result enabling to derive regret guarantees in some non-trivial situations. We leave the intricate question of extending Lemma 1 and 2 to the *fully general* case as an open problem.

Acknowledgements This work was supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 306638 (SUPREL) and the Technion.

Références

ABBASI-YADKORI Y., PÁL D. & SZEPESVÁRI C. (2011). Improved algorithms for linear stochastic bandits. In J. SHAWE-TAYLOR, R. S. ZEMEL, P. L. BARTLETT, F. C. N. PEREIRA & K. Q. WEINBERGER, Eds., *Advances in Neural Information Processing Systems*, p. 2312–2320.

ADOMAVICIUS G. & TUZHILIN A. (2005). Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, **17**(6), 734–749.

AGRAWAL R., TENEKETZIS D. & ANANTHARAM V. (1989). Asymptotically Efficient Adaptive Allocation Schemes for Controlled I.I.D. processes. *IEEE Transactions on Automatic Control*, **34**(3), 258–267.

AUER P. (2003). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, **3**, 397–422.

AVNER O., MANNOR S. & SHAMIR O. (2012). Decoupling exploration and exploitation in multi-armed bandits. In *Proceedings of the 29th International conference on Machine Learning* : Omnipress.

BURNETAS A. N. & KATEHAKIS M. N. (1996). Optimal adaptive policies for sequential allocation problems. *Adv. Appl. Math.*, **17**(2), 122–142.

CAPPÉ O., GARIVIER A., MAILLARD O.-A., MUNOS R. & STOLTZ G. (2013). Kullback–leibler upper confidence bounds for optimal sequential allocation. *Ann. Statist.*, **41**(3), 1516–1541.

DANI V., HAYES T. P. & KAKADE S. M. (2008). Stochastic Linear Optimization under Bandit Feedback. In R. A. SERVEDIO, T. ZHANG, R. A. SERVEDIO & T. ZHANG, Eds., *COLT*, p. 355–366 : Omnipress.

FILIPPI S., CAPPÉ O., CÉROT F. & MOULINES E. (2008). A near optimal policy for channel allocation in cognitive radio. In S. GIRGIN, M. LOTH, R. MUNOS, P. PREUX & D. RYABKO, Eds., *Recent Advances in Reinforcement Learning*, volume 5323 of *Lecture Notes in Computer Science*, p. 69–81. Springer Berlin Heidelberg.

GARIVIER A. & MOULINES E. (2008). On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems. *ArXiv e-prints*. Technical report, LTCI.

GHESHLAGHI AZAR M., LAZARIC A. & BRUNSKILL E. (2013). Sequential transfer in multi-armed bandit with finite set of models. In C. BURGES, L. BOTTOU, M. WELLING, Z. GHAHRAMANI & K. WEINBERGER, Eds., *Advances in Neural Information Processing Systems 26*, p. 2220–2228. Curran Associates, Inc.

HAZAN E. & MEGIDDO N. (2007). Online learning with prior knowledge. In *Proceedings of the 20th annual conference on Learning theory*, COLT'07, p. 499–513, Berlin, Heidelberg : Springer-Verlag.

LAI T. L. & ROBBINS H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, **6**(1), 4–22.

LANGFORD J. & ZHANG T. (2007). The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information. In J. C. PLATT, D. KOLLER, Y. SINGER, S. T. ROWEIS, J. C. PLATT, D. KOLLER, Y. SINGER & S. T. ROWEIS, Eds., *NIPS* : MIT Press.

LI L., CHU W., LANGFORD J. & SCHAPIRE R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, WWW '10, p. 661–670, New York, NY, USA : ACM.

LI L., CHU W., LANGFORD J. & WANG X. (2011). Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, p. 297–306, New York, NY, USA : ACM.

LU T., PÁL D. & PAL M. (2010). Contextual multi-armed bandits. *Journal of Machine Learning Research - Proceedings Track*, **9**, 485–492.

MAILLARD O.-A. & MANNOR S. (2013). Latent bandits. *HAL/Open archive*.

SALOMON A. & AUDIBERT J.-Y. (2011). Deviations of stochastic bandit regret. In *Proceedings of the 22nd international conference on Algorithmic learning theory*, ALT'11, p. 159–173, Berlin, Heidelberg : Springer-Verlag.

SLIVKINS A. (2011). Contextual bandits with similarity information. In *Proceedings of the 24th annual conference on Learning theory*, COLT'11 : Springer-Verlag.