Latent Bandits.

Odalric-Ambrym Maillard, Shie Mannor

Latent Bandits.

Odalric-Ambrym Maillard ODALRIC-AMBRYM.MAILLARD@ENS-CACHAN.ORG
The Technion, Faculty of Electrical Engineering 32000 Haifa, ISRAEL

Shie Mannor SHIE@EE.TECHNION.AC.IL
The Technion, Faculty of Electrical Engineering 32000 Haifa, ISRAEL

Abstract

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 several difficult scenarios.

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

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

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

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

$\Delta_{a,c} = \mu_{\star_c,c} - \mu_{a,c}$. 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).

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 supplementary material.

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:

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

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 $\widehat{\mu}_{a,b}(n)$.

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 $0.2, 0.6, 0.8$ for one cluster, and Bernoulli distributions of parameter $0.8, 0.1, 0.5$ for the second one. This modifies the achievable learning 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 $\mathcal{A}_- = \mathcal{A} \setminus$*

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

$\{\star_{\mathbf{c}}\}$ *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 $C(\mathbf{c}) = \Big\{c' \in \mathcal{C} : \nu_{\star_{\mathbf{c}},c'} = \nu_{\star_{\mathbf{c}},\mathbf{c}}$ *and* $\star_{\mathbf{c}} \neq \star'_c\Big\}$.

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

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

Although theoretically appealing, it may be in general expensive to compute the quantities $\{\omega^\star_c\}_{c\in\mathcal{C}}$, 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 performance bound:

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*

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

`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 $\mathcal{C}_{n-1} = \Big\{c \in \mathcal{C}, \forall a \in \mathcal{A} : \mu_{a,c} \in S_{a,\mathcal{B}}(n-1)\Big\}$ 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)
$$a_n = \operatorname*{argmax}_{a=\star_c, c\in C_{n-1}} \mu_{\star_c,c}\,. \quad (2)$$
6: **end for**

Regret bound Such algorithm enjoys the following regret performance:

Theorem 4 *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 $\mathcal{A}^\star = \Big\{a \in \mathcal{A} : \exists c \in \mathcal{C}$ *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 *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*

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

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.

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

Require: The cluster distributions $\{\nu_{a,c}\}_{a\in\mathcal{A},c\in\mathcal{C}}$.
for $n=1...N$ **do**
  Receive $b=b_n\sim\Upsilon$.
  Define the set of admissible classes $\mathcal{C}_{n-1}(b)=\Big\{c\in\mathcal{C},\forall a\in\mathcal{A}:\mu_{a,c}\in S_{a,b}(n-1)\Big\}$.
  Define the set of "elite" admissible arms $\mathcal{A}^\star_{n-1}=\{a\in\mathcal{A};\exists c\in\mathcal{C}_{n-1}(b_n)\ \star_c=a\}$.
  Choose the most optimistic "elite" arm
  $$a_n=\operatorname*{argmax}_{a=\star_c,\,c\in\mathcal{C}_{n-1}(b_n)}\mu_{\star_c,c}\,.$$
end for

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.

Theorem 6 *The regret of* `Multiple-K-UCB` *satisfies*

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

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

for $n=1...N$ **do**
  Receive $b_t\sim\Upsilon$.
  Compute the empirical means $\widehat{\mu}_{a,b}(n-1)$.
  Choose the next arm (breaks ties arbitrary)
  $$a_n=\operatorname*{argmax}_{a\in\mathcal{A}}U_{a,b_n}(n-1)\,. \quad (3)$$
end for

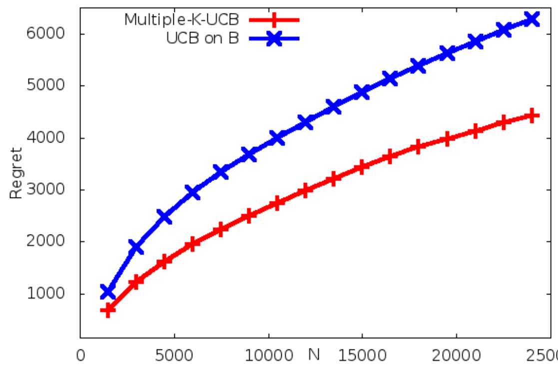

Figure 1. Theoretical regret bounds for `Multiple-K-UCB` (Theorem 6) and `UCB` on $\mathcal{B}$ (Theorem 7) for one problem characterized by $|\mathcal{A}|=3,|\mathcal{B}|=50,|\mathcal{C}|=4$ and

		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

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 cluster distributions, we obtained regret bounds that may significantly improve on their equivalent agnostic version. We now detail a second improvement that seems even more effective and is applicable both in case cluster distributions are known or not.

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 a (pseudo-)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 a `UCB` on $\mathcal{B}$ algorithm, to that of the pseudo-oracle `UCB` on $\mathcal{C}$.

Theorem 7 *The expected regret at time N of the algorithm* `UCB` *on $\mathcal{B}$ is upper bounded by*

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

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*

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

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 substantially 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 $\sum_{c=1}^{C}\sum_{a\in\mathcal{A}}\Delta_{a,c}^{-1}$, that can be substantial, since typically C is much smaller than B. Note of course that the partition $\mathcal{C}$ is unknown in practice. 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 use. 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 set $S_{a,B}(n)$, where

$$\widehat{\nu}_{a,B}(n) = \frac{\sum_{b'\in\mathcal{B}}\widehat{\nu}_{a,b'}(n)N_{a,b'}(n)\mathbb{I}\{b'\in B\}}{\sum_{b'\in\mathcal{B}}N_{a,b'}(n)\mathbb{I}\{b'\in B\}},$$

$$\mu_{a,B}(n) = \frac{\sum_{b'\in\mathcal{B}}\mu_{a,b'}N_{a,b'}(n)\mathbb{I}\{b'\in B\}}{\sum_{b'\in\mathcal{B}}N_{a,b'}(n)\mathbb{I}\{b'\in B\}}.$$

Note that for $B=\mathcal{B}_c$, then $\mu_{a,\mathcal{B}_c}(n)=\mu_{a,c}$, which may not hold for other sets B since there may be a bias when the $\{\mu_{a,b'}\}_{b'\in\mathcal{B}}$ are distinct. However, the speed of convergence of the group depends on $N_{a,B}(n)=\sum_{b'\in\mathcal{B}}N_{a,b'}(n)\mathbb{I}\{b'\in\mathcal{B}\}$, which is typically much faster than that of a single point b (that depends on $N_{a,b}(n)$). Thus $S_{a,B}(n)=[L_{a,B}(n),U_{a,B}(n)]$ 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 estimation 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, some 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

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

and then $S_{a,b}(n;1+\varepsilon)=[L_{a,b}(n;1+\varepsilon),U_{a,b}(n;1+\varepsilon)]$. Now, if $\mu_{a,b}=\mu_{a,b'}$ and $G_{a,b'}(n)\leqslant\frac{\varepsilon}{2}G_{a,b}(n)$[2], we must have $S_{a,b'}(n)\subset S_{a,b}(n;1+\varepsilon)$ with high probability. Finally, we focus only on mean-based procedures for clarity, but it is possible to use empirical distributions $\widehat{\nu}_{a,b}(n)$ to remove b' with obvious mismatch in Kullback-Leibler divergence. We do not discuss this.

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:

$$\mathfrak{B}_b(n)\stackrel{\text{def}}{=}\Big\{B\subset\mathcal{B}:\forall a\in\mathcal{A}\,\forall b',b''\in B\ S_{a,b'}(n)\subset S_{a,b''}(n;1+\varepsilon)\cap b\in B\cap\forall B',B''\subset B,\ S_{a,B''}(n)\cap S_{a,B'}(n)\neq\emptyset\Big\},$$
$$\mathfrak{B}_b^+(n)\stackrel{\text{def}}{=}\operatorname*{Argmax}_{B\in\mathfrak{B}_b(n)} B\quad(\textit{for the relation}\subset). \tag{4}$$

(Note that Argmax returns a set, contrary to argmax.)

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.

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

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

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

7: **end for**

Proposition 1 *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 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 $\eta \in (\alpha, 1]$,*

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

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

That is, in all cases the total number of pulls, for either the current user b_n or its class $\mathbf{c}_n$, of a chosen suboptimal arm is controlled.

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 $\mathcal{B}$, and can be significantly better by competing occasionally with UCB on $\mathcal{C}$. This is highlighted on Figure 5. 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 $N_{\star_{\mathbf{c}_n},\mathcal{B}_{\mathbf{c}_n}}(n-1) > O(\log(n))$, 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 $N_{\star_{\mathbf{c}_n},b}(n)$ 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 The reason for having an adaptive ε and not just a constant $\varepsilon = 1$ is that a constant ε does not always ensure that $\mathcal{B}_{c_n}$ is admissible (that is Ω_n happens) with high probability, but only that a *subset* of $\mathcal{B}_{c_n}$ is admissible at round n. To better understand the number of such points that are gathered in $S_{a,b}(n; 1+\varepsilon)$ we introduce the following quantity, that only depends on the law of arrivals Υ:

Definition 1 *The γ-balance of $\mathcal{B}$ with respect to cluster c, for point $b \in \mathcal{B}_c$ is defined by*

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

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 with $b \in \mathcal{B}$, which 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 scales with γ); more precisely

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

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

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 *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 $\alpha = 1/2$.

[3]In Appendix C.2, we show a slightly stronger result, though more difficult to interpret.

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 pseudo-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.

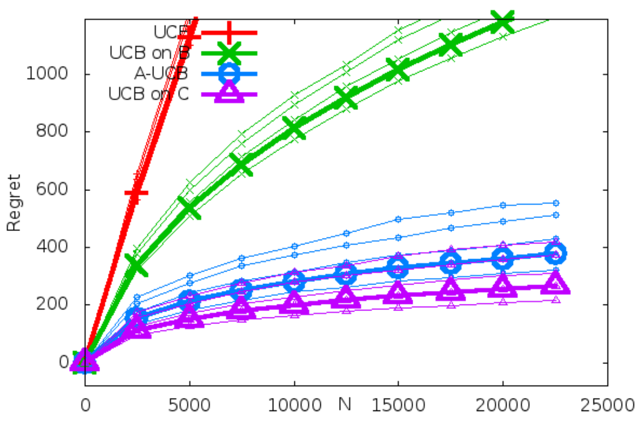

Figure 2. Regret of several algorithms in the following scenario with $|\mathcal{A}| = 3, |\mathcal{B}| = 50, |\mathcal{C}| = 4$ and

$\mu_{a,c}$	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 2 presents an expected situation, where both the naive UCB and UCB on $\mathcal{B}$ perform poorly with respect to the pseudo-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 pseudo-oracle, and all are defeated by UCB, which is not surprising since here one arm is the best in all contexts.

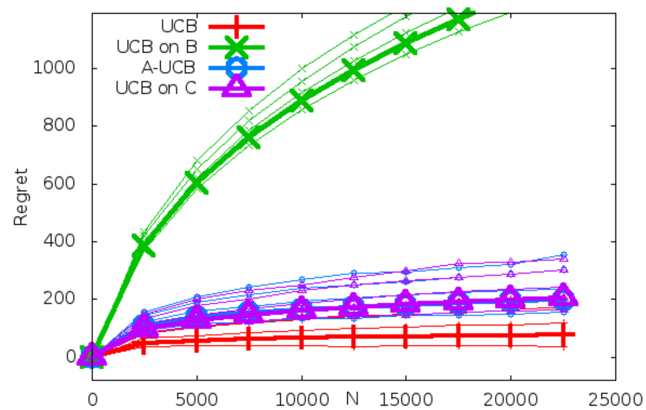

Figure 3. Regret of several algorithms in the following scenario with $|\mathcal{A}| = 3, |\mathcal{B}| = 50, |\mathcal{C}| = 4$ and

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

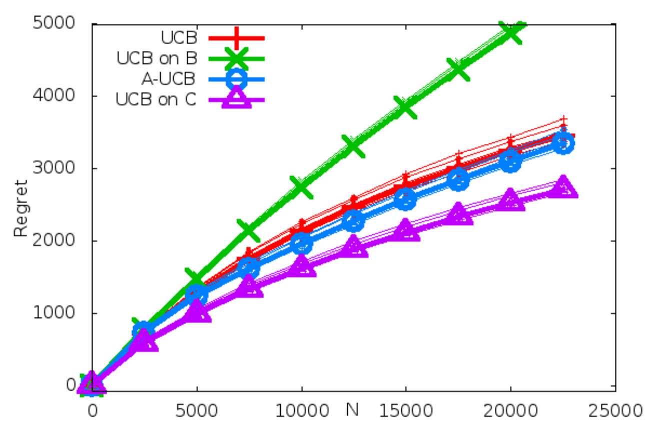

Figure 4. Regret of several algorithms in some randomly generated situation with $|\mathcal{A}| = 50, |\mathcal{B}| = 50, |\mathcal{C}| = 4$.

Figure 5 presents a variant when $\mathcal{B}$ 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).

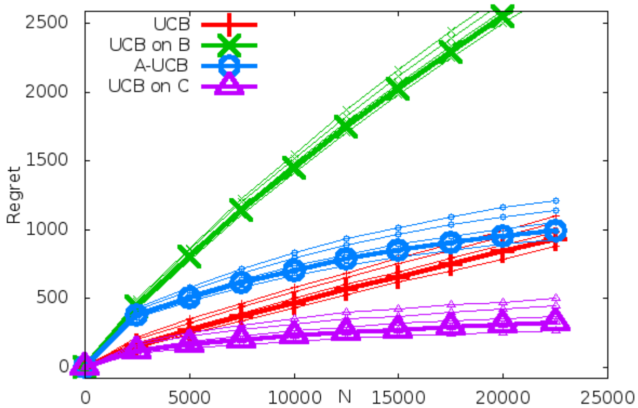

Figure 5. Regret of several algorithms in the following scenario with $|\mathcal{A}| = 3, |\mathcal{B}| = 500, |\mathcal{C}| = 4$ and

$\mu_{a,c}$	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

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

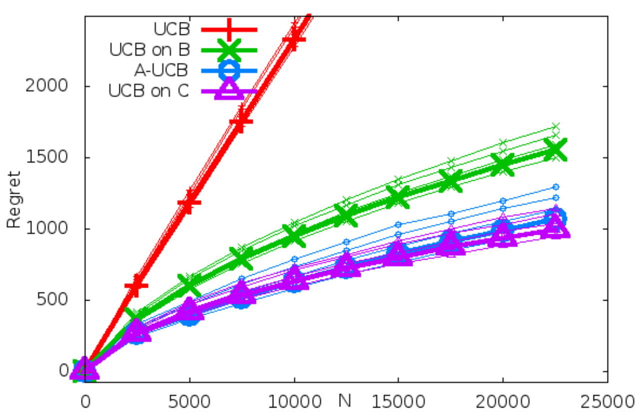

Figure 6. Regret of several algorithms in some randomly generated situation with $|\mathcal{A}| = 3, |\mathcal{B}| = 100, |\mathcal{C}| = 50$.

In all these experiments, we see 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.

5. Discussion

We introduced a novel setting for sequential decision making problem where there are some latent variables, such as 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.

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 situation. 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.

References

Abbasi-Yadkori, Yasin, Pál, Dávid, and Szepesvári, Csaba. Improved algorithms for linear stochastic bandits. In Shawe-Taylor, John, Zemel, Richard S., Bartlett, Peter L., Pereira, Fernando C. N., and Weinberger, Kilian Q. (eds.), *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Adomavicius, Gediminas and Tuzhilin, Alexander. 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, jun 2005. ISSN 1041-4347.

Agrawal, Rajeev, Teneketzis, Demosthenis, and Anantharam, Venkatachalam. Asymptotically Efficient Adaptive Allocation Schemes for Controlled I.I.D. processes. *IEEE Transactions on Automatic Control*, 34(3):258–267, 1989.

Auer, Peter. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, mar 2003. ISSN 1532-4435.

Auer, Peter, Cesa-Bianchi, Nicolò, Freund, Yoav, and Schapire, Robert E. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2003.

Avner, Orly, Mannor, Shie, and Shamir, Ohad. Decoupling exploration and exploitation in multi-armed bandits. In *Proceedings of the 29th International conference on Machine Learning*. Omnipress, 2012.

Bubeck, Sébastien. *Bandits Games and Clustering Foundations*. PhD thesis, Université Lille 1, 2010.

Burnetas, Apostolos N. and Katehakis, Michael N. Optimal adaptive policies for sequential allocation problems. *Adv. Appl. Math.*, 17(2):122–142, jun 1996. ISSN 0196-8858.

Cappé, Olivier, Garivier, Aurélien, Maillard, Odalric-Ambrym, Munos, Rémi, and Stoltz, Gilles. Kullback–leibler upper confidence bounds for optimal sequential allocation. *Ann. Statist.*, 41(3):1516–1541, 2013. ISSN 0090-5364.

Cesa-Bianchi, Nicolò and Lugosi, Gábor. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

Dani, Varsha, Hayes, Thomas P., and Kakade, Sham M. Stochastic Linear Optimization under Bandit Feedback. In Servedio, Rocco A., Zhang, Tong, Servedio, Rocco A., and Zhang, Tong (eds.), *COLT*, pp. 355–366. Omnipress, 2008.

Filippi, Sarah, Cappé, Olivier, Cérot, Fabrice, and Moulines, Eric. A near optimal policy for channel allocation in cognitive radio. In Girgin, Sertan, Loth, Manuel, Munos, Rémi, Preux, Philippe, and Ryabko, Daniil (eds.), *Recent Advances in Reinforcement Learning*, volume 5323 of *Lecture Notes in Computer Science*, pp. 69–81. Springer Berlin Heidelberg, 2008.

Garivier, Aurélien and Moulines, Eric. On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems. *ArXiv e-prints*, may 2008. Technical report, LTCI.

Hazan, Elad and Megiddo, Nimrod. Online learning with prior knowledge. In *Proceedings of the 20th annual conference on Learning theory*, COLT'07, pp. 499–513, Berlin, Heidelberg, 2007. Springer-Verlag.

Lai, Tze L. and Robbins, Herbert. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Langford, John and Zhang, Tong. The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information. In Platt, John C., Koller, Daphne, Singer, Yoram, Roweis, Sam T., Platt, John C., Koller, Daphne, Singer, Yoram, and Roweis, Sam T. (eds.), *NIPS*. MIT Press, 2007.

Li, Lihong, Chu, Wei, Langford, John, and Schapire, Robert E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pp. 661–670, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8.

Li, Lihong, Chu, Wei, Langford, John, and Wang, Xuanhui. 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, pp. 297–306, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0493-1.

Lu, Tyler, Pál, Dávid, and Pal, Martin. Contextual multi-armed bandits. *Journal of Machine Learning Research - Proceedings Track*, 9:485–492, 2010.

Salomon, Antoine and Audibert, Jean-Yves. Deviations of stochastic bandit regret. In *Proceedings of the 22nd international conference on Algorithmic learning theory*, ALT'11, pp. 159–173, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-24411-7.

Slivkins, Aleksandrs. Contextual bandits with similarity information. In *Proceedings of the 24th annual conference on Learning theory*, COLT'11. Springer-Verlag, 2011.

A. Technical details for the case of Known cluster distributions with single cluster arrivals.

In this section, we prove Theorem 4.

A.1. Proof of Theorem 4

Step 1. Let us define the event Ω_n that the class $\mathbf{c} \in \mathcal{C}$ corresponding to the incoming $b_n \in \mathcal{B}$ is admissible, that is

$$\Omega_n = \Big\{ \mathbf{c} \in \mathcal{C}_{n-1} \Big\} .$$

We have the property that

$$\begin{aligned}
\mathbb{P}(\overline{\Omega}_n) &= \mathbb{P}\Big(\exists a \in \mathcal{A}^\star : \mu_{a,c} \notin S_{a,\mathcal{B}}(n-1) \Big) \\
&\leqslant \sum_{a \in \mathcal{A}^\star} \mathbb{P}\Big(|\widehat{\mu}_{a,\mathcal{B}}(n-1) - \mu_{a,c}| > \frac{1}{2} G_{a,\mathcal{B}}(n-1) \Big) \\
&\leqslant \sum_{a \in \mathcal{A}^\star} \mathbb{P}\Big(\bigcup_{k=1}^{n-1} |\widehat{\mu}_{a,\mathcal{B},k} - \mu_{a,c}| > \frac{1}{2} G_{a,\mathcal{B},k}^{n-1} \Big) \\
&\leqslant \sum_{a \in \mathcal{A}^\star} \sum_{k=1}^{n-1} \frac{2}{(n-1)^3} = \frac{2A}{(n-1)^2} .
\end{aligned}$$

Here in the second line, we introduced $\widehat{\mu}_{a,\mathcal{B},k}$ and $G_{a,\mathcal{B},k}^{n-1}$ to be the values of $\widehat{\mu}_{a,\mathcal{B}}(n-1)$ and $G_{a,\mathcal{B}}(n-1)$ when $N_{a,\mathcal{B}}(n-1) = k$, and in the last line, we made use of a simple concentration of measure property for sub-Gaussian random variables. Let us note that with our writing conventions

$$G_{a,\mathcal{B},k}^{n-1} = 2R\sqrt{\frac{2\log\big((n-1)^3\big)}{k}} .$$

Step 2. Now, under the event Ω_n, let us consider that the `Single-K-UCB` algorithm plays $a_n = \star_{\tilde{c}}$ for some $\tilde{c} \in \mathcal{C}_{n-1}$. Note that since $\tilde{c}, \mathbf{c} \in \mathcal{C}_{n-1}$, we must have

$$\begin{aligned}
\mu_{\star_{\tilde{c}},\tilde{c}} - \mu_{\star_{\tilde{c}},\mathbf{c}} &\leqslant |\mu_{\star_{\tilde{c}},\tilde{c}} - \widehat{\mu}_{\star_{\tilde{c}},\mathcal{B}}(n-1)| \\
&\quad + |\widehat{\mu}_{\star_{\tilde{c}},\mathcal{B}}(n-1) - \mu_{\star_{\tilde{c}},\mathbf{c}}| \\
&\leqslant G_{\star_{\tilde{c}},\mathcal{B}}(n-1) .
\end{aligned}$$

Note also that by the optimistic choice of arm, we have $\mu_{\star_{\tilde{c}},\tilde{c}} \geqslant \mu_{\star_{\mathbf{c}},\mathbf{c}} \geqslant \mu_{\star_{\tilde{c}},\mathbf{c}}$. Thus, if $a_n = a \neq \star_{\mathbf{c}}$, we conclude that under Ω_n we must have $G_{a,\mathcal{B}}(n-1) \geqslant \Delta^+_{a,\mathbf{c}}$ where

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

Step 3. We now decompose the number of pulls of a sub-optimal arm $a \neq \star_{\mathbf{c}}$, for any $u_a \in \mathbb{N} \setminus \{0\}$, into

$$\begin{aligned} N_{a,\mathbf{c}}(N) &\leqslant u_a + \sum_{n=u_a+1}^{N} \mathbb{I}\{a_n = a \cap N_{a,\mathbf{c}}(n-1) > u_a\} \\ &\leqslant u_a + \sum_{n=u_a+1}^{N} \mathbb{I}\{a_n = a \cap N_{a,\mathbf{c}}(n-1) > u_a \cap \Omega_n\} \\ &\quad + \sum_{n=2}^{N} \mathbb{I}\{\overline{\Omega}_n\}\,. \end{aligned}$$

Now from the result of Step 1, we obtain

$$\mathbb{E}\Big[\sum_{n=2}^{N} \mathbb{I}\{\overline{\Omega}_n\}\Big] \leqslant \frac{A\pi^2}{3}\,.$$

On the other hand, from the result of step 2, we can choose u_a to be

$$u_a = \left\lceil \frac{24R^2\log(N)}{\Delta_{a,\mathbf{c}}^{+\,2}} \right\rceil,$$

which ensures that $N_{a,\mathbf{c}}(n-1) > u_a$ can not happen simultaneously with $G_{a,\mathcal{B}}(n-1) \geqslant \Delta^+_{a,\mathbf{c}}$ under Ω_n. Thus, we deduce that

$$\mathbb{E}\Big[N_{a,\mathbf{c}}(N)\Big] \leqslant \frac{24R^2\log(N)}{\Delta_{a,\mathbf{c}}^{+\,2}} + 1 + \frac{A\pi^2}{3}\,,$$

and then the final regret bound

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

B. Technical details for the case of Known cluster distributions with multiple cluster arrivals.

In this section, we prove the lower bound of Theorem 5, and then the regret bound of Theorem 6.

B.1. Proof of Theorem 5

It is actually easier to start proving the result in the agnostic case, and then extend it to the setting when we know the cluster distributions (which is *a priori* harder, since we give more information to the learner). Indeed in the agnostic case the proof of the lower bound is a simple adaptation of the lower bound from (Cesa-Bianchi & Lugosi, 2006) for standard stochastic multi-armed bandits. The main insight is that in case of a uniform distribution of users, then $\mathbb{E}\Big[N_{\mathcal{B}_c}(N)\Big] = N\Upsilon(\mathcal{B}_c) = \frac{N}{C}$, and $N_{\mathcal{B}_c}(N)$ is also close to $\mathbb{E}\Big[N_{\mathcal{B}_c}(N)\Big]$ in high probability.

More precisely from (Cesa-Bianchi & Lugosi, 2006), it suffices to introduce Bernoulli distributions with parameter $\frac{1\pm\varepsilon_c}{2}$ for each class c (with only one arm such having parameter $\frac{1+\varepsilon_c}{2}$ in each class). Now by application of Lemma 2.2 in (Bubeck, 2010), we have for any forecaster the following inequality:

$$\begin{aligned} &\sup_{a_c} \mathbb{E}\sum_{n=1}^{N} \mu_{a_c,c} - \mu_{a_n,c} \\ &\geqslant N\varepsilon_c\Big(1 - \frac{1}{A} - \sqrt{\frac{N\varepsilon_c}{2A}\log\big(\frac{1+\varepsilon_c}{1-\varepsilon_c}\big)}\Big). \end{aligned}$$

Using such a result, it is then possible to optimize ε_c and get the final lower bound after some tedious computations, similarly to Theorem 2.6 in (Bubeck, 2010), see also (Auer et al., 2003).

Now the reason why the bound also holds in the setting when we know the cluster distributions in advance is that since $C > A$, one can always choose arms such that the set of elite arms is not smaller than $\mathcal{A}$, that is no arm can be removed a-priori. In such a case, the learner still has to sample enough to be able to separate the optimal arms from others, so that its a priori information does not help reducing the regret.

B.2. Proof of Theorem 6

We analyze here the regret of `Multiple-K-UCB`. Let us denote for convenience $\mathcal{A}^\star = \{a \in \mathcal{A} : \exists c \in \mathcal{C}, \star_c = a\}$ the set of elite arms. Note that by definition only arms in $\mathcal{A}^\star$ can be pulled by the algorithm.

Step 1. We first control the number of arrivals of each $b \in \mathcal{B}$. By a simple application of Bernstein's inequality, we have the property that for all $\delta \in [0,1]$,

$$\mathbb{P}\bigg(\Big|N_b(n) - n\Upsilon(b)\Big| \geqslant \sqrt{\Upsilon(b)(1-\Upsilon(b))2n\log(1/\delta)} + \frac{\log(1/\delta)}{3}\bigg) \leqslant 2\delta.$$

Thus, if we introduce for convenience the two following quantities

$$\begin{aligned} n^+_{\Upsilon(b)}(\delta) &= n\Upsilon(b) + \sqrt{\Upsilon(b)(1-\Upsilon(b))2n\log(1/\delta)} + \frac{\log(1/\delta)}{3}, \\ n^-_{\Upsilon(b)}(\delta) &= n\Upsilon(b) - \sqrt{\Upsilon(b)(1-\Upsilon(b))2n\log(1/\delta)} - \frac{\log(1/\delta)}{3}, \end{aligned}$$

then we have the property that with probability higher than $1-2\delta$,

$$n^-_{\Upsilon(b)}(\delta) \leqslant N_b(n) \leqslant n^+_{\Upsilon(b)}(\delta)\,.$$

Now, by a simple concentration bound for sub-Gaussian random variables together with a rough union bound, we have that for all $a \in \mathcal{A}, b \in \mathcal{B}$,

$$\begin{aligned}
&\mathbb{P}\Big(\mu_{a,c_b} \notin S_{a,b}(n)\Big)\\
&\leqslant \mathbb{P}\Big(\bigcup_{k=1}^{N_b(n)} \frac{1}{k}\sum_{s=1}^{k}\Big|X_{a,b}(s) - \mu_{a,b}\Big| \geqslant R\sqrt{\frac{6\log\big(N_b(n)\big)}{k}}\Big)\\
&\leqslant \mathbb{P}\Big(\bigcup_{k=1}^{n^+_{\Upsilon(b)}(\delta)} \frac{1}{k}\sum_{s=1}^{k}\Big|X_{a,b}(s) - \mu_{a,b}\Big| \geqslant R\sqrt{\frac{6\log\big(n^-_{\Upsilon(b)}(\delta)\big)}{k}}\Big)\\
&\quad + 2\delta\\
&\leqslant \sum_{k=1}^{n^+_{\Upsilon(b)}(\delta)} n^-_{\Upsilon(b)}(\delta)^{-3} + 2\delta = n^+_{\Upsilon(b)}(\delta)n^-_{\Upsilon(b)}(\delta)^{-3} + 2\delta\,.
\end{aligned}$$

Note that one could derive even tighter bounds resorting to self-normalized bounds, at the price of a more complicated expression for the high-probability level. We do not use this here since our goal is simply to achieve a regret bound with right order of magnitude but not necessarily optimized constants. Finally, the same bound holds for the symmetric sum $\frac{1}{N_{a,b}(n)}\sum_{t=1}^{n}\Big(\mu_{a,b} - X_t\Big)\mathbb{I}\{a_t = a, b_t = b\}$ as well.

Step 2. Let us define the event Ω_n that the class $c_n \in \mathcal{C}$ corresponding to the incoming $b_n \in \mathcal{B}$ is admissible, that is

$$\Omega_n = \Big\{c_n \in \mathcal{C}_{n-1}(b_n)\Big\}\,.$$

By the result of Step 1 and using a union bound over $a \in \mathcal{A}$ it holds that for all δ', then

$$\mathbb{P}\Big(\overline{\Omega}_{n+1} \cap b_{n+1} = b\Big) \leqslant |\mathcal{A}^\star| n^+_{\Upsilon(b)}(\delta')n^-_{\Upsilon(b)}(\delta')^{-3} + 2\delta'\,,$$

where we get $2\delta'$ and not $2\delta'|\mathcal{A}^\star|$ since this quantity corresponds to the same event for all $a \in \mathcal{A}^\star$, and where the restriction from to $\mathcal{A}$ to $\mathcal{A}^\star$ comes from the fact that since the algorithm never pulls not elite arms, then the corresponding range of the confidence interval is infinite and thus the corresponding probability is null.

Step 3. Now, under the event Ω_n, we known that if $b_n = b \in \mathcal{B}_c$, then $c_b \in \mathcal{C}_{n-1}(b)$ is admissible, and thus the selected class $\widehat{c}_n$ can only has larger optimal mean than c_b, that is $\mu_{\star_{\widehat{c}_n},\widehat{c}_n} \geqslant \mu_{\star_{c_b},c_b} (\geqslant \mu_{\star_{\widehat{c}_n},c_b})$.

Note also that if $c' \in \mathcal{C}_{n-1}(b_n)$, since $c_{b_n} \in \mathcal{C}_{n-1}(b_n)$ under Ω_n, then we must have for all $a \in \mathcal{A}$

$$\begin{aligned}
\mu_{a,c'} - \mu_{a,c_{b_n}} &\leqslant |\mu_{a,c'} - \widehat{\mu}_{a,b_n}(n-1)|\\
&\quad + |\widehat{\mu}_{a,b_n}(n-1) - \mu_{a,c_{b_n}}|\\
&\leqslant G_{a,b_n}(n-1)\,,
\end{aligned}$$

that is $N_{a,b_n}(n-1)$ is small for all $a \in \mathcal{A}$, and in particular for a_n. We thus have for $a = a_n$,

$$\begin{aligned}
&G_{a,b_n}(n-1) \geqslant \mu_{a_n,\widehat{c}_n} - \mu_{a_n,c_{b_n}}\\
&\quad \geqslant \inf_{c' \in \mathcal{C}}\Big\{\mu_{a,c'} - \mu_{a,c_{b_n}} : \star_{c'} = a \cap \mu_{\star_{c'},c'} \geqslant \mu_{\star_{c_{b_n}},c_{b_n}}\Big\}\,,
\end{aligned}$$

where we recognize the definition of $\Delta^+_{a,c_{b_n}}$.

Step 4. At this point, we now look at the total number of pulls $N_{a,b}(N)$ of sub-optimal arm $a \neq \star_{c_b}$ in environment b after N rounds. It satisfies for all integer $u_{a,b} \in \mathbb{N}$

$$\begin{aligned}
N_{a,b}(N) &\leqslant u_{a,b}\\
&+ \sum_{n=u_{a,b}+1}^{N}\mathbb{I}\{a_n = a \cap b_n = b \cap N_{a,b_n}(n-1) > u_{a,b}\}\\
&\leqslant u_{a,b} + \sum_{n=u_{a,b}+1}^{N}\mathbb{I}\{b_n = b \cap \overline{\Omega}_n(c_b)\}\\
&+ \sum_{n=u_{a,b}+1}^{N}\mathbb{I}\{a_n = a \cap b_n = b \cap N_{a,b_n}(n-1) > u_{a,b} \cap \Omega_n(c_b)\}\,.
\end{aligned}$$

We now consider a confidence level $\delta' = \delta'_n = 1/n^2$ at round n. Now, we know that if we choose $u_{a,b} = \Big\lceil\frac{24R^2\log(N^+_{\Upsilon(b)}(\delta_N))}{\Delta^{+\,2}_{a,c_b}}\Big\rceil$, then the events $a_n = a$ and $N_{a,b_n}(n-1) > u_{a,b}$ can not hold simultaneously on Ω_n (we used here that $n \to n^+_{\Upsilon(b)}(\delta_n)$ is increasing with n). Thus, we deduce that

$$\begin{aligned}
\mathbb{E}\Big[N_{a,b}(N)\Big] &\leqslant 1 + \frac{24R^2\log(N^+_{\Upsilon(b)}(\delta_N))}{\Delta^{+\,2}_{a,c_{b_n}}}\\
&+ \sum_{n=u_{a,b}}^{N-1}\Big(2n^+_{\Upsilon(b)}(\delta_{n+1})n^-_{\Upsilon(b)}(\delta_{n+1})^{-3}|\mathcal{A}^\star| + 2\delta_{n+1}\Big)\Upsilon(b)\,.
\end{aligned}$$

Step 5. Now choosing $\delta_n = 1/n^2$ and since $n^+_{\Upsilon(b)}(\delta)n^-_{\Upsilon(b)}(\delta_n)^{-3} = O\Big(n^{-2}\Upsilon(b)^{-2}\Big)$, then we deduce that

$$\begin{aligned}
\mathbb{E}\Big[N_{a,b}(N)\Big] &\leqslant 1 + \frac{24R^2\log(N^+_{\Upsilon(b)}(\delta_N))}{\Delta^{+\,2}_{a,c_{b_n}}}\\
&\quad + \frac{\pi^2}{3}\Big(O\big(\Upsilon(b)^{-1}\big) + \Upsilon(b)\Big)\,.
\end{aligned}$$

Finally, since on the other end we must have in all cases $\mathbb{E}\Big[N_{a,b}(N)\Big] \leqslant \mathbb{E}\Big[N_b(N)\Big] = N\Upsilon(b)$, we deduce that the regret is given by[4]

$$\begin{aligned}\mathfrak{R}_N &= \sum_{b\in\mathcal{B}}\sum_{a\in\mathcal{A}}\Delta_{a,c_b}\mathbb{E}\Big[N_{a,b}(N)\Big]\\ &\leqslant \sum_{b\in\mathcal{B}}\sum_{a\in\mathcal{A}^\star}\min\Big\{\frac{24R^2\Delta_{a,c_b}\log(N\Upsilon(b))}{\Delta^{+\,2}_{a,c_b}}\\ &\quad +O\big(\Upsilon(b)^{-1}\big),\Delta_{a,c_b}N\Upsilon(b)\Big\}.\end{aligned}$$

C. Technical details for the fully agnostic case

In this section, we prove Theorem 7 about the regret of benchmark algorithms. We then prove Proposition 1 and then a sketch of proof for Lemma 1 and Lemma 2.

C.1. Proof of Theorem 7

We analyze here the regret of `UCB` on $\mathcal{B}$. The regret bound for `UCB` on $\mathcal{C}$ is obtained following exactly the same line of proof.

Step 1. The algorithm chooses at step n, for incoming $b_n\in\mathcal{B}$ an arm a_n maximizing the upper bound

$$U_{a,b_n}(n) = \widehat{\mu}_{a,b_n}(n) + R\sqrt{\frac{6\log\big(N_{b_n}(n)\big)}{N_{a,b_n}(n)}}\,.$$

Now let us denote $b=b_n$ the current environment, and let us introduce the event

$$\Omega_a(n) = \Big\{L_{a,b_n}(n)\leqslant \mu_{a,b} \quad\cap\quad \mu_{\star_{b_n},b_n}\leqslant U_{\star_{b_n},b_n}(n)\Big\}.$$

If the algorithm chooses a sub-optimal arm, that is $a_n = a\neq \star_{b_n} = a'$, then we must have on the event $\Omega_a(n-1)$ that

$$\begin{aligned}&\mu_{a,b}+2R\sqrt{6\frac{\log(N_b(n-1))}{N_{a,b}(n-1)}}\\ &\geqslant \widehat{\mu}_{a,b}(n-1)+R\sqrt{6\frac{\log(N_b(n-1))}{N_{a,b}(n-1)}}\\ &\geqslant \widehat{\mu}_{a',b}(n-1)+R\sqrt{6\frac{\log(N_b(n-1))}{N_{a',b}(n-1)}}\\ &\geqslant \mu_{a',b}\,,\end{aligned}$$

and since we also have $\{N_b(n-1)\leqslant (n-1)^+_{\Upsilon(b_n)}(\delta)\}\subset\Omega_a(n-1)$, we deduce that we must have under this event

$$\begin{aligned}N_{a,b_n}(n-1) &\leqslant \frac{24R^2\log(N_b(n-1))}{\Big(\mu_{\star_{b_n},b_n}-\mu_{a,b_n}\Big)^2}\\ &\leqslant \frac{24R^2\log((n-1)^+_{\Upsilon(b_n)}(\delta))}{\Delta^2_{a,b_n}}\,.\end{aligned}$$

Step 2. At this point, we now look at the total number of pulls $N_{a,b}(N)$ of sub-optimal arm $a\neq\star_b$ in environment b after N rounds. It satisfies for all integer $u_{a,b}\in\mathbb{N}$

$$\begin{aligned}N_{a,b}(N) &\leqslant u_{a,b}+\sum_{n=u_{a,b}+1}^{N}\mathbb{I}\{a_n=a\cap b_n=b\\ &\qquad\cap N_{a,b_n}(n-1)>u_{a,b}\}\\ &\leqslant u_{a,b}+\sum_{n=u_{a,b}+1}^{N}\mathbb{I}\{U_{a_n,b}(n)\geqslant U_{\pi^\star(b),b}(n)\\ &\qquad\cap b_n=b\cap N_{a,b_n}(n-1)>u_{a,b}\}\end{aligned}$$

Step 3. We now consider a confidence level $\delta=\delta_n=1/n^2$ at round n. Now, we know that if we choose $u_{a,b}=\lceil\frac{6R^2\log(N^+_{\Upsilon(b)}(\delta_N))}{\Delta^2_{a,b}}\rceil$, then $a_n=a$ and $N_{a,b_n}(n-1)>u_{a,b}$ can not hold simultaneously on $\Omega_a(n)$ (we used here that $n\to n^+_{\Upsilon(b)}(\delta_n)$ is increasing with n). Thus, we deduce that

$$\begin{aligned}\mathbb{E}\Big[N_{a,b}(N)\Big] &\leqslant 1+\frac{24R^2\log(N^+_{\Upsilon(b)}(\delta_N))}{\Delta^2_{a,b}}\\ &\quad+\sum_{n=u_{a,b}+1}^{N}\Big(2n^+_{\Upsilon(b)}(\delta)n^-_{\Upsilon(b)}(\delta_n)^{-3}+2\delta_n\Big)\Upsilon(b)\,.\end{aligned}$$

Now, since $n^+_{\Upsilon(b)}(\delta)n^-_{\Upsilon(b)}(\delta_n)^{-3}=O\Big(n^{-2}\Upsilon(b)^{-2}\Big)$, we deduce that

$$\begin{aligned}\mathbb{E}\Big[N_{a,b}(N)\Big] &= 1+\frac{24R^2\log(N\Upsilon(b))}{\Delta^2_{a,b}}\\ &\quad+\frac{\pi^2}{3}\Big(O\Big(\Upsilon(b)^{-1}\Big)+\Upsilon(b)\Big)\,.\end{aligned}$$

Noticing that, on the other hand, we must have $\mathbb{E}\Big[N_{a,b}(N)\Big]\leqslant\mathbb{E}\Big[N_b(N)\Big]=N\Upsilon(b)$, we deduce that the final bound on the regret is given by

$$\begin{aligned}\mathfrak{R}_N^{\texttt{UCB on }\mathcal{B}} &= \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}\Delta_{a,b}\mathbb{E}\Big[N_{a,b}(N)\Big]\\ &\leqslant \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}\min\Big\{\frac{24R^2\log(N\Upsilon(b))}{\Delta^2_{a,b}}\\ &\quad+O\Big(\Upsilon(b)^{-1}\Big),\ \Delta_{a,b}N\Upsilon(b)\Big\}.\end{aligned}$$

[4] Note that the event $b_n=b\cap\Omega_n(c_b)^c$ is actually common to all arms $a\in\mathcal{A}^\star$, so we could in principle improve slightly on the result.

C.2. Proof of Proposition 1

We actually prove the following result that is slightly stronger than Proposition 1:

Proposition 2 *Let us consider the event* $\Omega_n = \left\{\mathcal{B}_{c_n} \in \mathfrak{B}_{b_n}(n-1)\right\}$*, where* $c_n = c$ *is the current class and* $a_n = a$ *the chosen action. Under this event, there exists a maximally compatible super-set* $\tilde{\mathcal{B}}_{c_n}$ *that contains* $\mathcal{B}_{c_n}$*. Now let us define the pseudo-gaps*

$$\tilde{\Delta}_{a,c} = \max\left\{\mu_{a,c(b)} - \mu_{a,c} : b \in \tilde{\mathcal{B}}_c \cap \max_{B\in\mathfrak{B}_{b_n}(n-1)} U_{a,B} \geqslant U_{\star_c,\tilde{\mathcal{B}}_c}\right\},$$

then the events $\Gamma_n = \left\{G_{a,\mathcal{B}_c}(n-1) > \tilde{\Delta}_{a,c}\right\}$ *and* $\mathcal{E}_n$ *the event that the confidence interval of the optimal arm of cluster* $\mathcal{B}_c$ *is small enough (or that* $\tilde{\mathcal{B}}_c$ *is not too biased),*

$$\mathcal{E}_n = \left\{\min\left\{G_{\star_c,\mathcal{B}_c}(n-1), \max_{b\in\tilde{B}} \mu_{\star_c,c} - \mu_{\star_c,c(b)}\right\} < \alpha\Delta_{a_n,c}\right\}$$

Then, under $\Omega_n \cap \Gamma_n$*, it happens that*

$$N_{a_n,\mathcal{B}_{c_n}}(n-1) < \frac{24R^2}{\tilde{\Delta}^2_{a_n,c_n}} \log(N_{\mathcal{B}_{c_n}}(n-1))\,.$$

Now, under $\Omega_n \cap \mathcal{E}_n$*, it happens that for all* $\eta \in (\alpha, 1]$*, either*

$$N_{a_n,b_n}(n-1) < \left(1+\frac{\varepsilon}{2}\right)^2 \frac{24R^2 \log(N_{b_n}(n-1))}{(\eta-\alpha)^2\Delta^2_{a_n,c_n}}\,,$$
$$\text{or } N_{a_n,\mathcal{B}_{c_n}}(n-1) < \frac{24R^2 \log(N_{\mathcal{B}_{c_n}}(n-1))}{(1-\eta)^2\Delta^2_{a_n,c_n}}\,.$$

Step 1. Let us consider the event that at round n, where $b_n \in \mathcal{B}$ has class $c = c_n \in \mathcal{C}$, then the set $\mathcal{B}_{c_n}$ is compatible with b_n and chosen action $a_n \in \mathcal{A}$ that is

$$\Omega_n = \left\{\mathcal{B}_{c_n} \in \mathfrak{B}_{b_n}(n-1)\right\}.$$

In that case, then, there is a maximally compatible super-set $\tilde{\mathcal{B}}_{c_n}$ that contains $\mathcal{B}_{c_n}$.

Step 2. Since the algorithm chooses to play some action a_n proposed by some chosen maximally compatible set $B_n \ni b_n$, it satisfies

$$U_{a_n,B_n}(n-1) \geqslant U_{a_{\tilde{\mathcal{B}}_{c_n}},\tilde{\mathcal{B}}_{c_n}}(n-1) \geqslant U_{\star_{c_n},\tilde{\mathcal{B}}_{c_n}}(n-1)\,.$$

Now on the other end, any set $B \in \mathfrak{B}_{b_n}(n-1)$ compatible with b_n must satisfy for $a = a_n$

$$\begin{aligned} U_{a,B}(n-1) - U_{a,\tilde{\mathcal{B}}_{c_n}}(n-1) &\leqslant U_{a,b_n}(n-1; 1+\varepsilon) - L_{a,b_n}(n-1) \\ &= \left(1+\frac{\varepsilon}{2}\right) G_{a,b_n}(n-1)\,. \end{aligned}$$

Thus, we deduce that under Ω_n, any model B that is chosen must satisfy

$$U_{a_B,B}(n-1) - U_{a_B,\tilde{\mathcal{B}}_{c_n}}(n-1) \leqslant \left(1+\frac{\varepsilon}{2}\right) G_{a_B,b_n}(n-1)$$
$$\text{and} \quad U_{a_B,B}(n-1) \geqslant U_{\star_{c_n},\tilde{\mathcal{B}}_{c_n}}(n-1)\,.$$

Step 3. In case $\tilde{\mathcal{B}}_{c_n} = \mathcal{B}_{c_n}$, then $U_{\star_{c_n},\tilde{\mathcal{B}}_{c_n}}(n-1) \geqslant \mu_{\star_{c_n},c_n}$. However, in general $\tilde{\mathcal{B}}_{c_n}$ may contain other points that cause the confidence interval to shrink, and creates a possible bias. Now if the corresponding bias is positive, then we still have $U_{\star_{c_n},\tilde{\mathcal{B}}_{c_n}}(n-1) \geqslant \mu_{\star_{c_n},c_n}$. Otherwise, on the one hand we have $U_{\star_{c_n},\tilde{\mathcal{B}}_{c_n}}(n-1) \geqslant \mu_{\star_{c_n},c_n} - \max_{b\in\tilde{\mathcal{B}}_{c_n}} d_{\star_{c_n},c_n,c(b)}$, and on the other hand, in the worst case since $\mathcal{B}_{c_n} \subset \tilde{\mathcal{B}}_{c_n}$, the interval is shrink on the lower-side of $S_{\star_{c_n},\mathcal{B}_{c_n}}(n-1)$ and we thus get $U_{\star_{c_n},\tilde{\mathcal{B}}_{c_n}}(n-1) \geqslant \mu_{\star_{c_n},c_n} - G_{\star_{c_n},\mathcal{B}_{c_n}}(n-1)$. Thus, we have in all cases

$$U_{\star_{c_n},\tilde{\mathcal{B}}_{c_n}}(n-1) \geqslant \mu_{\star_{c_n},c_n} - \underline{G}_{\star_{c_n},\mathcal{B}_{c_n},\tilde{\mathcal{B}}_{c_n}}\,,$$

where we introduced for convenience the notation

$$\underline{G}_{\star_{c_n},\mathcal{B}_{c_n},\tilde{\mathcal{B}}_{c_n}} = \min\left\{G_{\star_{c_n},\mathcal{B}_{c_n}}(n-1), \max_{b\in\tilde{\mathcal{B}}_{c_n}} d_{\star_{c_n},c_n,c(b)}\right\}.$$

Whenever $\underline{G}_{\star_{c_n},\mathcal{B}_{c_n},\tilde{\mathcal{B}}_{c_n}}$ is small enough, this bias is controlled and we are essentially done. Since for the same reason we have

$$U_{a_B,\tilde{\mathcal{B}}_{c_n}}(n-1) \leqslant \mu_{a_B,c_n} + G_{a_B,\mathcal{B}_{c_n}}(n-1)\,,$$

where we introduced this time

$$\overline{G}_{a_B,\mathcal{B}_{c_n},\tilde{\mathcal{B}}_{c_n}} = \min\left\{G_{a_B,\mathcal{B}_{c_n}}(n-1), \max_{b\in\tilde{\mathcal{B}}_{c_n}} d_{a,c(b),c_n}\right\},$$

we deduce that under Ω_n, we must have for the chosen $a_n = a_B$

$$\begin{aligned} \Delta_{a_B,c_n} \leqslant\ & \underline{G}_{\star_{c_n},\mathcal{B}_{c_n},\tilde{\mathcal{B}}_{c_n}} + \overline{G}_{a_B,\mathcal{B}_{c_n},\tilde{\mathcal{B}}_{c_n}} \\ & + \left(1+\frac{\varepsilon}{2}\right) G_{a_B,b_n}(n-1)\,. \end{aligned}$$

Step 4. In the case when $\tilde{\mathcal{B}}_{c_n}$ is negatively biased, if a is chosen and $c_n = c$, then a must belong with high probability to the set

$$\tilde{\mathcal{A}}_c(n-1) = \Big\{a \in \mathcal{A} : \exists B, a = a_B, \forall a' \in \mathcal{A} : U_{a,B}(n-1) \geqslant L_{a',\mathcal{B}_c}(n-1) - \underline{G}_{\star_c,\mathcal{B}_c,\tilde{\mathcal{B}}_c} \text{ and } U_{a,\tilde{\mathcal{B}}_c}(n-1) \leqslant U_{a',\mathcal{B}_c}(n-1) + \overline{G}_{a,\mathcal{B}_c,\tilde{\mathcal{B}}_c} + \Big(1+\frac{\varepsilon}{2}\Big) G_{a_B,b_n}(n-1)\Big\}.$$

Indeed for instance otherwise for a' such that it doesn't hold, we get with high probability that

$$\begin{aligned} U_{a_B,B}(n-1) &< L_{a',\mathcal{B}_c}(n-1) - \underline{G}_{\star_c,\mathcal{B}_c,\tilde{\mathcal{B}}_c} \\ &\leqslant \mu_{a',c} - \tilde{G}_{\star_c,\mathcal{B}_c,\tilde{\mathcal{B}}_c}, \end{aligned}$$

which contradicts the fact that $U_{a_B,B}(n-1) \geqslant U_{\star_c,\tilde{\mathcal{B}}_c}(n-1) \geqslant \mu_{\star_c,c} - \tilde{G}_{\star_c,\mathcal{B}_c,\tilde{\mathcal{B}}_c}$.

Step 5. At this point, two situations may appear. In the first case, we have for the chosen arm a

$$\begin{aligned} G_{a,\mathcal{B}_{c_n}}(n-1) &> \max_{b \in \tilde{\mathcal{B}}_{c_n}} d_{a,c(b),c_n} \\ &= \max\Big\{\mu_{a,c(b)} - \mu_{a,c_n} : b \in \tilde{\mathcal{B}}_{c_n} \\ &\qquad \cap \max_{B \in \mathfrak{B}_{b_n}} U_{a,B} \geqslant U_{\star_{c_n},\tilde{B}_{c_n}}\Big\}, \end{aligned}$$

and thus its total number of visits $N_{a,\mathcal{B}_{c_n}}(n-1)$ must be small. We naturally the quantity on the right hand side of previous inequality $\tilde{\Delta}_{a,c_n}$ and introduce the corresponding event $\Gamma_n = \Big\{G_{a_n,\mathcal{B}_{c_n}}(n-1) > \tilde{\Delta}_{a_n,c_n}\Big\}$. In the second case, it holds anyway that $\overline{G}_{a_n,\mathcal{B}_{c_n},\tilde{\mathcal{B}}_{c_n}} \leqslant G_{a_n,\mathcal{B}_{c_n}}(n-1)$, thus under the event $\Omega_n \cap \mathcal{E}_n$, where $\mathcal{E}_n = \Big\{\underline{G}_{\star_{c_n},\mathcal{B}_{c_n},\tilde{\mathcal{B}}_{c_n}} \leqslant \alpha\Delta_{a_n,c_n}\Big\}$ we deduce that

$$(1-\alpha)\Delta_{a_n,c_n} \leqslant G_{a_n,\mathcal{B}_{c_n}}(n-1) + \Big(1+\frac{\varepsilon}{2}\Big)G_{a_n,b_n}(n-1).$$

In such a case, for any $\eta > \alpha$, two situations may occur. First, if $\Big(1+\frac{\varepsilon}{2}\Big)G_{a_n,b_n}(n-1) < (\eta-\alpha)\Delta_{a_n,c_n}$, then $(1-\eta)\Delta_{a_n,c} \leqslant G_{a_n,\mathcal{B}_c}(n-1)$, that is $N_{a,\mathcal{B}_c}(n-1)$ must be small as well. Otherwise, $\Big(1+\frac{\varepsilon}{2}\Big)G_{a_n,b_n}(n-1) \geqslant (\eta-\alpha)\Delta_{a_n,c_n}$ and thus $N_{a_n,b_n}(n-1)$ must be small.

Step 6. All in all, under $\Omega_n \cap \Gamma_n$, we obtain that

$$N_{a_n,\mathcal{B}_c}(n-1) < \frac{24R^2}{\tilde{\Delta}^2_{a_n,c}} \log(N_{\mathcal{B}_c}(n-1)).$$

Now under $\Omega_n \cap \mathcal{E}_n$ it holds that for all $\eta > \alpha$ either

$$N_{a_n,b_n}(n-1) < \Big(1+\frac{\varepsilon}{2}\Big)^2 \frac{24R^2}{(\eta-\alpha)^2\Delta^2_{a_n,c}} \log(N_{b_n}(n-1)),$$

or

$$N_{a_n,\mathcal{B}_c}(n-1) < \frac{24R^2}{(1-\eta)^2\Delta^2_{a_n,c}} \log(N_{\mathcal{B}_c}(n-1)).$$

C.3. Sketch of proof of Lemma 1

In this proof our goal is to show that if $b_n \in \mathcal{B}_c$, then $\mathfrak{B}_{b_n}(n-1)$ contains at least $\mathcal{B}_c(b_n;\gamma)$ for the chosen γ. Thus for big enough γ, all the set $\mathcal{B}_c$ will be admissible, and thus we get a maximal speed-up. We present now a sketch of proof relying on some event (see (6)). We provide intuition in Step 5 that this intuitive event indeed holds, even if a precise proof is more technical than it seems. Thus for clarity and to avoid a not illuminating tedious analysis the proof of this last step is not reproduced here.

Step 1. Proceeding similarly to the analysis in Step 1 of the proof of Theorem 6, we note that the event $\forall b \in \mathcal{B} \forall a \in \mathcal{A}\, \mu_{a,b} \in S_{a,b}(n-1)$ happens with probability at least $1 - \Big(\sum_{b\in\mathcal{B}} n^+_{\Upsilon(b)}(\delta')n^-_{\Upsilon(b)}(\delta')^{-3}A + 2\delta' B\Big)$ for all δ'. Now under this event, we note that if $b' \in \mathcal{B}_c$ is such that $S_{a,b'}(n-1) \subset S_{a,b_n}(n-1;1+\varepsilon)$ for all $a \in \mathcal{A}$, since it is also such that $S_{a,b'}(n-1) \cap S_{a,b_n}(n-1) \neq \emptyset$ and satisfies other conditions for being admissible, we deduce that the set of all such b' is admissible, that is we have the inclusion

$$\Big\{b' \in \mathcal{B}_c : \forall a \in \mathcal{A}\, G_{a,b'}(n-1) \leqslant (1+\varepsilon)G_{a,b_n}(n-1)\Big\} \in \mathfrak{B}_{b_n}(n-1),$$

Step 2. At this point, one can consider an adaptive value for ε. The value $\varepsilon = \varepsilon_{a,b_n,b',n-1}$ used in the algorithm is given by

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

Now upon noting that when it happens that for all $b, b' \in \mathcal{B}_c$, for all $a \in \mathcal{A}$ $N_{a,b}(n-1)/(N_{a,b'}(n-1) \vee 1) \leqslant 2\gamma$ by the definition of the confidence gaps it holds that $\sqrt{\frac{2\gamma \log N_{b'}(n-1)}{\log N_{b_n}(n-1)}} \geqslant \frac{G_{a,b'}(n-1)}{G_{a,b_n}(n-1)}$, we deduce in such a case that we have the following inclusion

$$\mathcal{B}_c \in \mathfrak{B}_{b_n}(n-1).$$

Step 3. Now note also that on the event $\forall b \in \mathcal{B} \forall a \in \mathcal{A}\, \mu_{a,b} \in S_{a,b}(n-1)$, by the confidence bounds on $N_{b'}(n-1)$ and $N_{b_n}(n-1)$, it holds that $N_b(n-1)/N_{b'}(n-1) = \Upsilon(b)/\Upsilon(b') + O((n-1)^{-1/2})$, and thus whenever $\gamma \geqslant \gamma_c + O((n-1)^{-1/2})$, it is enough to show that

$$\frac{N_{a,b}(n-1)}{N_{a,b'}(n-1) \vee 1} \leqslant 2\frac{N_b(n-1)}{N_{b'}(n-1)}, \tag{6}$$

in order to get the desired inclusion. Note that for $\gamma < \gamma_c + O((n-1)^{-1/2})$, we would get only the inclusion $\mathcal{B}_c(b_n;\gamma) \in \mathfrak{B}_{b_n}(n-1)$.

Step 4. Now choosing $\delta' = 1/n^2$ we get that $n^+_{\Upsilon(b)}(\delta')n^-_{\Upsilon(b)}(\delta')^{-3} = O\Big(n^{-2}\Upsilon(b)^{-2}\Big)$ and thus for $\gamma \geqslant \gamma_c + O((n-1)^{-1/2})$, upon controlling the ratio (6) we get

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

Step 5. It now remains to show that for $b, b' \in \mathcal{B}_c$, for all $a \in \mathcal{A}$, then with high probability

$$\frac{N_{a,b'}(n-1)}{N_{a,b}(n-1) \vee 1} \leqslant 2\frac{N_b(n-1)}{N_{b'}(n-1)}.$$

Up to the distortion due the different number of visits of b and b', this is essentially saying that the algorithm behaves in the same way for two b and b' belonging to the same class. Although this is quite intuitive in case the algorithm behaves completely independently on b and b' (like `UCB` on $\mathcal{B}$), this is much less clear when decisions are coupled. However, since the b_n are sampled independently from Υ and not in an adversarial way, and that b and b' are statistically not distinguishable by the algorithm, it is actually possible to show that the ratio between the number of pulls of the same arm is indeed bounded by some small constant.

C.4. Sketch of proof of Lemma 2

The detailed proof is tedious and involved with many intricate sub-cases. For this conference version of the paper, we focus on indicating the tricky parts and how to solve them, which we think to be more interesting, but do not provide all the detailed of the proof.

The actual results hold for the slightly more elaborate event E_n defined in Proposition 2.

Step 1. Let us consider some $c' \neq c$. By definition of the compatible sets used in `A-UCB` for some $B' \subset \mathcal{B}_{c'}$, if it exists some action $a \in \mathcal{A}$ such that

$$G_{a,B'} + G_{a,\mathcal{B}_c} < d_{a,c,c'},$$

then B' is not compatible with $\mathcal{B}_c$. Now in the case when $d_{a,c,c'} < 2G_{a,\mathcal{B}_c}$ it means that $N_{a,\mathcal{B}_c}$ is small, and that every B' such that $G_{a,B'} \geqslant G_{a,\mathcal{B}_c}$ is compatible. In the other cases, B' must satisfy at least $G_{a,B'} \geqslant d_{a,c,c'}/2$ to be compatible with $\mathcal{B}_c$. Thus we deduce that for each $c' \neq c$, the sets $B' \subset \mathcal{B}_{c'}$ compatible with $\mathcal{B}_c$ must satisfy that

$$\forall a \in \mathcal{A}\Big(G_{a,B'} \geqslant d_{a,c,c'}/2 \quad \text{or} \quad d_{a,c,c'} < 2G_{a,\mathcal{B}_c}\Big).$$

We use this observation in order to control the bias of $\tilde{\mathcal{B}}_{c_n}$. In the sequel, we work under the event Ω_n and focus on the situation when $\tilde{\mathcal{B}}_{c_n} = \mathcal{B}_{c_n} \cup B'$, with $B' \subset \mathcal{B}_{c'}$. The general situation when B' can be spread in different classes c', can be handled similarly. Now three important situations may occur that we consider in the following three steps.

Step 2.1 First, for a compatible $B' \subset \mathcal{B}_{c'}$, if there is no $a \in \mathcal{A}$ such that $d_{a,c,c'} < 2G_{a,\mathcal{B}_c}$, then it means that $\forall a \in \mathcal{A}, G_{a,B'} \geqslant d_{a,c,c'}/2 \geqslant G_{a,\mathcal{B}_c}$. In such case, we deduce that B' induces a very small bias on these arms. Indeed

$$\begin{aligned} N_{a,B'} &\leqslant \frac{32R^2\log(N^3_{B'})}{d^2_{a,c,c'}} \\ &\leqslant \frac{\log(N_{B'})}{\log(N_{\mathcal{B}_c})}N_{a,\mathcal{B}_c}, \end{aligned}$$

with $N_{B'} \simeq N_{\mathcal{B}_c}$ due to the uniform assumption on Γ. Thus the maximal bias that B' can create on the mean $\mu_{a,c}$ of $\mathcal{B}_c$ for arm a, that is $\mu_{a,\tilde{\mathcal{B}}_{c_n}} - \mu_{a,c}$ is essentially less than $d_{a,c,c'}/2 \leqslant \Big(G_{a,B'} + G_{a,\mathcal{B}_c}\Big)/2$ (note that only a positive bias can be bad, a negative bias on sub-optimal arms is always helpful), which happens when $N_{a,B'}$ is large and thus $G_{a,B'}$ is small. Now if $N_{a,B'}$ is much smaller than $N_{a,\mathcal{B}_c}$, the bias is much smaller than $d_{a,c,c'}/2$ as well. Elaborating on these ideas, we can actually show that the bias is controlled by $\beta G_{a,\mathcal{B}_c}$ for some constant β.

Step 2.2 The second situation is when there is some $a \neq \star_c$ with $d_{a,c,c'} > 0$ such that $d_{a,c,c'} < 2G_{a,\mathcal{B}_c}$. In that case we immediately deduce that

$$N_{a,\mathcal{B}_c} \leqslant \frac{32R^2\log(N^3_{\mathcal{B}_c})}{d^2_{a,c,c'}}.$$

On the other hand, if all $a \neq \star_c$ such that $d_{a,c,c'} < 2G_{a,\mathcal{B}_c}$ satisfy $d_{a,c,c'} = 0$, no bias is introduced on these arms. Thus either there is no bias or $N_{a,\mathcal{B}_c}$ is small.

Step 2.3 Finally, it remains to handle the case when $d_{\star_c,c,c'} < 2G_{\star_c,\mathcal{B}_c}$. Let us consider all the $a' \neq \star_c$ such that $d_{a',c,c'} \geqslant 2G_{a',\mathcal{B}_c}$ (it may be an empty set, in which case it means that either $d_{a',c,c'} = 0$ or $N_{a',\mathcal{B}_c}$

is small for all $a' \neq \star_c$). By the result of Step 2.1, the bias on $\mu_{a,c}$ is controlled by $G_{a,\mathcal{B}_c}$ for all these arms. We want to show that the bias on $\mu_{\star_c,c}$ must be also small. Note that if $d_{\star_c,c,c'} < 1/2\Delta_{a,c}$ for all a (following the assumption of the lemma), then $\mathcal{E}_n$ happens and thus the regret is controlled. Similarly $\mathcal{E}_n$ happens if $G_{\star_c,\mathcal{B}_c} < \alpha d_{\star_c,c,c'}$. We thus focus on the case when $d_{\star_c,c,c'} \geqslant 2(1+\eta)\Delta_{a,c} > 2\Delta_{a,c}$ for some $a \neq \star_c$, and $G_{\star_c,\mathcal{B}_c} \geqslant \alpha d_{\star_c,c,c'}$. We need to consider more sub-cases, depending on the size of $G_{\star_c,B'}$.

2.3.1 If $G_{\star_c,B'} = \kappa G_{\star_c,\mathcal{B}_c}$, with $\kappa > 1$, then by definition of $G_{\star_c,B'}$, we deduce that

$$N_{\star_c,B'} \leqslant \frac{8R^2\log(N_{B'}^3)}{\kappa^2 G^2_{\star_c,\mathcal{B}_c}} = N_{\star_c,\mathcal{B}_c}\frac{\log(N_{B'})}{\kappa^2\log(N_{\mathcal{B}_c})}\,,$$

thus if we introduce for convenience the notation $\alpha = \frac{\log(N_{B'})}{\kappa^2\log(N_{\mathcal{B}_c})}$, which is close to κ^{-2} if $N_{B'} \simeq N_{\mathcal{B}_c}$ then it holds that

$$\begin{aligned}
&\widehat{\mu}_{\star_c,\mathcal{B}_c\cup B'}\\
&= \widehat{\mu}_{\star_c,\mathcal{B}_c}\frac{1}{1+\alpha} + \widehat{\mu}_{\star_c,B'}\frac{\alpha}{1+\alpha}\\
&\geqslant \Big(\mu_{\star_c,c} - \frac{1}{2}G_{\star_c,\mathcal{B}_c}\Big)\frac{1}{1+\alpha} + \Big(\mu_{\star_c,c'} - \frac{1}{2}G_{\star_c,B'}\Big)\frac{\alpha}{1+\alpha}\\
&= \mu_{\star_c,c} - d_{\star_c,c,c'}\frac{\alpha}{1+\alpha} - \frac{1}{2(1+\alpha)}\Big(G_{\star_c,\mathcal{B}_c} + \alpha G_{\star_c,B'}\Big)\\
&= \mu_{\star_c,c} - d_{\star_c,c,c'}\frac{\alpha}{1+\alpha} - \frac{1+\kappa\alpha}{2(1+\alpha)}G_{\star_c,\mathcal{B}_c}\,.
\end{aligned}$$

Now on the other hand,

$$G_{\star_c,\mathcal{B}_c\cup B'} \geqslant \sqrt{\frac{\log(N_{B'}+N_{\mathcal{B}_c})}{\frac{\log(N_{B'})}{G^2_{\star_c,B'}} + \frac{\log(N_{\mathcal{B}_c})}{G^2_{\star_c,\mathcal{B}_c}}}} \geqslant \frac{1}{\sqrt{\alpha+1}}G_{\star_c,\mathcal{B}_c}\,.$$

Thus, we deduce that the upper bound on $\star_c$ must be high, which indicates that $\star_c$ must be the chosen arm

$$\begin{aligned}
U_{\star_c,\mathcal{B}_c\cup B'} \geqslant\ & \mu_{\star_c,c} - d_{\star_c,c,c'}\frac{\alpha}{1+\alpha}\\
&+\Big(\frac{1}{\sqrt{\alpha+1}} - \frac{1+\kappa\alpha}{2(1+\alpha)}\Big)G_{\star_c,\mathcal{B}_c}\,.
\end{aligned}$$

Indeed, using the approximation $\alpha \simeq \kappa^{-2}$, we get

$$\begin{aligned}
U_{\star_c,\mathcal{B}_c\cup B'} \gtrsim\ & \mu_{\star_c,c} - d_{\star_c,c,c'}\frac{1}{1+\kappa^2}\\
&+\frac{\kappa}{1+\kappa^2}\Big(\sqrt{1+\kappa^2} - \frac{1+\kappa}{2}\Big)G_{\star_c,\mathcal{B}_c}
\end{aligned} \tag{7}$$

The term in the second line is always positive, and if $\kappa > 1$, then we already get

$$U_{\star_c,\mathcal{B}_c\cup B'} \geqslant \mu_{\star_c,c} - \frac{1}{2}d_{\star_c,c,c'} + \frac{1}{2\sqrt{2}}G_{\star_c,\mathcal{B}_c}\,.$$

A maximal regret is incurred when the chosen arm is such that $\mu_{a,c}$ is minimum, under constraint that $d_{\star_c,c,c'} > 3\Delta_{a,c}/2$, that is when $d_{\star_c,c,c'} = 3\Delta_{a,c}/2$. Also by hypothesis $G_{\star_c,\mathcal{B}_c} \geqslant \alpha d_{\star_c,c,c'} > 3\alpha\Delta_{a,c}/2$ (and can be much larger), thus in this situation we still have $U_{\star_c,\mathcal{B}_c\cup B'} \geqslant \mu_{a,c} + (1+\frac{3\alpha}{\sqrt{2}})\frac{\Delta_{a,c}}{4}$, which is enough to discard playing arm a, since by previous analysis the bias on arm a is controlled by $\beta G_{a,\mathcal{B}_c}$. Now it is possible to extend this intuition to show that indeed $\star_c$ must be chosen with high probability if both κ and $G_{\star_c,\mathcal{B}_c}$ are big.

2.3.2 Now, we turn to the case when $G_{\star_c,B'} = \kappa G_{\star_c,\mathcal{B}_c}$ with $\kappa < 1$ (still with the assumption $G_{\star_c,\mathcal{B}_c} \geqslant \alpha d_{\star_c,c,c'}$). Using the bound (7) and the fact that $G_{\star_c,\mathcal{B}_c} \geqslant d_{\star_c,c,c'}$, we first note that it is enough to show that

$$\frac{\kappa\alpha}{1+\kappa^2}\Big(\sqrt{1+\kappa^2} - \frac{1+\kappa}{2} - \frac{1}{\kappa\alpha}\Big)d_{\star_c,c,c'} > -\Delta_{a,c}\,, \tag{8}$$

to ensure that the bias is controlled (this is actually a slight refinement of $\mathcal{E}_n$), which is essentially saying that $\kappa > 0.6$ (or more precisely $\sqrt{\frac{\log(N_{\mathcal{B}_c})}{\log(N_{B'})}}\kappa > 0.6$) is enough in the scenario when $d_{\star_c,c,c'} = 3\Delta_{a,c}/2$ and $\alpha = 1/2$.

Now, let us show that indeed κ cannot be too small. First, note that we have $\frac{\log(N_{B'})}{G^2_{\star_c,B'}} = \sum_{b'\in B'}\frac{\log(N_{b'})}{G^2_{\star_c,b'}}$, and under the assumption that Υ is uniform and all classes have equal size, then we get

$$\begin{aligned}
&\sum_{b'\in B'}\frac{\log(N_{b'})}{G^2_{\star_c,b'}}\\
&\leqslant \frac{1}{|\mathcal{B}_c|}\sum_{b\in\mathcal{B}_c}\Big(\sum_{b'\in B':G_{\star_c,b'}>G_{\star_c,b}}\frac{\log(N_{b'})}{G^2_{\star_c,b}}\\
&\qquad + \sum_{b'\in B':G_{\star_c,b'}\leqslant G_{\star_c,b}}\frac{\log(N_b)}{G^2_{\star_c,b}}2\gamma\Big)\\
&\lesssim \frac{1}{|\mathcal{B}_c|}\sum_{b\in\mathcal{B}_c}\Big(\frac{\log(N_b)}{G^2_{\star_c,b}}|B'|2\gamma\Big)\\
&\leqslant \sum_{b\in\mathcal{B}_c}\frac{2\gamma\log(N_b)}{G^2_{\star_c,b}}\,,
\end{aligned}$$

where we used the definition of $\varepsilon_{b',b}$ in the second line, and the fact that $N'_b \simeq N_b$ and that $|B'| \leqslant |\mathcal{B}_{c'}|$ in the last two lines. Thus, we deduce that we must have

$$G_{\star_c,B'} \gtrsim \sqrt{\frac{\log(N_{B'})}{\log(N_{\mathcal{B}_c})}}\frac{G_{\star_c,\mathcal{B}_c}}{\sqrt{2\gamma}}\,. \tag{9}$$

Now it is easy to check that for $\gamma \simeq \gamma_c$, since $\gamma_c = 1$ by assumption, then indeed condition (8) holds, which concludes this sketch of proof. It is then possible to refine and make each of the steps and intuitions of this proof precise, though at the price of a more tedious and cautious analysis that is not reproduced here.