Categorized Bandits

Matthieu Jedor
CMLA, ENS Paris-Saclay & Cdiscount
jedor@cmla.ens-cachan.fr

Jonathan Louëdec
Cdiscount
jonathan.louedec@cdiscount.com

Vianney Perchet
CMLA, ENS Paris-Saclay & Criteo AI Lab
perchet@cmla.ens-cachan.fr

Abstract

We introduce a new stochastic multi-armed bandit setting where arms are grouped inside "ordered" categories. The motivating example comes from e-commerce, where a customer typically has a greater appetence for items of a specific well-identified but unknown category than any other one. We introduce three concepts of ordering between categories, inspired by stochastic dominance between random variables, which are gradually weaker so that more and more bandit scenarios satisfy at least one of them. We first prove instance-dependent lower bounds on the cumulative regret for each of these models, indicating how the complexity of the bandit problems increases with the generality of the ordering concept considered. We also provide algorithms that fully leverage the structure of the model with their associated theoretical guarantees. Finally, we have conducted an analysis on real data to highlight that those ordered categories actually exist in practice.

1 Introduction

In the multi-armed bandit problem, an agent has several possible decisions, usually referred to as "arms", and chooses or "pulls" sequentially one of them at each time step. This generates a sequence of rewards and the objective is to maximize their cumulative sum. The performance of a learning algorithm is then evaluated through the "regret", which is the difference between the cumulative reward of an oracle (that knows the best arm in expectation) and the cumulative reward of the algorithm. There is a clear trade-off arising between gathering information on uncertain arms (by pulling them more often) and using this information (by choosing greedily the best decision so far). This tradeoff is usually called "exploration vs exploitation". Although originally introduced for adaptive clinical trials [37], multi-armed bandits now play an important role in recommender systems [30]. However, the traditional bandit model (see Bubeck and Cesa-Bianchi [6] for more details and variants) must be adapted to specific applications to unleash its full power.

Consider for instance e-commerce. One of the core optimization problem is to decide which products to recommend, or display, to a user landing on a website, in the objective of maximizing the click-through-rate or the conversion rate. Arms of recommender systems are the different products that can be displayed. The number of products, even if finite, is prohibitively huge as the regret, i.e. the learning cost, typically scale linearly with the number of arms. So agnostic bandit algorithms take too much time to complete their learning phase. Thankfully, there is an inherent structure behind a typical catalogue: products are gathered into well defined categories. As customers are generally interested in only one or a few of them, it seems possible and profitable to gather information across products to speed up the learning phase and, ultimately, to make more refined recommendations.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Our results We introduce and study the idea of *categorized bandits*. In this framework, arms are grouped inside known categories and we assume the existence of a partial yet unknown order between categories. We aim at leveraging this additional assumption to reduce the linear dependency in the total number of arms. We present three different partial orders over categories inspired by different notions of stochastic dominance between random variables. We considered gradually weaker notions of ordering in order to cover more and more bandit scenarios. On the other hand, the stronger the assumption, the more "powerful" the algorithms are, i.e. their regret is smaller. Those assumptions are motivated and justified by real data gathered on the e-commerce website Cdiscount. We first prove asymptotic instance-dependent lower bounds on the cumulative regret for each of these models, with a special emphasis on how the complexity of the bandit problems increases with the generality of the ordering concept considered. We then proceed to develop two generic algorithms for the categorized bandit problem that fully leverage the structure of the model; the first one is devised from the principle of optimism in the face of uncertainty [3] when the second one is from the Bayesian principle [37]. Finite-time instance-dependent upper bounds on the cumulative regret are provided for the former algorithm. Finally, we conduct numerical experiments on different scenarios to illustrate both finite-time and asymptotic performances of our algorithms compared to algorithms either agnostic to the structure or only taking it partly into account.

Related works The idea of clustering is not novel in the bandit literature [34, 5, 15, 24, 31] yet they mainly focus on clustering users based on their preferences. Li et al. [32] extended these work to the clustering of items as well. Katariya et al. [19] considered a problem where the goal is to sort items according to their means into clusters. Similar in spirit are bandit algorithms for low-rank matrix completion [39, 21, 18]. Maillard and Mannor [33] studied a multi-armed bandit problem where arms are partitioned into latent groups. Valko et al. [38] and Kocák et al. [22] proposed algorithms where the features of items are derived from a known similarity graph over the items. However, none of these works consider the known structure of categories in which the items are gathered. The model fits in the more general structured stochastic bandit framework i.e. where expected reward of arms can be dependent, see e.g., [28, 13, 2, 25, 35]. More recently, Combes et al. [8] proposed an asymptotically optimal algorithm for structured bandits relying on forced exploration (similarly to [29]) and a tracking mechanism on the number of draws of sub-optimal arms. However, these approaches forcing exploration are too conservative as the linear dependency only disappears asymptotically. There exist two other ways to tackle the bandit problem with arms grouped inside categories. The first one could rely on tree search methods, popularized by the celebrated UCT algorithm [23]. Alternative hierarchical algorithms [9] could also be used. The second one could be linear bandits [10, 36, 1] where we introduce a "categorical" feature that indicates in which category the arm belongs. However, these approaches are also not satisfactory as they do not leverage the full structure of the problem.

2 Model

We now present the variant of the multi-armed bandit model we consider. As usual, a decision maker sequentially selects (or pulls) an arm at each time step $t \in \{1, \dots, T\} =: [T]$. As motivated in the introduction, the total number of possible arms can be prohibitively large, but we assume that this large number of arms are grouped in a small number M of categories. For the sake of presentation, we are going to assume that each category has the same number of arms K, yet all of our assumptions and results immediately generalize to different number of arms. We emphasize again that the M categories of K arms each form a known partition of the set of arms (of cardinality MK). At time step $t \in [T]$, the agent selects a category C_t and an arm $A_t \in C_t$ in this category. This generates a reward $X_{A_t}^{C_t} = \mu_{A_t}^{C_t} + \eta_t$ where η_t is some independent 1 sub-Gaussian white noise and μ_k^m is the unknown expected reward of the arm k of category m. For notational convenience, we will assume that arms are ordered inside each category, i.e. $\mu_1^m > \mu_2^m \geq \cdots \geq \mu_{K-1}^m > \mu_K^m$ for all category m and that category 1 is the best category, with respect to a partial order defined below.[1] We stress out that, in the partial orders we consider, the maximum of μ_k^m over m and k is necessarily μ_1^1. As in any multi-armed bandit problem, the overall objective of an agent is to maximize her expected cumulative reward until time horizon T or identically, to minimize her expected cumulative

[1] To be precise, since the order is only partial, some categories might not be pairwise comparable, but we assume that the optimal category is comparable to, and dominates, all the others.

regret $\mathbb{E}[R_T] = T\mu_1^1 - \mathbb{E}[\sum_{t=1}^T \mu_{A_t}^{C_t}]$, or equivalently, $\mathbb{E}[R_T] = \sum_{m,k} \Delta_{m,k}\, \mathbb{E}[N_k^m(T)]$, where $\Delta_{m,k} \coloneqq \mu_1^1 - \mu_k^m$ is the difference, usually called "gap", between the expected rewards of the best arm and the k^{th} arm of category m and $N_k^m(t) \coloneqq \sum_{s<t} \mathbf{1}\{C_s = m, A_s = k\}$ denotes the number of times this arm has been pulled up to (not including) time step t.

Relations of dominance The main assumption to leverage is that the set of categories is partially ordered with a unique maximal element. Those partial orders are quite similar to the standard ones induced by stochastic dominance [17, 4] over random variables. We are going to consider three notions of dominance (inducing three different partial orders) that are gradually weaker so that the bandit setting is more and more general. Consequently, the regret should be higher and higher.

Definition 1. Let $\mathcal{A} = \{\mu_1^{\mathcal{A}}, \dots, \mu_K^{\mathcal{A}}\} \subset \mathbb{R}$ and $\mathcal{B} = \{\mu_1^{\mathcal{B}}, \dots, \mu_K^{\mathcal{B}}\} \subset \mathbb{R}$ be a pair of categories,

Group-sparse dominance $\mathcal{A}$ group-sparsely dominates $\mathcal{B}$, denoted by $\mathcal{A} \succeq_s \mathcal{B}$, if each element of $\mathcal{A}$ are non-negative and at least one is positive, and each element of $\mathcal{B}$ are non-positive, i.e., $\max_{k\in[K]} \mu_k^{\mathcal{A}} > \min_{k\in[K]} \mu_k^{\mathcal{A}} \geq 0 \geq \max_{k\in[K]} \mu_k^{\mathcal{B}}$.

Strong dominance $\mathcal{A}$ strongly dominates $\mathcal{B}$, denoted by $\mathcal{A} \succeq_0 \mathcal{B}$, if each element of $\mathcal{A}$ is bigger than any element of $\mathcal{B}$, i.e., $\min_{k\in[K]} \mu_k^{\mathcal{A}} \geq \max_{k\in[K]} \mu_k^{\mathcal{B}}$.

First-order dominance $\mathcal{A}$ first-order dominates $\mathcal{B}$, denoted by $\mathcal{A} \succeq_1 \mathcal{B}$, if $\sup_{x\in\mathbb{R}} F_{\mathcal{A}}(x) - F_{\mathcal{B}}(x) \leq 0$, where $F_{\mathcal{A}}(x) = \frac{1}{K}\sum_{k=1}^K \mathbf{1}\{\mu_k^{\mathcal{A}} \leq x\}$ is the cumulative distribution function of a uniform random variable over $\mathcal{A}$ (and similarly for $\mathcal{B}$).

The first notion of dominance is inspired by the classical (group-)sparsity concept in machine learning, that already emerged in variants of multi-armed bandits [26, 7]. It is quite a strong assumption as it implies the knowledge of a threshold[2] between two categories. The second notion weakens this assumption as the threshold is unknown. The third notion is even weaker. The second and third notions of dominance are similar to the zeroth (also called strong) and first-order of stochastic dominances between two random variables respectively uniform over $\mathcal{A}$ and $\mathcal{B}$. Hence, the three concepts of dominance immediately generalize to categories with different number of elements, with the very same definitions. Furthermore, one can weaken even more the dominance, e.g. introducing a second-order variant, but we will not consider it in this paper.

Example To illustrate the concepts of dominance, we have represented, in Figure 1, 3 categories of 3 arms each. It can be easily checked that, for the first-order dominance, *CAT 1* $\succeq_1$ *CAT 2* $\succeq_1$ *CAT 3* as, if they have the same number of elements, $\mathcal{A}$ first-order dominates $\mathcal{B}$ if the k^{th} largest elements of $\mathcal{A}$ is greater than the k^{th} largest element of $\mathcal{B}$, for any k. Moreover, for the strong dominance, *CAT 1* $\succeq_0$ *CAT 3* since the worst mean of *CAT 1* is higher than the best mean of *CAT 3*. Moreover, if this common value was known, then the dominance would even be group-sparse.

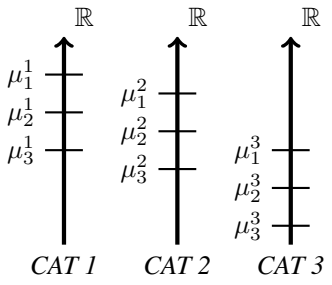


Figure 1: Illustration of dominances

Lemma 1. *Let $\mathcal{A}_1, \dots, \mathcal{A}_M$ be finite categories. If there is a category $\mathcal{A}^*$ that dominates all the other ones for any of the partial orders defined above, then $\mathcal{A}^*$ contains the maximal element of the union $\mathcal{A}_1 \cup \mathcal{A}_2 \cup \ldots \cup \mathcal{A}_M$. Moreover, if $\mathcal{A}$ group-sparsely dominates $\mathcal{B}$, then the dominance also holds in the strong sense. Similarly, if $\mathcal{A}$ strongly dominates $\mathcal{B}$, then the dominance also holds in the first-order sense.*

2.1 Empirical evidence of dominance

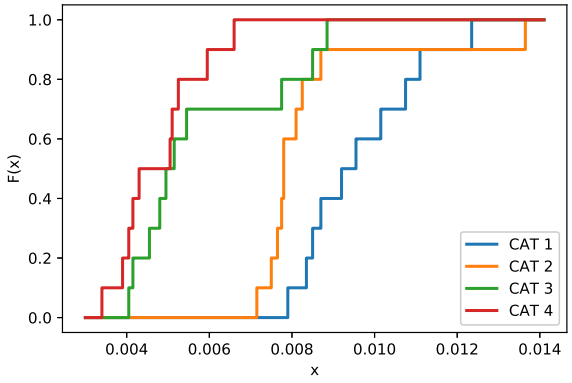
We illustrate these assumptions on a real dataset. We have collected the CTR of products in four different categories over one month on the e-commerce website Cdiscount, one of the leading e-commerce companies in France, gathered in Table 2a. *CAT* 1 to 3 are three of the largest categories[3] in terms of revenue while *CAT 4* is a smaller category. The following dominances can be highlighted.

[2] This threshold is fixed at 0 for convenience, but it could have any value.

[3] For privacy reason, the exact content of the different categories cannot be revealed.

CAT 1	CAT 2	CAT 3	CAT 4
0.0133	0.0140	0.0089	0.0069
0.0114	0.0088	0.0086	0.0063
0.0108	0.0083	0.0078	0.0053
0.0107	0.0082	0.0056	0.0051
0.0096	0.0078	0.0052	0.0051
0.0095	0.0078	0.0050	0.0044
0.0088	0.0078	0.0049	0.0042
0.0086	0.0077	0.0047	0.0041
0.0084	0.0076	0.0042	0.0040
0.0080	0.0074	0.0041	0.0038

(a) Click-through rates



(b) Cumulative distribution functions

Figure 2: Illustrations of dominance on a real dataset

Strong dominance *CAT 1* strongly dominates *CAT 4* as its minimum CTR is 0.008 compared to the maximum CTR of 0.0069 of the other. Similarly, *CAT 2* strongly dominates *CAT 4*.

First-order dominance *CAT 2* first-order dominates *CAT 3* as the CTR of each line of the second column are bigger than those of the third column. This dominance is not strong as 0.0074 is smaller than 0.0089. *CAT 3* first-order but not strongly dominates *CAT 4*.

Uncomparable categories *CAT 1* and *CAT 2* are not comparable with respect to any partial order.

Notice that, had the first item of *CAT 2* performed only 5% worse than observed,[4] then *CAT 1* would have been optimal with respect to the first-order dominance. So even if the dominance assumption is not satisfied during that specific month, assuming it would still give good empirical results. The relations of dominance can be easier to determine based on the representation of the associated cdf of Figure 2b. As the cdf of the random variable uniform on *CAT 4* is, pointwise, the biggest one, this means that this category is first-order dominated by all the other ones. Moreover, it reaches 1 while the cdf of *CAT 1* and *CAT 2* are still at 0. This implies that the dominance of these two categories is even strong. This analysis motivates and validates our assumption.

3 Lower bounds

In this section, we provide lower bounds on the regret that any "reasonable" algorithm (the precise definition is given below) must incur in a multi-armed bandit problem, where arms are grouped into partially ordered categories (with a dominating one). To simplify the exposition, we assume here that noises are drawn from Gaussian distribution with unit variance. The class of algorithms we consider are consistent [27] with respect to a given a class of possible bandit problems $\mathcal{M} = \big\{\mu = (\mu_1, \dots, \mu_{MK}) \in \mathbb{R}^{MK}\big\}$. We recall that an algorithm is consistent with $\mathcal{M}$ if, for any admissible reward vector $\mu \in \mathcal{M}$ and any parameter $\alpha \in (0, 1]$, the regret of that algorithm is asymptotically negligible compared to T^α, i.e., $\sup_{\alpha\in(0,1)} \limsup_{T\to\infty} \frac{\mathbb{E}_\mu[R_T]}{T^\alpha} = 0$. Graves and Lai [16] proved that any algorithm consistent with $\mathcal{M}$ has a regret scaling at least logarithmically in T, with a leading constant c_μ depending on μ (and $\mathcal{M}$) i.e., $\liminf_{T\to\infty} \frac{\mathbb{E}_\mu[R_T]}{\log(T)} \geq c_\mu$; moreover, c_μ is the solution of some auxiliary optimization problem. In our setting, it rewrites as

$$c_\mu = \min_{N\geq 0} \sum_{m,k} N_k^m \Delta_{m,k} \quad \text{subject to} \quad \sum_{m,k} N_k^m \left(\mu_k^m - \lambda_k^m\right)^2 \geq 2, \forall\, \lambda \in \Lambda(\mu)\,,$$

where $\Lambda(\mu) = \big\{\lambda \in \mathcal{M}; \mu_1^1 = \lambda_1^1, \lambda_1^1 < \max_{m,k} \lambda_k^m\big\}$. We point out that the assumption of dominance is hidden in the class of bandit problem $\mathcal{M}$. In the remaining and with a slight abuse of notation, we are going to call an algorithm consistent with a dominance assumption if it is consistent with the set of all possible vectors of means satisfying this dominance assumption.

[4] The CTR of the best item of *CAT 2* is so higher than the second one, we could expect it is actually an outlier, i.e., an artefact of the choice of that specific month and category.

Group-sparse dominance In this case, the above optimization problem has a closed-form solution.

Theorem 3.1. *An algorithm consistent with the group-sparse dominance satisfies* $c_\mu = \sum_{k=2}^{K} \frac{2}{\Delta_{1,k}}$.

The proof of this result (and the subsequent ones) is postponed to the Appendix. This lower bound indicates that all arms in the optimal category (and only those) should be pulled a logarithmic number of times, hence the regret should only scale asymptotically linearly in the number of arms in the optimal category instead of linearly with the total number of arms. We want to stress out here that Theorem 3.1 might have a misleading interpretation. Although the asymptotic regret scales with K and independently of M, the finite-stage minimax regret is still of the order of $\sqrt{MKT}$, as with usual bandits. This is simply because the lower-bound proof [6] of the standard multi-armed bandit case uses set of parameters of the form $(0, \dots, 0, \varepsilon, 0, \dots, 0)$ which respect the group-sparse assumption. As a result, the asymptotic lower bound of Theorem 3.1 is hiding some finite-time dependency in MK (possibly of the form of an extra-term in $\sum_{m,k} 1/\Delta_{m,k}$, yet independent of $\log(T)$) that non-asymptotic algorithms[5] would not be able to remove.

Strong dominance In the case of strong dominance, a similar closed-form expression can be stated.

Theorem 3.2. *With strong dominance, a consistent algorithm verifies* $c_\mu = \sum_{k=2}^{K} \frac{2}{\Delta_{1,k}} + \sum_{m=2}^{M} \frac{2}{\Delta_{m,K}}$.

This lower bound indicates that the dominance assumption can be leveraged to replace the asymptotic linear dependency in the total number of arms category into a linear dependency in the number of arms of the optimal category plus the number of categories. With M categories of K arms each, the dependency in MK is replaced into $M + K$. However, as before and for the same reasons, the finite-time minimax lower bound will still be of the order $\sqrt{MKT}$. The lower bound of Theorem 3.2 seems to indicate that an optimal algorithm should be pulling only the arms of the optimal category and the **worst** arm (not the best!) of the other categories, at least asymptotically and logarithmically. Yet again, there is no guarantee that non-asymptotic algorithms can achieve this highly-demanding (and rather counter-intuitive) lower bound.

First-order dominance There are no simple closed form expression of c_μ with the first-order dominance assumption, see nonetheless Appendix A.3 for some variational expression. However, for the sake of illustration, we provide a closed-form solution for a specific case.

Theorem 3.3. *With first-order dominance and* $M = K = 2$ *and assuming that arms are intertwined, i.e.* $\mu_1^1 > \mu_1^2 > \mu_2^1 > \mu_2^2$*, a consistent algorithm satisfies*

$$c_\mu = \frac{2}{\Delta_{1,2}} + \frac{2}{\Delta_{2,2}} + \frac{2}{\Delta_{2,1}} \left(1 - \frac{(\Delta_{2,2} - \Delta_{1,2})^2}{(\Delta_{1,2})^2 + (\Delta_{2,2})^2} \right).$$

It is quite interesting to compare this lower bound to the corresponding ones with group-sparsity where $c_\mu = \frac{2}{\Delta_{1,2}}$, with strong dominance where $c_\mu = \frac{2}{\Delta_{1,2}} + \frac{2}{\Delta_{2,2}}$ and without structure at all where $c_\mu = \frac{2}{\Delta_{1,2}} + \frac{2}{\Delta_{2,2}} + \frac{2}{\Delta_{2,1}}$. Clearly, lower bounds are, as expected, decreasing with additional structure. More interestingly, the first-order lower bound somehow interpolates between this two by multiplying the term $\frac{2}{\Delta_{2,1}}$ by a factor $\rho \in (0, 1)$; $\rho = 0$ corresponding to the stronger assumption of strong dominance and $\rho = 1$ to the absence of dominance assumption.

4 Algorithms and upper bounds

4.1 Optimism principle

Our first algorithm is based on the principle of optimism in the face of uncertainty and is summarized in Algorithm 1. It behaves in three different ways depending on the number of categories that are called

[5] We call an algorithm non-asymptotic if its worst-case regret is of the order of $\sqrt{MKT}$, maybe up to some additional polynomial dependency in M and K. In particular, classical algorithms for structured bandits [8, 29] are only asymptotical.

"active". The definition of an active category will depend on the assumption of dominance. Formally, let $\delta \in (0,1)$ be a confidence level (fixing the confidence level actually requires that the horizon T is known, but there exist well understood anytime version of all these results [12]). At time step t, it computes the set of active categories, denoted $\mathcal{A}(t,\delta)$. The three states of Algorithm 1 are then as follows:

1. $|\mathcal{A}(t,\delta)| = 0$: no category is active; the algorithm pulls all arms.
2. $|\mathcal{A}(t,\delta)| = 1$: only one category is active; the algorithm performs UCB(δ) in it.
3. $|\mathcal{A}(t,\delta)| > 1$: several categories are active; the algorithm pulls all arms inside those.

We now detail what we called an active category for each notion of dominance defined previously along with theorems upper bounding the regret of the CATSE algorithm.

Algorithm 1: CATSE(δ)

```
Pull each arm once
while t ≤ T do
    Compute set of active categories A(t, δ)
    if |A(t, δ)| = 0 then
        Pull all arms
    else if |A(t, δ)| = 1 then
        Perform UCB(δ) in the active
         category
    else
        Pull all arms in active categories
    end
end
```

Group-sparse dominance Under this assumption, we say a category is active if it has an active arm. Following the idea of sparse bandits [26] or bounded regret [7], we say that the arm k of category m is active if

$$\widehat{\mu}_k^m(t) := \frac{\sum_{s<t;(C_t,A_t)=(m,k)} X_{A_t}^{C_t}}{N_k^m(t)} \geq 2\sqrt{\frac{\log N_k^m(t)}{N_k^m(t)}}\,.$$

This condition ensures that the expected number of times an arm with positive mean is non active is finite in expectation. Similarly, the expected number of times an arm with non positive mean is active is also finite. Those conditions will ensure that the expected number of times a suboptimal category is pulled is also finite. Then, the set of active categories, denoted $\mathcal{A}(t)$ is simply

$$\mathcal{A}(t) \coloneqq \left\{ m \in [M]; \exists\, k \in [K], \widehat{\mu}_k^m(t) \geq 2\sqrt{\frac{\log N_k^m(t)}{N_k^m(t)}} \right\}.$$

Theorem 4.1. *In the group-sparse dominance setting, the expected regret of* CATSE *verifies with probability at least* $1 - 2\delta KT$*,*

$$\mathbb{E}[R_T] \leq \sum_{k=2}^{K} \frac{8\log\frac{1}{\delta}}{\Delta_{1,k}} + \sum_{m,k} \Delta_{m,k} + \frac{40}{(\mu_1^1)^2}\log\frac{16}{(\mu_1^1)^2}\sum_{m,k}\Delta_{m,k} + (M-1)K\frac{\pi^2}{6}\sum_{m,k}\Delta_{m,k}\,.$$

The first term is the bound of the UCB algorithm while the third term is the regret incurred when the optimal category is non active and the last term comes from a suboptimal category being active. As a result, CATSE is asymptotically optimal, up to a multiplicative factor. A trick to improve empirically the performance of the algorithm is to replace the round-robin sampling phase (when $|\mathcal{A}(t)| = 0$) by choosing an arm with a higher probability the closer it is to be active. This idea was analyzed in [7] with additional assumptions. Yet this can only improve the second term of the regret, which is already constant w.r.t. T (so we chose to not focus on it). For example, a possibility is to pull arm (m,k) at time t with probability $p_k^m(t) \propto \left(\sqrt{\frac{4\log N_k^m(t)}{N_k^m(t)}} - \widehat{\mu}_k^m(t)\right)^{-2}$. Another possible improvement is to eliminate categories in which there exist an arm whose upper bound is less than 0. Again, this only improves a term constant w.r.t. T.

Strong dominance In this setting, CATSE will use the information gathered by all arms. The overall idea is to construct confidence region for the mean vector and to eliminate a category as soon as it is clearly dominated by another one. The statistical test to perform in order to determine which categories to eliminate is based on the following alternative characterization of dominance.

Let $\Delta(K) \coloneqq \{\mathbf{x} \in \mathbb{R}_+^K; \|\mathbf{x}\|_1 = 1\}$ be the K-simplex and $\mu^m \coloneqq (\mu_k^m)_k$ be the vector of means.

Proposition 1. $\mathcal{A}$ strongly dominates $\mathcal{B}$ if and only if $\forall\, \mathbf{x} \in \Delta(K), \forall\, \mathbf{y} \in \Delta(K), \langle \mathbf{x}, \mu^{\mathcal{A}} \rangle \geq \langle \mathbf{y}, \mu^{\mathcal{B}} \rangle$.

At the end of the p-th round of the phase of successive elimination of categories, each arm has been pulled p times. A natural estimator of $\mu^m \in \mathbb{R}^K$ is the coordinate wise empirical average of rewards,

i.e., $\mu_k^m(p) = \frac{1}{p}\sum_{r=1}^{p} X_k^m(r)$, where (with a slight abuse of notation), $X_k^m(r)$ is the reward gathered by the r-th pull of arm k of category m. We now describe the statistical run at the end of round $p \in \mathbb{N}$; category $n \in [M]$ is eliminated by category $m \in [M]$ if it holds that

$$L_m^+(p,\delta) \coloneqq \max_{\mathbf{x}\in\Delta(K)} \langle \mathbf{x}, \widehat{\mu}^m(p)\rangle - \|\mathbf{x}\|_2\,\beta(p,\delta) > \min_{\mathbf{y}\in\Delta(K)} \langle \mathbf{y}, \widehat{\mu}^n(p)\rangle + \|\mathbf{y}\|_2\,\beta(p,\delta) =: L_n^-(p,\delta)\,, \tag{1}$$

where $\beta(p,\delta) = \sqrt{\frac{2}{p}\left(K\log 2 + \log\frac{1}{\delta}\right)}$. The set of active categories is then define as follows

$$\mathcal{A}(t,\delta) = \left\{m \in [M]; \forall\, n \neq m, L_n^+(t,\delta) \leq L_m^-(t,\delta)\right\}\,.$$

Theorem 4.2. *In the strong dominance case, the regret of* CATSE *satisfies w.p. at least* $1-\delta MT$*,*

$$R_T \leq \sum_{k=2}^{K} \frac{8\log\frac{1}{\delta}}{\Delta_{1,k}} + \sum_{m,k}\Delta_{m,k} + 8\big(K\log 2 + \log\frac{1}{\delta}\big)\sum_{m=2}^{M}\min_{\mathbf{x},\mathbf{y}\in\Delta(K)}\Big(\frac{\|\mathbf{x}\|_2 + \|\mathbf{y}\|_2}{\langle \mathbf{x},\mu^1\rangle - \langle \mathbf{y},\mu^m\rangle}\Big)^2 \sum_{k=1}^{K}\Delta_{m,k}$$

First-order dominance CATSE will proceed with first-order dominance as with strong dominance, the major difference is the statistical test. Let us first characterize the notion of first-order dominance.

Proposition 2. $\mathcal{A}$ first-order dominates $\mathcal{B}$ if and only if $\forall\,\mathbf{x}\in\Delta(K), \langle \mathbf{x},\mu^{\mathcal{A}}\rangle \geq \langle \mathbf{x},\mu^{\mathcal{B}}\rangle$.

The statistical test is then: category $n \in [M]$ is eliminated by category $m \in [M]$ at round p if

$$D_{m,n}(p,\delta) \coloneqq \max_{\mathbf{x}\in\Delta(K)} \frac{\langle \mathbf{x}, \widehat{\mu}_\sigma^m(p) - \widehat{\mu}_\tau^n(p)\rangle}{\|\mathbf{x}\|_2} > 2\gamma(p,\delta)\,, \tag{2}$$

where $\widehat{\mu}_\sigma^m(p)$ and $\widehat{\mu}_\tau^n(p)$ represent respectively the reordering of $\widehat{\mu}^m(p)$ and $\widehat{\mu}^n(p)$ in decreasing order and $\gamma(p,\delta) = \frac{1}{\sqrt{2p}}\big(\sqrt{K\log\frac{1}{\delta}} + \sqrt{1+(K+1)\log K}\big)$. We emphasis the permutation is specific to both a category and a round. This statistical test yields the following set of active categories

$$\mathcal{A}(t,\delta) = \{m \in [M]; \forall\, n \neq m, D_{m,n}(t,\delta) \leq 2\gamma(t,\delta)\}\,.$$

Theorem 4.3. *Under the additional assumption that* $X_k^m \in [0,1]$ *for all category* m *and arm* k*, in the first-order dominance, the regret of* CATSE *verifies with probability at least* $1-\delta MT$*,*

$$R_T \leq \sum_{k=2}^{K}\frac{8\log\frac{1}{\delta}}{\Delta_{1,k}} + \sum_{m,k}\Delta_{m,k} + 16\left(K\log\frac{1}{\delta} + K\log K + \log K + 1\right)\sum_{m=2}^{M}\frac{\sum_{k=1}^{K}\Delta_{m,k}}{\|\mu^1-\mu^m\|_2^2}\,.$$

4.2 Bayesian principle

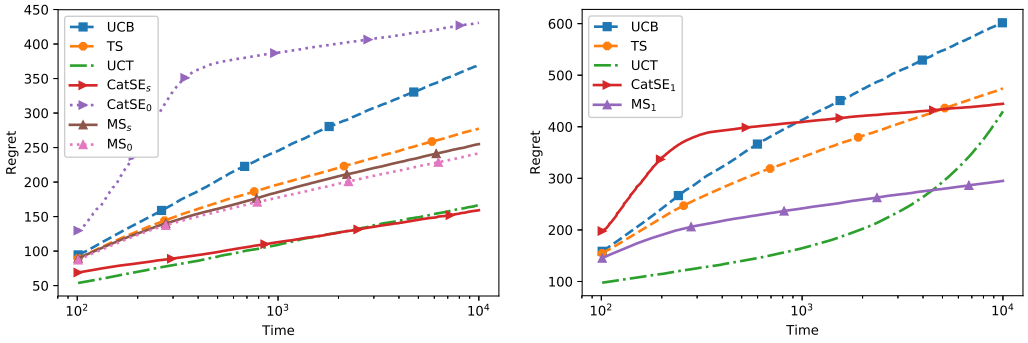
The MURPHY SAMPLING (MS) algorithm [20] was originally developed in a pure exploration setting. Conceptually, it is derived from THOMPSON SAMPLING (TS) [37], the difference is that the sampling respects some inherent structure of the problem. To define MS, we denote by $\mathcal{F}(t) = \sigma\left(A_1, X_1, \ldots, A_t, X_t\right)$ the information available after t steps and $\mathcal{H}_d$ the assumption of dominance considered. Let $\Pi_t = \mathbb{P}\left(\cdot|\mathcal{F}_t\right)$ be the posterior distribution of the means parameters after t rounds. The algorithm samples, at each time step, from the posterior distribution $\Pi_{t-1}\left(\cdot|\mathcal{H}_d\right)$ and then pulls the best arm, which, by definition, is in the best category sampled at this time step. In comparison, TS would sample from Π_{t-1} without taking into account any structure. To implement this algorithm, we use that independent conjugate priors will produce independent posteriors, making the posterior sampling tractable. The required assumption, i.e. the structure of our problem, is then attained using rejection sampling. We do not provide theoretical guarantees on its regret but we will illustrate empirically on simulated data that it is highly competitive compared to the other algorithms.

Algorithm 2: MURPHY SAMPLING

while $t \leq T$ **do**
  Sample $\theta(t) \sim \Pi_{t-1}\left(\cdot|\mathcal{H}_d\right)$
  Pull $(C_t, A_t) \in \arg\max_{(m,k)} \theta_k^m(t)$
end

5 Experiments

In this section, we present numerical experiments illustrating the performance of the algorithms we have introduced. We also compare them with two families of algorithms. The first one is algorithms



(a) In the sparse and strong dominance scenario

(b) In the first-order dominance scenario

Figure 3: Regret of various algorithms as a function of time

for the multi-armed bandit framework, namely UCB [3] and TS [37]; they are agnostic to the structure of the arms. The second family of algorithms is adapted to tree search, namely UCT [23]; they partially take into account the inherent structure. Specifically, they will just use the fact that arms are grouped into categories but not that one category dominates the others. We consider two scenarios for the different dominance hypothesis. In all experiments, rewards are drawn from Gaussian distribution with unit variance and we report the average regret as a function of time, in log-scale. To implement TS and MS, we pulled each arm once and then sampled using a Gaussian prior. The simulations were ran until time horizon 10,000 and results were averaged over 100 independent runs.

Group-sparse & strong dominance We start by grouping the experiments in the group-sparse and strong dominance setting, as we recall that the only difference between the two concepts is the knowledge of a threshold between the best category and the others. In this first scenario, we analyze a problem with five categories and five arms per category. Precisely, in the first category the optimal arm has expected reward 1, and the four suboptimal arms consist of one group of three (stochastically) identical arms each with expected reward 0.5 and one arm with expected reward 0. The four suboptimal category are identical and are composed of two arms with expected rewards 0 and -1, respectively and a group of three arms with expected reward -0.5. We used the subscript s and 0 to denote the assumption of dominance the algorithm exploited. CATSE$_s$ and CATSE$_0$ were run with $\delta = \frac{1}{t}$ and $\delta = \frac{1}{Mt}$, respectively. Results are presented on Figure 3a. In the case of group-sparse dominance, CATSE$_s$ outperforms both UCB and UCT; MS$_s$ asymptotically performs as well yet with a slightly higher regret. Interestingly, UCT performs well in the beginning; thanks to the lack of an exploration phase compared to CATSE$_s$. In the case of strong dominance, MS$_0$ and CATSE$_0$ asymptotically perform alike and slightly better than UCT. However, the regret of CATSE$_0$ is much higher due to its round-robin sampling phase; this can be seen in the beginning as CATSE$_0$ is still in the search of the optimal category. If we compare the two versions of each algorithm between them, we can notice two points. Firstly, for CATSE, the result of the potential sampling improvement is significant. Secondly, for MS, the regret in the group-sparse case is slightly worse than in the strong dominance case even though it is stronger. This is simply due to our implementation and the difficulty of the posterior sampling, in particular the rejection sampling phase.

First-order dominance Finally, we consider the first-order dominance setting. In this scenario, we look upon a problem with five categories and ten arms per category. Precisely, in the optimal category, the best arm has expected reward 5 while the nine suboptimal arms consist of three group of five, three and one arms, with expected rewards 4, 3 and 2, respectively. The four suboptimal categories are composed of two arms with expected rewards 4.5 and 0, respectively, and eight arms with expected reward 3. CATSE was run with $\delta = \frac{1}{Mt}$ and the results are presented on Figure 3b. Once again, MS and CATSE outperform baseline algorithms and both appear to have the same slope asymptotically with a significant difference between their regret, again due to the exploration phase of CATSE. It is interesting to observe that UCT performed poorly; as noticed in [9], the convergence can be sluggish. Indeed, the main issue occurs when the best arm is underestimated. In that case, it is pulled a logarithmic number of times the optimal category is pulled, which is a logarithmic

number of times, since the second best arm overall is in suboptimal categories. Hence, it would take an exponential of exponentials number of time for the optimal arm to become the best again.

6 Conclusion

Two problems remain open: the first one is a better exploration phase in CATSE since it heavily impact the regret and as noted in [14], ETC algorithms are necessarily suboptimal; and the second is an upper bound on the regret of the MS algorithm since it is highly competitive in practice. We believe that it is asymptotically optimal and that it can be applied to other setting of structured bandits.

Acknowledgments

This work was supported in part by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH, in a joint call with Gaspard Monge Program for optimization, operations research and their interactions with data sciences.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[2] Naoki Abe and Philip M Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pages 3–11, 1999.

[3] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[4] Vijay S Bawa. Optimal rules for ordering uncertain prospects. *Journal of Financial Economics*, 2(1):95–121, 1975.

[5] Guy Bresler, George H Chen, and Devavrat Shah. A latent source model for online collaborative filtering. In *Advances in Neural Information Processing Systems*, pages 3347–3355, 2014.

[6] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[7] Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, pages 122–134, 2013.

[8] Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 1761–1769, 2017.

[9] Pierre-Arnaud Coquelin and Rémi Munos. Bandit algorithms for tree search. *arXiv preprint cs/0703062*, 2007.

[10] Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pages 355–366, 2008. URL `http://colt2008.cs.helsinki.fi/papers/80-Dani.pdf`.

[11] Herbert Aron David and Haikady Navada Nagaraja. *Order statistics*. Wiley, third edition, 2003.

[12] Rémy Degenne and Vianney Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In *International Conference on Machine Learning*, pages 1587–1595, 2016.

[13] Rémy Degenne and Vianney Perchet. Combinatorial semi-bandit with known covariance. In *Advances in Neural Information Processing Systems*, pages 2972–2980, 2016.

[14] Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems*, pages 784–792, 2016.

[15] Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *International Conference on Machine Learning*, pages 757–765, 2014.

[16] Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws incontrolled markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997.

[17] Josef Hadar and William R Russell. Rules for ordering uncertain prospects. *The American economic review*, 59(1):25–34, 1969.

[18] Sumeet Katariya, Branislav Kveton, Csaba Szepesvari, Claire Vernade, and Zheng Wen. Stochastic rank-1 bandits. *arXiv preprint arXiv:1608.03023*, 2016.

[19] Sumeet Katariya, Lalit Jain, Nandana Sengupta, James Evans, and Robert Nowak. Adaptive sampling for coarse ranking. *arXiv preprint arXiv:1802.07176*, 2018.

[20] Emilie Kaufmann, Wouter Koolen, and Aurelien Garivier. Sequential test for the lowest mean: From thompson to murphy sampling. *arXiv preprint arXiv:1806.00973*, 2018.

[21] Jaya Kawale, Hung H Bui, Branislav Kveton, Long Tran-Thanh, and Sanjay Chawla. Efficient thompson sampling for online matrix-factorization recommendation. In *Advances in neural information processing systems*, pages 1297–1305, 2015.

[22] Tomáš Kocák, Michal Valko, Rémi Munos, and Shipra Agrawal. Spectral thompson sampling. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[23] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer, 2006.

[24] Nathan Korda, Balázs Szörényi, and Li Shuai. Distributed clustering of linear bandits in peer to peer networks. In *Journal of machine learning research workshop and conference proceedings*, volume 48, pages 1301–1309. International Machine Learning Societ, 2016.

[25] Joon Kwon and Vianney Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. *The Journal of Machine Learning Research*, 17(1):8106–8137, 2016.

[26] Joon Kwon, Vianney Perchet, and Claire Vernade. Sparse stochastic bandits. In *30th Annual Conference on Learning Theory - COLT 2017, Amsterdam, Netherlands, July 7-10, 2017*, pages 355–366, 2017.

[27] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[28] Tor Lattimore and Rémi Munos. Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems*, pages 550–558, 2014.

[29] Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *20th International Conference on Artificial Intelligence and Statistics*, pages 728–737, 2017.

[30] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[31] Shuai Li, Claudio Gentile, and Alexandros Karatzoglou. Graph clustering bandits for recommendation. *arXiv preprint arXiv:1605.00596*, 2016.

[32] Shuai Li, Alexandros Karatzoglou, and Claudio Gentile. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 539–548. ACM, 2016.

[33] Odalric-Ambrym Maillard and Shie Mannor. Latent bandits. In *International Conference on Machine Learning*, pages 136–144, 2014.

[34] Trong T Nguyen and Hady W Lauw. Dynamic clustering of contextual multi-armed bandits. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1959–1962. ACM, 2014.

[35] Pierre Perrault, Vianney Perchet, and Michal Valko. Finding the bandit in a graph: Sequential search-and-stop. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1668–1677, 2019.

[36] Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

[37] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[38] Michal Valko, Rémi Munos, Branislav Kveton, and Tomáš Kocák. Spectral bandits for smooth graph functions. In *International Conference on Machine Learning*, pages 46–54, 2014.

[39] Xiaoxue Zhao, Weinan Zhang, and Jun Wang. Interactive collaborative filtering. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1411–1420. ACM, 2013.