Stochastic bandits with groups of similar arms

Fabien Pesquerel, Hassan Saber, Odalric-Ambrym Maillard

Stochastic bandits with groups of similar arms

Fabien Pesquerel* **Hassan Saber** **Odalric-Ambrym Maillard**
fabien.pesquerel@inria.fr hassan.saber@inria.fr odalric.maillard@inria.fr

Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9198-CRIStAL, F-59000 Lille, France

Abstract

We consider a variant of the stochastic multi-armed bandit problem where arms are known to be organized into different groups having the same mean. The groups are unknown but a lower bound q on their size is known. This situation typically appears when each arm can be described with a list of categorical attributes, and the (unknown) mean reward function only depends on a subset of them, the others being redundant. In this case, q is linked naturally to the number of attributes considered redundant, and the number of categories of each attribute. For this structured problem of practical relevance, we first derive the asymptotic regret lower bound and corresponding constrained optimization problem. They reveal the achievable regret can be substantially reduced when compared to the unstructured setup, possibly by a factor q. However, solving exactly the exact constrained optimization problem involves a combinatorial problem. We introduce a lower-bound inspired strategy involving a computationally efficient relaxation that is based on a sorting mechanism. We further prove it achieves a lower bound close to the optimal one up to a controlled factor, and achieves an asymptotic regret q times smaller than the unstructured one. We believe this shows it is a valuable strategy for the practitioner. Last, we illustrate the performance of the considered strategy on numerical experiments involving a large number of arms.

1 Introduction

The finite stochastic multi-armed bandit problem is a popular framework for studying sequential decision making problems in which a learner sequentially samples from a finite set of distributions called arms. It was first introduced in the context of medical trials [Thompson, 1933b, 1935] and later formalized under this name by Robbins in Robbins [1952]. We refer the interested reader to Lattimore and Szepesvári [2020] for a recent survey. This is one of the simplest theoretical framework in which one can study the notion of *exploration-exploitation* tradeoff. This tension between *exploration* and *exploitation* arises from the sequential optimization problem a learner is trying to perform while being uncertain about the very problem it is optimizing.

Formally, a multi-armed bandit configuration is specified by a set of unknown real-valued probability distributions $\nu = (\nu_a)_{a\in\mathcal{A}}$ with means $(\mu_a)_{a\in\mathcal{A}}$, indexed by a set of arms $\mathcal{A}$. We hereafter consider a finite $\mathcal{A}$, and that all ν_a, $a \in \mathcal{A}$ belong to the same family of distributions $\mathcal{F}$ (e.g. Bernoulli, Gaussian, etc.), that is $\nu \in \mathcal{F}^{\mathcal{A}}$. The bandit game proceeds as follows. At each time $t \in \mathbb{N}$, the learner chooses an arm $a_t \in \mathcal{A}$ based on the past observations and decisions, then receives and observes a sample X_t (called the reward), conditionally independent, sampled from ν_{a_t}. Her goal is to maximize the cumulative reward received over time. The mean of each arm is unknown, which makes the problem non-trivial, hence the learner should adjust her sampling strategy based on past information obtained from drawing different arms in order to maximize the expected sum of rewards. The maximal expected value of a finite bandit configuration is denoted by $\mu_\star$, defined as $\mu_\star = \max_{a\in\mathcal{A}} \mu_a$. The performance of the strategy used by the agent is measured by the (pseudo) *regret*, that compares the

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

expected sum of rewards obtained by an oracle that would constantly pull an optimal arm and the ones obtained by the learner, up to some time horizon T (that we assume is unknown to the learner).

Definition 1 (Regret)**.** *The regret incurred by a sampling strategy after T time steps on a bandit configuration ν is given by:*

$$\mathcal{R}(\nu, T) = \mathbb{E}_\nu \left(\sum_{t=1}^{T} (\mu_* - \mu_{a_t}) \right) = \sum_{a \in \mathcal{A}} (\mu_* - \mu_a)\, \mathbb{E}_\nu \left(N_a(T) \right),$$

where $N_a(T) = \sum_{t=1}^{T} \mathbb{I}\{a_t = a\}$ denotes the number of selection of arm a after T time steps.

Group of similar arms Motivated by various practical reasons, one may want to restrict to a subset $\mathcal{B} \subset \mathcal{F}^{\mathcal{A}}$ of allowed bandit configurations instead of the full set $\mathcal{F}^{\mathcal{A}}$. In this paper, we study a variant of the multi-armed bandit problem in which the reward function, $\mu : a \in \mathcal{A} \to \mu_a$, is assumed to satisfy a cluster-like structural property. A bandit configuration ν is said to satisfy the **q-equivalence property** if for every arm $a \in \mathcal{A}$, there are at least $q-1$ distinct arms having the same expected value:

$$\forall a \in \mathcal{A}, \quad |\{a' \in \mathcal{A} | \mu_{a'} = \mu_a\}| \geqslant q.$$

Assuming the set of arms $\mathcal{A}$ and base distributions $\mathcal{D}$ are known to the learner, we denote by $\mathcal{B}_q$ the set of bandit configurations having the q-equivalence property. We also denote by $\mathcal{B}_q(\mu)$ the set of all expected values in $\mathcal{B}_q$. Formally, $\mathcal{B}_q(\mu)$ is the image of $\mathcal{B}_q$ under the μ mapping.

Definition 2 (Arm equivalence and equivalence class)**.** *Given a bandit configuration ν, two arms $a, a' \in \mathcal{A}$ are said to be equivalent if their associated distributions have the same expected values:*

$$a \sim a' \Leftrightarrow \mu_a = \mu_{a'}.$$

An equivalence class c in ν is a maximal subset of arms in $\mathcal{A}$ having the same mean, i.e., for all arms a, a' in c, $\mu_a = \mu_{a'}$ and for all arm $a \in c$ and $a' \in \mathcal{A} \setminus c$, $\mu_a \neq \mu_{a'}$.

This situation typically appears in practical situations when each arm can be described with a list of categorical attributes, and the (unknown) mean reward function only depends on a subset of them, the others being redundant. In this case, q is naturally linked to the number of attributes considered redundant (or useless descriptors), and the number of categories of each attribute. Precisely, $q = \prod_{i \in \mathcal{R}} c_i$ where $\mathcal{R}$ is the set of redundant attributes and c_i the number of categories for attribute i. The learner may know that there exists such a structure while not knowing a closed form formula mapping the list of categorical attributes to the significant subset. In this case, q might be a lower bound on the sizes of the class since the set $\mathcal{R}$ might not be the largest possible one or because the number of redundant attributes depends on the number of relevant attributes. In all cases, the smallest possible number of redundant attributes can be naturally linked to q. We hereafter consider the learner only knows q but would like to exploit the prior knowledge of this structure in a bandit problem.

Regret lower bounds overview In order to assess the performance of a bandit algorithm on a set of configurations $\mathcal{B}$, one naturally studies the best guarantee achievable by a uniformly efficient algorithm on $\mathcal{B}$, i.e with sub-linear regret on any instance $\nu \in \mathcal{B}$ of the bandit problem. When $\mathcal{B} = \mathcal{F}^{\mathcal{A}}$, such a guarantee was first provided by Lai and Robbins [1985] for parametric families $\mathcal{F}$, and then extended by Burnetas and Katehakis [1996] for more general families. It states that any algorithm that is uniformly efficient[1] on a family of distributions $\mathcal{F}$ must satisfy

$$\liminf_{T \to \infty} \frac{\mathcal{R}(\nu, T)}{\log(T)} \geqslant \sum_{a : \mu_\star > \mu_a 0} \frac{\mu_\star - \mu_a}{\mathcal{K}_{\mathcal{F}}(\nu_a \| \mu^\star)}, \quad \mathcal{K}_{\mathcal{F}}(\nu_a \| \mu^\star) = \inf_{G \in \mathcal{F}} \{ \mathrm{KL}(\nu_a \| G) : \mathbf{E}_G(X) > \mu^\star \}. \quad (1)$$

This popular result entails that any strategy having the desirable property to have sub-linear regret on any instance in $\mathcal{F}$ must incur a non-trivial minimal regret. When $\mathcal{B}$ is a strict subset of $\mathcal{F}^{\mathcal{A}}$, the bandit problem is called structured, as in this case pulling an arm may reveal information that makes it possible to refine estimation of other arms (e.g. think of the set of bandit configurations having Lipschitz mean function with respect to $\mathcal{A} \subset \mathbb{R}^d$). The presence of structure may considerably modify the achievable lowest regret, as shown in Burnetas and Katehakis [1996] and Graves and Lai

[1] Formally, for each bandit on $\mathcal{F}$, for each arm k with $\Delta_k > 0$, then $\mathbf{E}[N_k(T)] = o(T^\alpha)$ for all $\alpha \in (0, 1]$.

[1997], who extended the (unstructured) lower bounds to arbitrarily structured bandit problems (and beyond). These lower bound take the generic form $\liminf_{T\to\infty} \dfrac{\mathcal{R}(\nu,T)}{\log(T)} \geqslant \mathfrak{C}_{\mathcal{B}}(\nu)$, where $\mathfrak{C}_{\mathcal{B}}(\nu)$ is a constant term solution of a constrained linear-optimization problem. A bandit algorithm is then called *asymptotically optimal* for a set $\mathcal{B}$ when its regret asymptotically matches this lower bound.

Existing strategies In order to minimize the regret, a learner faces the classical exploration/exploitation dilemma: it needs to balance *exploration*, that is gaining information about the expected values of the arms by sampling them, and *exploitation*, that is playing the most promising arm sufficiently often. Many algorithms have been proposed to solve the multi-armed bandits problem (see Lattimore and Szepesvári [2020] for a survey). The study of the lower bounds had a crucial impact on the development of provably asymptotically optimal strategies. In the case of *unstructured* bandit $\mathcal{B} = \mathcal{F}^{\mathcal{A}}$, this includes strategies that build on the concept of *Optimism in Face of Uncertainty* (the most celebrated of which being the Upper Confidence Bound (UCB) algorithms Agrawal [1995], Auer et al. [2002]), such as `KLUCB` [Lai, 1987, Cappé et al., 2013, Maillard, 2018], `DMED` and `IMED` Honda and Takemura [2011, 2015], that are proven asymptotically optimal for various families $\mathcal{F}$ (e.g. one-dimensional exponential families), and directly exploit the lower bound in their structure. Alternative asymptotically optimal strategies include the Thompson Sampling (TS) Thompson [1933a], Agrawal and Goyal [2012], which uses a Bayesian posterior distribution given a specific prior, whose optimality was shown in Korda et al. [2013]. See also Kveton et al. [2019] for other randomized algorithms and Kveton et al. [2020], Chan [2020], Baudry et al. [2020] for recent non-parametric extensions using re-sampling methods. Further, some authors also allow many optimal arms, see de Heide et al. [2021], or even countably many arms, see Kalvit and Zeevi [2020]. However, these works do not consider nor exploit a constraint on the level-sets of the mean function and follow an optimistic paradigm while we follow an information minimization targeting optimality. On the other hand, several instances of structured bandits received considerable attention in the last few years. This is the case for instance of linear bandits, see [Abbasi-Yadkori et al., 2011, Srinivas et al., 2010, Durand et al., 2017, Kveton et al., 2020] and Lattimore and Szepesvari [2017], Lipschitz bandits Magureanu et al. [2014], Wang et al. [2020], Lu et al. [2019], unimodal bandits Yu and Mannor [2011], Combes and Proutiere [2014], Saber et al. [2020], or combinatorial bandits Kveton et al. [2015], Magureanu [2018], and more recently Cuvelier et al. [2021b]. A generic asymptotically optimal algorithm, called `OSSB` (Optimal Structured Stochastic Bandit), has been introduced in the work of Combes et al. [2017], and proven to be asymptotically optimal for all structures satisfying some weak properties that include all the aforementioned structures. Although being asymptotically optimal this algorithm often suffers from a long burn-in phase that may hinder its finite practical performance. It further comes with high computational price as it requires to solve an empirical version of the optimization problem $\mathfrak{C}_{\mathcal{B}}(\nu)$ at each step. This motivates the quest for alternative strategies, perhaps less generic but better suited to a specific structure. Inspired by combinatorial structures for which computing $\mathfrak{C}_{\mathcal{D}}(\nu)$ is simply not feasible, a relaxation of the generic constrained optimization problem was recently proposed in Cuvelier et al. [2021a]. The authors show that this comes at the price of trading-off regret optimality for computational efficiency. Indeed in some structure, combinatorial properties are at stake and asymptotically optimal algorithms may require solving combinatorial optimization problems (see Cuvelier et al. [2021a]) related to $\mathfrak{C}_{\mathcal{B}}(\nu)$. In order to exploit the combinatorial structures in a numerically efficient way, research has been made in how to relax these combinatorial optimization problems while preserving theoretical properties on the regret of the relaxed algorithms (see Cuvelier et al. [2021b,a]). Our work consider similar computational issues, with a different perspective.

Goal For the structure $\mathcal{B}_q$, as we show in Theorem 1 below, the term $\mathfrak{C}_{\mathcal{B}_q}(\nu)$ unfortunately makes appear in general a combinatorial optimization problem. This makes resorting to OSSB or any strategy targeting exact asymptotic optimality a daunting task for the practitioner. In this paper, our goal is to provide a computationally efficient strategy adapted to the structure $\mathcal{B}_q$, that is able to reach optimality up to controlled error term.

Outline and contributions The rest of this paper is organized as follows. In section 2, we derive a lower bound on the regret for the structured set of bandit configurations $\mathcal{B}_q$. This bound makes appear two components, one that we call *non-combinatorial* as optimizing it can be done efficiently, and a second term that we term *combinatorial* as it involves solving a combinatorial problem. Interestingly, using in Lemma 1 and Theorem 3 that the contribution of the combinatorial part of the lower bound

can be controlled. Owing to this key insight, we introduce in section 3, IMED-EC, an adaptation of the IMED strategy from Honda and Takemura [2015] to the structured set $\mathcal{B}_q$. One advantage of IMED over a KL-UCB alternative is its reduced complexity, which translates to the equivalence class setup. At each time step, the complexity of computing the next arm to be pulled by IMED-EC is no more than the one of sorting a list of $|\mathcal{A}|$ elements once the IMED indexes have been computed, which is only $\log |A|$ times larger than looking for the minimal IMED index. In Section 4, we prove that IMED-EC achieves a controlled asymptotic regret that matches the non-combinatorial part of the lower bound and is at most (less than) a factor of 2 times the optimal regret bound. Last, we illustrate the benefit of the IMED-EC over its unstructured version in section 5, where it shows a substantial improvement. Our experiments also highlights the robustness of the algorithm to a misspecified parameter q, which is a desirable feature for the practitioner.

2 A regret lower bound with combinatorial and non-combinatorial parts

In this section, we derive a lower bound on the number of pulls of suboptimal arms that involves a combinatorial optimization problem. Using that lower bound, we derive a simple algorithm, IMED-EC, that does not involve any optimization problem. While not being asymptotically optimal, we will show in the next section that our algorithm has an upper bound on its regret that is no more than a fraction of the unstructured regret.

The proof of Theorem 1 is based on the concept of **most confusing instance**. Most confusing instances allow to assess the intrinsic difficulty of a bandit problem and allow to compute lower bounds on the number of times suboptimal arms are pulled. The lower bound informs us on the minimal amount of exploration one needs to do to solve a bandit problem. More formally, a confusing instance ν' associated to a suboptimal arm a for a bandit problem ν is a bandit instance with the same set of arms as the original one, but in which μ_a has been changed to $\mu'_a > \mu_*$. An optimal sampling strategy (one that does not sample suboptimal arms too much) should behave differently on the two problems. Studying this difference, we can compute the minimal amount of exploration performed by an optimal strategy on arm a in the original problem ν. Doing so for all suboptimal arms allows to bound the number of samples of suboptimal arms and therefore characterize the intrinsic complexity of a bandit instance ν.

In a structured setting, a confusing instance also has to respect the structure. In our case, it means that a confusing instance cannot have a class with less than q arms. We will therefore consider confusing instances associated to classes rather than individual arms.

Definition 3 (Confusing instance)**.** *Given a bandit configuration $\nu \in \mathcal{B}_q$, a real number λ and a subset $c_q \subseteq \mathcal{A}$ of q equivalent arms in ν, we denote by $\mathcal{B}_q(\nu, c_q, \lambda)$ the set of all bandit configurations having the same set of arms as ν and such that for all $\nu' \in \mathcal{B}_q(\nu, c_q, \lambda)$, $\nu' \in \mathcal{B}_q$ and for every arm a in c_q, $\mu'_a \geqslant \lambda$.*

*When $\lambda > \mu_*$, and c_q is a subset of a suboptimal class, a bandit configuration in $\mathcal{B}_q(\nu, c_q, \lambda)$ is called a **confusing instance** of ν.*

Similarly to the notation introduced above, we will use the notation $\mathcal{B}_q(\mu, c_q, \lambda)$ to specify the set of means of bandit configurations in $\mathcal{B}_q(\mu, c_q, \lambda)$.

The aim of an asymptotic lower bound on the number of pulls of a suboptimal arm is to mathematically understand the minimal asymptotic amount of exploration an algorithm should perform.

Assumption 1: The family $\mathcal{F}$ is such that for all $\kappa \in \mathcal{F}$, $\mu \mapsto \mathcal{K}_{\mathcal{F}}(\kappa\|\mu)$ and $\mu \mapsto \mathcal{K}_{eq}(\kappa\|\mu)$ are continuous, where $\mathcal{K}_{eq}(\kappa, \mu) = \inf_{G\in\mathcal{F}}\{\mathrm{KL}(\kappa, G) : \mathbf{E}_G(X) = \mu\}$ with KL being a notation for the relative entropy or Kullback-Leibler divergence.

Assumption 2: The family $\mathcal{F}$ is an exponential family of dimension 1. Therefore the KL divergences are parameterized by the mean and we may write the KL as a function of the means, $\forall \kappa, \chi \in \mathcal{F}, \mathrm{KL}(\kappa\|\chi) = \mathrm{KL}(\mathbb{E}_{X\sim\kappa}(X)\,\|\mathbb{E}_{X\sim\chi}(X))$ (identification of the KL with its parameterization by the means).

Theorem 1 (Asymptotic lower bound)**.** *Let $q \in \mathbb{N}_*$ be a positive integer and $\nu \in \mathcal{B}_q$ be a bandit configuration having the q-equivalence property. Let $c \subset \mathcal{A}$ be a suboptimal equivalence class in ν.*

Assuming uniform consistency, for all suboptimal arms a,

$$\forall \alpha > 0, \ \lim_{T \to +\infty} \mathbb{E}\left(\frac{N_a(T)}{T^\alpha}\right) = 0,$$

assuming assumption 1, we have the following asymptotic bandit dependent lower bound on the number of pulls of arms in c:

$$\liminf_{T \to \infty} \frac{\min\limits_{c_q \subseteq c} \sum\limits_{a \in c_q} \mathbb{E}_\nu(N_a(T))\, \mathcal{K}_{\mathcal{F}}\left(\nu_a \| \mu_*\right) + \inf\limits_{\mu' \in \mathcal{B}_q(\mu, c_q, \mu_*)} \sum\limits_{a \notin c_q} \mathbb{E}_\nu(N_a(T))\, \mathcal{K}_{eq}(\nu_a \| \mu'_a)}{\log T} \geqslant 1\,, \quad (2)$$

where c_q is any subset of c having q distincts arms within it.

We briefly sketch how confusing instances are used in the proof of Theorem 1. We consider confusing instances in which q arms from a suboptimal class c are moved above the optimal one (*w.r.t.* the mean). If there are q arms in the class, then there are no remaining arms to move. If there are more than $2q$ arms, then moving q arms creates a reminder of size larger than q meaning that the crafted confusing instance respects the equivalence structure. However, if there are between $q+1$ and $2q-1$ arms, then the reminder is of size larger than 1 but strictly smaller than q. The created confusing instance does not respect the equivalence structure and we have to deal with the arms in the reminder (the *infimum* of equation (2)). There are $|c|$ choose q possible choices to move q arms from class c (the *minimum* of equation (2)). All in all, the lower bound involves a combinatorial optimization problem.

While this lower bound involves a combinatorial optimization term, one can distinguish between two regimes depending on the size of the suboptimal class. The *combinatorial regime* and the *non combinatorial regime*.

Non-combinatorial regime For a suboptimal class c, if $|c| = q$ or $|c| \geqslant 2q$, then the lower bound reduces to

$$\liminf_{T \to \infty} \frac{\min\limits_{c_q \subseteq c} \sum\limits_{a \in c_q} \mathbb{E}_\nu\left(N_a(T)\right) \mathcal{K}_{\mathcal{F}}\left(\nu_a \| \mu_*\right)}{\log T} \geqslant 1,$$

because the reminder is of size larger than q and the *infimum* from Theorem 1 disappears. Indeed, the *infimum* is always 0 as this quantity can be obtained by choosing $\mu'_a = \mu_a$ for all $a \in c \setminus c_q$. Furthermore, the minimum over all q-partitions of c is in fact the sum of the q smallest elements of $\left\{\mathbb{E}_\nu\left(N_a(T)\right) \mathcal{K}_{\mathcal{F}}\left(\nu_a \| \mu_*\right)\right\}_{a \in c}$. The search amongst all the q-partitions of c amounts to a research of the q smallest elements which is not more complex than sorting a list of $|c|$ elements. Hence, the problem is no more a combinatorial optimization one and we call this case the *non-combinatorial* regime.

Lemma 1. *Let $\nu \in \mathcal{B}_q$ be a bandit configuration having the q-equivalence property. Let c be a suboptimal class in the non-combinatorial regime, then, under assumption 1 and 2,*

$$\liminf_{T \to \infty} \frac{\sum\limits_{a \in c} \mathbb{E}_\nu\left(N_a(T)\right)}{\log T} \geqslant \frac{|c|}{q} \frac{1}{\mathcal{K}_{\mathcal{F}}\left(\nu_a \| \lambda\right)}. \quad (3)$$

While we do not have information about individual number of times an arm in a class has been sampled, Lemma 1 roughly tells us than on average, the lower bound on the minimal amount of exploration of an arm in a suboptimal class has been divided by q.

Lemma 2. *If all suboptimal classes are in the non-combinatorial regime, under assumption 1 and 2, the regret may be asymptotically lower bounded by*

$$\liminf_{T \to \infty} \frac{\mathcal{R}\left(\nu, T\right)}{\log T} \geqslant \frac{1}{q} \sum_{a \in \mathcal{A} \setminus \mathcal{A}_*} \frac{\mu_* - \mu_a}{\mathcal{K}_{\mathcal{F}}\left(\nu_a \| \lambda\right)}. \quad (4)$$

Lemma 2 informs us that in the non-combinatorial regime, the classical lower bound on the regret given by equation (1) has been divided by q.

Combinatorial regime For a suboptimal class c to be in the *combinatorial* regime, we need $q < |c| < 2q$, since the reminder is such that $0 < |c \setminus c_q| < q$ and the infimum in Theorem 1 is not 0. In that case, the lower bound (2) involves a combinatorial optimization problem. Two difficulties arise from the term

$$\inf_{\mu' \in \mathcal{B}_q(\mu, c_q, \lambda)} \sum_{a \notin c_q} \mathbb{E}_\nu \left(N_a(T)\right) \mathcal{K}_{eq} \left(\nu_a \| \mu'_a\right).$$

First, while we could have thought that summing on the reminder $c \setminus c_q$ would be enough, the summand has to be on $a \notin c_q$ as a whole. Indeed, the residual $c \setminus c_q$ may be of size $q-1$ meaning that it might cost less to move an arm from another class to the residual in order to complete it rather than moving all the reminder. Second, while we could have thought that moving elements from one class of ν to another might be enough, the *infimum* has to be taken on $\mathcal{B}_q(\mu, c_q, \lambda)$. Indeed, the residual $c \setminus c_q$ may be of size $q-1$ and the *nearest* class might be of size exactly q. In this case, it may cost less to move all the $2q-1$ distributions in between the two classes and create a new one rather than merging one of the two with the other.

Lemma 3. *Let $\nu \in \mathcal{B}_q$ be a bandit configuration having the q-equivalence property and c be a suboptimal class in the combinatorial regime. Then, under assumptions 1 and 2,*

$$\liminf_{T\to\infty} \frac{\sum_{a\in c} \mathbb{E}_\nu \left(N_a(T)\right)}{\log T} \geqslant \frac{1}{\frac{q}{|c|}\mathcal{K}_{\mathcal{F}} \left(\nu_a \| \mu_*\right) + \frac{|c|-q}{|c|} \min_{\kappa \in \nu} \mathcal{K}_{eq} \left(\nu_a \| \kappa\right)}, \tag{5}$$

$$\liminf_{T\to\infty} \frac{\sum_{a\in c} \mathbb{E}_\nu \left(N_a(T)\right)}{\log T} \geqslant \frac{1}{2q} \sum_{a\in c} \frac{1}{\mathcal{K}_{\mathcal{F}} \left(\nu_a \| \mu_*\right)}. \tag{6}$$

Those equations can be compared to the equation (3) from the non-combinatorial regime. We emphasize the fact that the lower bounds given by equations (5) and (6) are not the *largest* possible lower bound and hence do not provide as much information about the algorithmically achievable regret as the largest one given by equation (2). However, together with a regret upper bound on the algorithm `IMED-EC`, those quantities will help us control the asymptotic discrepancy between `IMED-EC`'s regret and the asymptotic lower bound given by Theorem 1.

3 Information Minimization for bandits with equivalence class

The algorithm we present, `IMED-EC`, depends on the (*weak*) indexes introduced in the `IMED` paper by Honda and Takemura [2015]. At each time step t, for each arm $a \in \mathcal{A}$, we can compute its IMED index as

$$I_a(t) = N_a(t) \mathcal{K}_{\mathcal{F}} \left(\widehat{\mu}_a(t) \| \widehat{\mu}^*(t)\right) + \log N_a(t),$$

where $\widehat{\mu}^*(t) = \max_{a\in\mathcal{A}} \widehat{\mu}_a(t)$ and for each arm $a \in \mathcal{A}$, $\widehat{\mu}_a(t)$ is the empirical mean of arm a computed with samples from this arm collected up to time t, $\widehat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} X_s \mathbb{1}\left\{a_s = a\right\}$, where X_s is the sample collected by the algorithm at time step s. Let $\nu \in \mathcal{B}_q$ be a bandit configuration having the q-equivalence property. We denote by $\mathcal{A}_*(t) = \arg\max_{a\in\mathcal{A}} \left\{\widehat{\mu}_a(t)\right\}$ the set of empirical optimal arms at time t. We will denote by $\mathcal{A}_q(t)$ the set of arms having the q smallest IMED indexes (breaking ties randomly so that this set has size q). We will also consider the two following quantities for each time t:

$$I^*(t) = \min_{a\in\mathcal{A}_*(t)} I_a(t) = \min_{a\in\mathcal{A}_*(t)} \log N_a(t),$$

$$I(t) = \min_{\substack{\mathcal{A}'\subset\mathcal{A} \\ |\mathcal{A}'|=q}} \sum_{a'\in\mathcal{A}'} I_{a'}(t) = \sum_{a'\in\mathcal{A}_q(t)} I_a(t).$$

$I(t)$ can be computed efficiently by summing the q smallest elements of the list of IMED indexes. Finding the q smallest elements can be done by maintaining a sorted array of IMED indexes while computing them. The procedure costs a constant factor of $\log|\mathcal{A}|$. Computing $I(t)$ therefore costs $\mathcal{O}\left(|\mathcal{A}| \log|A|\right)$, which is only $\log|A|$ times larger than looking for the minimal IMED index. Computing $|\mathcal{A}_*(t)|$ can be done by maintaining a set of arms having the best empirical mean (adds a constant factor). The `IMED-EC` algorithm is presented in Algorithm 1.

Algorithm 1 IMED-EC (IMED for Equivalent Classes)

Pull each arm once
for $t = |\mathcal{A}| \dots T-1$ **do**
  if $I^*(t) \leqslant I(t)$ **then**
    Pull $a_{t+1} \in \arg\min_{a \in \mathcal{A}_*(t)} N_a(t)$ (chosen arbitrarily)
  else
    Pull $a_{t+1} \in \arg\min_{a \notin \mathcal{A}_*(t)} I_a(t)$ (chosen arbitrarily)
  end if
end for

While the orginal problem involves combinatorial quantities, those are not involved in the IMED-EC algorithm. From a time complexity viewpoint, this makes this algorithm on par with other popular algorithms such as UCB, KLUCB, and IMED algorithm. On the contrary, the general structure algorithm OSSB involves solving a combinatorial optimization problem at each time step, which makes it numerically inefficient. We are not aware of any general relaxation method for this algorithm that we could compare IMED-EC with. It is interesting to note that in the case where $q = 1$, the IMED-EC algorithms coincide with the IMED algorithm.

Intuition For an arm a, $N_a(t)\mathcal{K}_{\mathcal{F}}\left(\widehat{\mu}_a(t) \| \widehat{\mu}^*(t)\right)$ may be interpreted as the opposite of a *log-likelihood of optimality* of that arm. $\log N_a(t)$ is linked to the log-frequency of play of that arm, the frequency of play of an arm being interpreted as the probability of pulling that arm is a sequence of length t. The IMED algorithm thus can be intuitively understood as an algorithm matching an empirical log-probability with a log-frequency of play. In our setting, there is at least q elements in each group. It therefore makes sense to test for the optimality of a group rather single elements. Since all arms are independent, it makes sense to sum the *log-likelihood of optimality* on all the q-partitions of the set of arms. Since we have the intuition that this first part is the logarithm of a product of probability, we may compare it to the product of the frequencies. Therefore, we get that important quantities are the sum of IMED indexes for each q partition of the arms, seen as a comparison between the optimality of this group of q elements and the associated frequency of play of that group. The minimal IMED index is the one whose frequency of play is the lowest compared to its *likelihood of optimality*, similarly for the sum of IMED indexes. Other intuitions regarding the fairness (frequency of pulls within the same class) of the algorithm are given in appendix D.

4 Regret analysis

In this section, we now detail the main bound on the regret of IMED-EC.

Theorem 2 (Upper bound on the number of pulls)**.** *Under the IMED-EC algorithms, under assumption 1 and 2, the number of pulls of a suboptimal arm a is upper bounded by:*

$$\mathbb{E}_\nu\left(N_a(T)\right) \leqslant \frac{\log T}{q\mathcal{K}_{\mathcal{F}}\left(\nu_a \| \mu_*\right)}\left(1 + \alpha(\varepsilon)\right) + f(\varepsilon), \tag{7}$$

where $0 < \varepsilon < \frac{1}{3}\min_{a \in \mathcal{A}\setminus\mathcal{A}_*}\left(\mu_* - \mu_a\right)$, f is function that depends on concentration properties on $\mathcal{F}$, and α tends to 0 as ε tends to 0.

Remark: α and f functions are mostly used for deriving theoretical guarantees in IMED-EC regret analysis. α is controlled thanks to property 2 as in the paper of Honda and Takemura [2015] for IMED regret analysis. A finite sample analysis can be derived from a careful analysis of the term f. Being more precise requires scrutinizing the properties of the considered family.

Corollary 1. *Under the IMED-EC algorithms, under assumptions 1 and 2, the number of pulls of a suboptimal arm a is upper bounded by:*

$$\min_{c_q \subseteq c} \sum_{a \in c_q} \mathbb{E}_\nu\left(N_a(T)\right)\mathcal{K}_{\mathcal{F}}\left(\nu_a \| \mu_*\right) \leqslant \left(1 + \alpha\left(\varepsilon\right)\right)\log T + g(\varepsilon). \tag{8}$$

where $0 < \varepsilon < \frac{1}{3}\min_{a \in \mathcal{A}\setminus\mathcal{A}_*}\left(\mu_* - \mu_a\right)$ α and α tends to 0 as ε tends to 0.

Theorem 3 (Asymptotic upper bound on the number of pulls)**.** *Under the* `IMED-EC` *algorithms, under assumption 1 and 2, the number of pulls of a suboptimal arm a is asymptotically upper bounded by:*

$$\liminf_{t\to+\infty} \frac{\mathbb{E}_\nu\left(N_a(T)\right)}{\log T} \leqslant \frac{1}{q\mathcal{K}_{\mathcal{F}}\left(\nu_a\|\mu_*\right)}. \tag{9}$$

Discussion This upper bound shows that in particular, the number of pulls of a suboptimal class, $\sum_{a\in c}\mathbb{E}_\nu\left(N_a(T)\right)$ is asymptotically no more than $\frac{|c|}{q\mathcal{K}_{\mathcal{F}}(\nu_a\|\mu_*)}\log T$. This hence matches the lower bound in the *non-combinatorial regime*. In the *combinatorial regime*, along with equation (6), this regret upper bound shows that

$$\frac{|c|}{q\mathcal{K}_{\mathcal{F}}\left(\nu_a\|\mu_*\right)} \geqslant \liminf_{T\to\infty}\sum_{a\in c}\frac{\mathbb{E}_\nu\left(N_a(T)\right)}{\log T} \geqslant \frac{1}{2}\cdot\frac{|c|}{q\mathcal{K}_{\mathcal{F}}\left(\nu_a\|\mu_*\right)},$$

proving that the regret of the proposed `IMED-EC` does not differ from the optimal lower bound by a factor more than 2. This is a striking result. Equation (6) can be used to have an even more precise control on the discrepancy to the optimal regret bound, as it shows the factor 2 can be actually replaced with $1+\frac{|c|-q}{q}\frac{\min_{\kappa\in\nu}\mathcal{K}_{eq}(\nu_a\|\kappa)}{\mathcal{K}_{\mathcal{F}}(\nu_a\|\mu_*)}$. Since the factor $\frac{|c|-q}{q}\frac{\min_{\kappa\in\nu}\mathcal{K}_{eq}(\nu_a\|\kappa)}{\mathcal{K}_{\mathcal{F}}(\nu_a\|\mu_*)}$ is strictly between 0 and 1 in the combinatorial regime that we are studying, the discrepancy between the lower bound and the regret of `IMED-EC` indeed is always bounded by 2. On the other hand, this refined error measurement is problem dependant while the factor of 2 is universal.

We provide the full proof of Theorem 3 and Theorem 2 in appendix C where we also discuss how to weaken assumption 2 and still get the result of theorem 2.

5 Experiments

In this section, we support our theoretical analysis by conducting three sets of experiments. The Python code used to perform those experiments is available on Github[2]. We support our empirical evidences using plots of cumulative regrets. In this section, all the experiments are conducted using gaussian distributions whose means are between 0 and 1 and of unit standard deviation. Those graphs are representative of all the experiments that we conducted and more plots and experiments may be found in the appendix D.

Balanced class, perfect knowledge In this set of experiments, see Figure 1, we focus on the bandit configurations in which all equivalence classes have the same cardinality and assume that we know the number of elements per class. This setting is interesting for two reasons. First, one can compute the theoretical lowerbound without solving a combinatorial optimization problem. Second, the theoretical analysis shows that `IMED-EC` is asymptotically optimal in this case. This setting will thus allow us to numerically grasp what happens in the most structured case. We compare `IMED-EC` to unspecialized bandit algorithm, `UCB`, `IMED` and `KLUCB`. To make the comparison fairer we also compare `IMED-EC` to `OSSB`, an algorithm specialized in structured bandit. Since `OSSB` has to solve a combinatorial optimization problem at each time step, we cannot carry experiments on large sets of arms while comparing `IMED-EC` to it. In this particular setting, we see that while `OSSB` and `IMED-EC` are provably asymptotically optimal, `IMED-EC` numerically performs better in finite time horizon. We recall that it is furthermore numerically more efficient since it does not involve any combinatorial optimization. Without too much surprises, `IMED-EC` also outperforms unspecialized algorithm.

Imperfect knowledge In the experiment plotted Figure 2, we leverage the knowledge hypothesis and assume that we only know a lower bound on the number of elements per class while the classes are still balanced. We compare `IMED-EC` to unspecialized bandit algorithm, `IMED` and `KLUCB`. We drop `OSSB` from our test bed due to the computational burden of solving a combinatorial optimization problem at each time step. We can see that the finite time cumulative regret of `IMED-EC` indeed is much smaller than the regret of the unspecialized algorithms.

Influence of the parameter q Here we show the numerical robustness of `IMED-EC` with respect to the lower bound parameter q on the number of elements per classes. On the same bandit problem,

[2]https://github.com/fabienpesquerel/stochastic-bandits-with-groups-of-similar-arms-neurips-2021

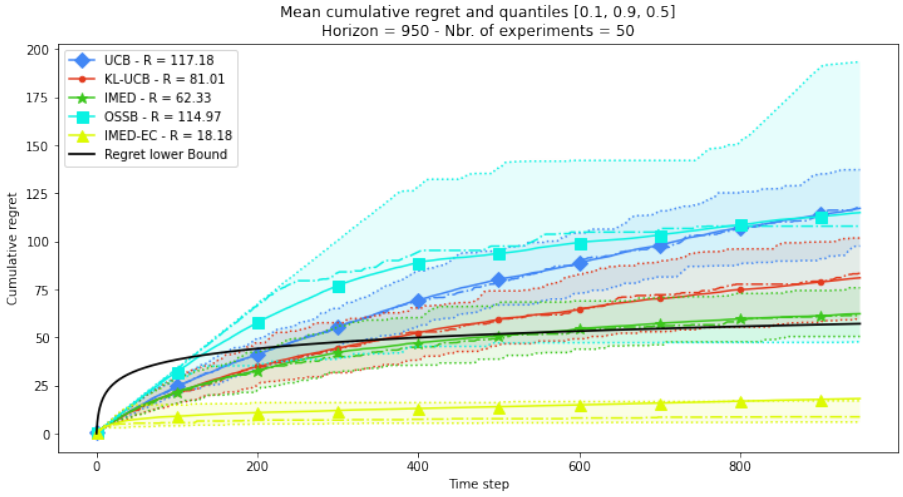

Figure 1: 3 classes, 3 distributions per class - set of means = $\{0.3, 0.5, 0.9\}$

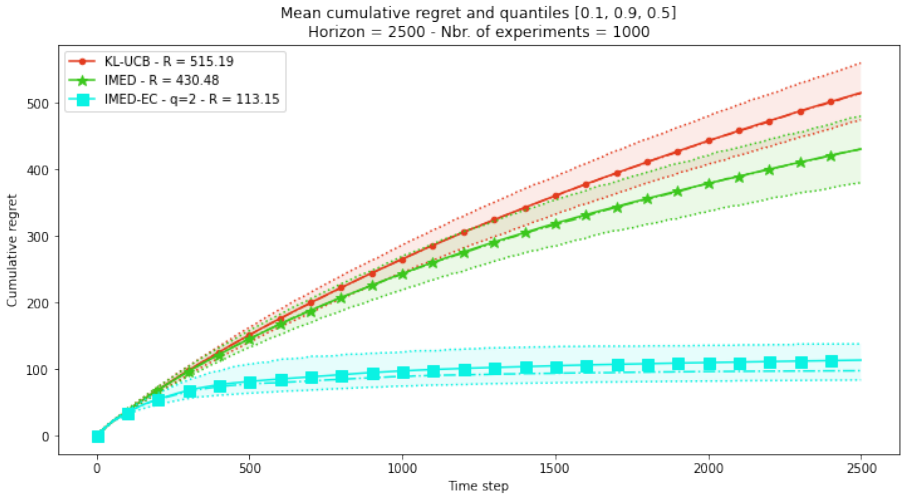

Figure 2: 7 classes, 8 distributions per class - set of means = $\{0.1, 0.3, 0.4, 0.5, 0.6, 0.75, 0.9\}$

we compare different instances of `IMED-EC` where different values of q are used. In the legend, *opt.* stands for optimal and corresponds to the largest valid lower bound on the number of elements per class, *i.e.* the minimal number of elements in a class. The experiment Figure 3 is performed on a bandit problem with 7 classes and an uneven number of distributions per class. The smallest class has 4 elements and the largest 23. While q increases up to the minimum cardinality of a class, we see that the performances of `IMED-EC` increases. It is rather remarkable that once we go beyond that theoretical threshold, the performances of `IMED-EC` do not deteriorate. We even found it difficult to find settings to deteriorate them at all. While the expected regret does not seem to deteriorate, we sometimes see that the tails of the regret widen as it can be seen on the plot Figure 3 for $q = 7$ and $q = 20$ since the 0.9 quantile curves are so large for those values of q. We interpret part of this robustness to the fact that the relaxation induced in `IMED-EC` makes the algorithm over explore compared to what the true lower bound suggests. Increasing q reduces the exploration and therefore may improve the performances of the algorithm. However, this robustness is observed even in the case where the classes are balanced. This interpretation thus does not explain everything about the numerical robustness of `IMED-EC`. This type of experiment does not take more than roughly 10 to 15 minutes on a notebook run in Google Colab depending on the number of arms, the horizon and the number of runs. This supports the numerical efficiency of the relaxation made in `IMED-EC`.

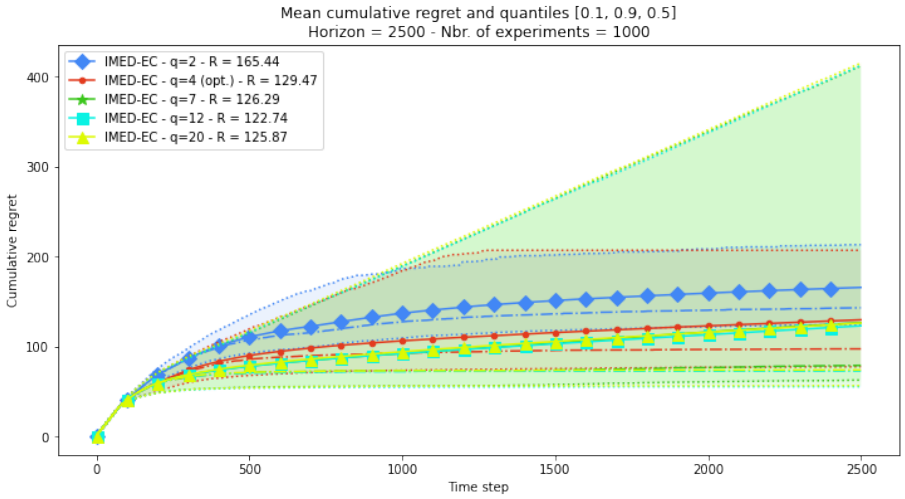

Figure 3: 7 classes, unbalanced - set of means = $\{0.1, 0.3, 0.4, 0.5, 0.6, 0.75, 0.9\}$

6 Conclusion

In this paper, we introduced `IMED-EC`, a numerically efficient algorithm to solve a structured bandit problem for which we derived a lower bound involving a combinatorial optimization problem. While not being asymptotically optimal, we proved that the asymptotic regret of `IMED-EC` is always smaller than the unstructured one and that we can control the discrepancy with respect to the structured regret lower bound by a factor of at most 2.

Acknowledgments and Disclosure of Funding

This work has been supported by the French Ministry of Higher Education and Research, Inria, Scool, the French Agence Nationale de la Recherche (ANR) under grant ANR-16-CE40-0002 (the BADASS project) the MEL and the I-Site ULNE regarding project R-PILOTE-19-004-APPRENF. The PhD of Fabien Pesquerel is supported by a grant from École Normale Supérieure.

References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

R. Agrawal. Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27, 1995.

S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47, 2002.

D. Baudry, E. Kaufmann, and O.-A. Maillard. Sub-sampling for Efficient Non-Parametric Bandit Exploration. In *NeurIPS 2020*, Vancouver, Canada, Dec. 2020. URL `https://hal.archives-ouvertes.fr/hal-02977552`.

A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, 2013.

H. P. Chan. The multi-armed bandit problem: An efficient nonparametric solution. *The Annals of Statistics*, 48(1):346–373, 2020.

R. Combes and A. Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *International Conference on Machine Learning*, 2014.

R. Combes, S. Magureanu, and A. Proutiere. Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 1763–1771, 2017.

T. Cuvelier, R. Combes, and E. Gourdin. Asymptotically optimal strategies for combinatorial semi-bandits in polynomial time. In *Algorithmic Learning Theory*, pages 505–528. PMLR, 2021a.

T. Cuvelier, R. Combes, and E. Gourdin. Statistically efficient, polynomial-time algorithms for combinatorial semi-bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 5(1):1–31, 2021b.

R. de Heide, J. Cheshire, P. Ménard, and A. Carpentier. Bandits with many optimal arms. In *Advances in Neural Information Processing Systems*, 2021. URL `https://arxiv.org/abs/2103.12452`.

A. Dembo and O. Zeitouni. Large deviations techniques and applications. *Elearn*, 1998.

A. Durand, O.-A. Maillard, and J. Pineau. Streaming kernel regression with provably adaptive mean, variance, and regularization. *arXiv preprint arXiv:1708.00768*, 2017.

A. Garivier, P. Ménard, and G. Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *arXiv preprint arXiv:1602.07182*, 2016.

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

J. Honda and A. Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85(3):361–391, 2011.

J. Honda and A. Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *Machine Learning*, 16:3721–3756, 2015.

A. Kalvit and A. Zeevi. From finite to countable-armed bandits. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8259–8269. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/5dbc8390f17e019d300d5a162c3ce3bc-Paper.pdf`.

N. Korda, E. Kaufmann, and R. Munos. Thompson Sampling for 1-dimensional Exponential family bandits. In *Advances in Neural Information Processing Systems*, 2013.

B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan. Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, pages 767–776. PMLR, 2015.

B. Kveton, C. Szepesvari, Z. Wen, M. Ghavamzadeh, and T. Lattimore. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *ICML*, 2019.

B. Kveton, M. Zaheer, C. Szepesvari, L. Li, M. Ghavamzadeh, and C. Boutilier. Randomized exploration in generalized linear bandits. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2066–2076. PMLR, 26–28 Aug 2020. URL `http://proceedings.mlr.press/v108/kveton20a.html`.

T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6, 1985.

T. L. Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114, 1987.

T. L. Lai. Boundary crossing problems for sample means. *The Annals of Probability*, pages 375–396, 1988.

T. Lattimore and C. Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737, 2017.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

S. Lu, G. Wang, Y. Hu, and L. Zhang. Optimal algorithms for Lipschitz bandits with heavy-tailed rewards. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4154–4163. PMLR, 09–15 Jun 2019. URL `http://proceedings.mlr.press/v97/lu19c.html`.

S. Magureanu. *Efficient Online Learning under Bandit Feedback*. PhD thesis, KTH Royal Institute of Technology, 2018.

S. Magureanu, R. Combes, and A. Proutiere. Lipschitz bandits: Regret lower bounds and optimal algorithms. *Machine Learning*, 35:1–25, 2014.

O.-A. Maillard. Boundary crossing probabilities for general exponential families. *Mathematical Methods of Statistics*, 27(1):1–31, 2018.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58:527–535, 1952.

H. Saber, P. Ménard, and O.-A. Maillard. Forced-exploration free strategies for unimodal bandits. *arXiv preprint arXiv:2006.16569*, 2020.

N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 1015–1022. Omnipress, 2010.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25, 1933a.

W. 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, 1933b.

W. R. Thompson. On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation. *The Annals of Mathematical Statistics*, 6(4):214–219, 1935.

T. Wang, W. Ye, D. Geng, and C. Rudin. Towards practical lipschitz bandits. *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, Oct 2020. doi: 10.1145/3412815.3416885. URL `http://dx.doi.org/10.1145/3412815.3416885`.

J. Y. Yu and S. Mannor. Unimodal bandits. In *ICML*, 2011.

A Notations and Assumptions

In this section, we first recall a few notations. Then, we provide details about the assumptions made about the distributions. In particular, we make precise the assumption about exponential families we consider, and we also provide an alternative set of assumptions under which our analysis hold. Indeed, our proof techniques naturally apply to the setup considered for the analysis of the IMED strategy. We formalize the corresponding assumptions below in (Assumption 1, 3).

Notations: Given a bandit configuration ν and an arm $a \in \mathcal{A}$, $\widehat{\mu}_a(T) = \frac{1}{N_a(T)} \sum_{t=1}^{T} X_t \mathbb{1}\{a_t = a\}$ where $N_a(T) \sum_{t=1}^{T} \mathbb{1}\{a_t = a\}$ and X_t is the reward obtained at time t. We will denote by $\widehat{\mu}_a^n$ the empirical mean of arm a obtained from n *i.i.d.* samples (and n is not a random variable).

Assumption 1: The family $\mathcal{F}$ is such that for all $\kappa \in \mathcal{F}$, $\mu \mapsto \mathcal{K}_{\mathcal{F}}(\kappa \| \mu)$ and $\mu \mapsto \mathcal{K}_{eq}(\kappa \| \mu)$ are continuous, where $\mathcal{K}_{eq}(\kappa \| \mu) = \inf_{G \in \mathcal{F}} \{\mathrm{KL}(\kappa \| G) : \mathbf{E}_G(X) = \mu\}$ with KL being a notation for the relative entropy or Kullback-Leibler divergence.

This property is also assumed in Honda and Takemura [2015]. This is especially relevant as they consider reward distributions having a semi-bounded support.

Assumption 2: The family $\mathcal{F}$ is an exponential family of dimension 1, *i.e.* admits the canonical decomposition with respect to some measure

$$p_\theta(x) = \exp\left(t(x)\theta - \psi(\theta) + k(x)\right),$$

where $\theta \in \Theta \subseteq \mathbb{R}$ is a parameter, k a real function, t is called the sufficient statistics and ψ, the log-partition function, is assumed to be twice differentiable. We assume that Θ is open and non-empty. We further assume that on Θ, the second derivative, ψ'' , of the log-partition function is bounded. Formally, there exists $M_{\psi,\Theta}$ such that:

$$\sup_{\theta \in \Theta} \psi''(\theta) \leqslant M_{\psi,\Theta}. \tag{10}$$

If p_θ and $p_{\theta'}$ are two distributions in $\mathcal{F}$, then:

$$\mathrm{KL}(p_\theta \| p_{\theta'}) = \psi(\theta) - \psi(\theta') - (\theta - \theta')\psi'(\theta').$$

Because $\mathcal{F}$ is an exponential family of dimension 1, $\psi'' > 0$ and ψ' is strictly increasing. This implies that there is a continuous bijection between the parameter space Θ and the space of expected values that can be taken between distributions in $\mathcal{F}$. Specifically, $\mu : \theta \mapsto \mathbb{E}_{X \sim p_\theta}(X)$ is a bijection on its co-domain, $\mu(\Theta)$. Therefore the KL divergences can be parameterized by the mean and we may write the KL as a function of the means, $\forall \kappa, \chi \in \mathcal{F}, \mathrm{KL}(\kappa \| \chi) = d\left(\mathbb{E}_{X \sim \kappa}(X) \| \mathbb{E}_{X \sim \chi}(X)\right)$.

Therefore, under assumption 2 $\mathcal{K}_{\mathcal{F}}$ identifies with the KL and assumption 1 holds by twice differentiability of ψ since it implies the continuity of ψ and ψ'. Note that exponential families of dimension 1 have been considered in several other works, see e.g. Korda et al. [2013], Cappé et al. [2013], or Maillard [2018] and Lai [1988] where restriction on ψ'' is also considered.

Assumption 3: The family $\mathcal{F}$ is the set of distributions with semi-bounded rewards, *i.e.* whose supports lies in $(-\infty, 1]$. Furthermore, the moment generating function $\mathbb{E}(\exp(\lambda X))$ of any distribution $X \in \mathcal{F}$ exists in a neighborhood of $\lambda = 0$. For technical reasons linked to the moment generating function, we also assume that the maximal expected rewards of the distributions in $\mathcal{F}$ is strictly smaller than 1. Those assumptions on $\mathcal{F}$ are the same than the one of the paper Honda and Takemura [2015].

Under assumption 2, the moment generating function can be written as a function of ψ,

$$\mathbb{E}_{X \sim p_\theta}(\exp(\lambda X)) = \exp\left(\psi(\theta + \lambda) - \psi(\theta)\right),$$

which is well defined as long as $\theta + \lambda \in \Theta$. This is the reason assumption 1 consider Θ to be *open* since in this case, θ can never be too close to the boundary, *i.e.* there always exists λ small enough such that $\theta + \lambda \in \Theta$, and the moment generating function always exists in a neighborhood of 0.

Remark: The regret lower bound can be proved under assumption 2 or alternatively under assumption 1 and 3. Most of the distributions studied in the bandit literature fall under one of those sets of assumptions. Assumption 2 includes for instance Gaussian bandits (known variance) and Bernoulli distributions. For Bernoulli distributions, the set of considered means should be of the form $(\varepsilon, 1-\delta)$ with $0 < \varepsilon < 1 - \delta < 1$ in order to upper bound the second derivative of the log-partition function. This situation is represented in Figure 4. Now for Gaussian distributions with known variance, (10) holds trivially.

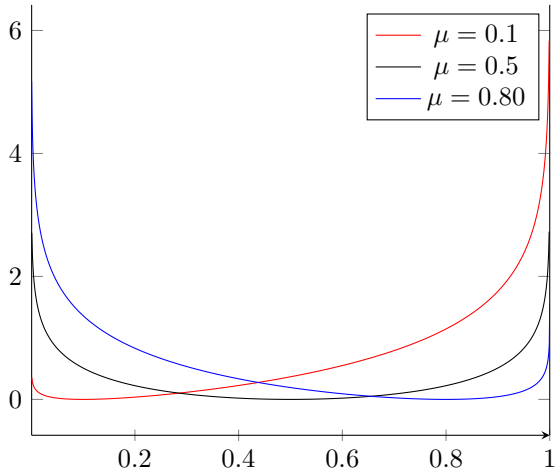

Figure 4: Plot of $x \mapsto \mathrm{kl}(\mu \| x) = x \log \frac{x}{\mu} + (1-x) \frac{1-x}{1-\mu}$ (Bernoulli)

Property 1 (Monotonicity of the KL)**.** *Under assumption 2, if we assume that,*

$$a \leqslant \mu + \varepsilon < \mu_* - \varepsilon \leqslant b,$$

then,

$$d(a\|b) \geqslant d(\mu + \varepsilon \| \mu_* - \varepsilon).$$

The paper of Honda and Takemura [2015] features a similar property for the $\mathcal{K}_{\mathcal{F}}$ under assumption 1 and assumption 3.

Property 2 (Upper approximation of the KL)**.** *Let ε be such that $\mu + \varepsilon < x - \varepsilon$. Under assumption 2, there exists α such that*

$$\frac{1}{d(\mu + \varepsilon \| x - \varepsilon)} \leqslant \frac{1}{d(\mu \| x)} (1 + \alpha(\varepsilon)), \tag{11}$$

with $\alpha(\varepsilon) \to 0$ as $\varepsilon \to 0$.

The paper of Honda and Takemura [2015] features a similar property for the $\mathcal{K}_{\mathcal{F}}$ under assumption 1 and assumption 3.

Proposition 1. *Let $\mathcal{F}$ be an exponential family $\mathcal{F}$ of dimension 1 satisfying the hypothesis of assumption 2. Let $X \sim p_{\theta_*} \in \mathcal{F}$ be a real random variable. Denote by θ_* its true natural parameter and $\widehat{\theta}_n$ it empirical parameter computed with n i.i.d. samples from the distribution of X. Let ψ be the log-partition function of the family $\mathcal{F}$. Let θ be the parameter corresponding to the distribution of $\mathcal{F}$ with mean $\mu < \mu_*$ and θ_* be the parameter corresponding to the disribution of $\mathcal{F}$ with mean μ_*, then*

$$\begin{aligned}\mathbb{P}\left(\mathrm{KL}\left(p_{\widehat{\theta}_n} \| p_\theta\right) \geqslant u, \widehat{\mu}_n \leqslant \mu\right) &\leqslant \mathbb{P}\left(\mathrm{KL}\left(p_{\widehat{\theta}_n} \| p_{\theta_*}\right) \geqslant u + \alpha_\psi(\theta_*, \theta)\sqrt{u}\right) \\ &\leqslant \exp\left(-n\left(u + \alpha_\psi(\theta_*, \theta)\sqrt{u}\right)\right),\end{aligned}$$

where $\alpha_\psi(\theta_*, \theta) > 0$.

Proof. For distributions in an exponential family, the Kullback-Leibler divergence can be written using ψ'', and therefore:

$$\mathrm{KL}\left(p_{\widehat{\theta}_n} \| p_\theta\right) = \frac{1}{2}\left(\widehat{\theta}_n - \theta\right)^2 \psi''\left(\gamma\widehat{\theta}_n + (1-\gamma)\theta\right) \geqslant u.$$

Under assumption 2, ψ'' is uniformly bounded, $\psi'' \leqslant M$, therefore:

$$\left|\theta - \widehat{\theta}_n\right| \geqslant \sqrt{\frac{2u}{M}}.$$

Without loss of generality, we can assume that

$$\theta - \widehat{\theta}_n \geqslant \sqrt{\frac{2u}{M}}. \tag{12}$$

because the following derivation will feature the term $\langle \theta - \widehat{\theta}_n | \psi'(\theta_*) - \psi'(\theta) \rangle$. This term is always positive because $\widehat{\mu}_n \leqslant \mu < \mu_*$ and $\mu : \theta \mapsto \mathbb{E}_{X \sim p_\theta}(X)$ is a continuous bijection. Therefore it has to be strictly monotonous. Since ψ' is strictly increasing (because ψ is strictly convex), it preserves the order. Therefore, $\theta - \widehat{\theta}_n$ and $\psi'(\theta_*) - \psi'(\theta)$ have the same sign.

Since $\mathrm{KL}(p_\theta \| p_{\theta'}) = \psi(\theta) - \psi(\theta) - (\theta - \theta')\psi'(\theta')$, we can derive the following sequence of implications:

$$\begin{aligned}
&\Leftrightarrow \mathrm{KL}\left(p_{\widehat{\theta}_n} \| p_\theta\right) \geqslant u \\
&\Leftrightarrow \mathrm{KL}\left(p_{\widehat{\theta}_n} \| p_\theta\right) + \mathrm{KL}\left(p_{\widehat{\theta}_n} \| p_{\theta_*}\right) \geqslant u + \mathrm{KL}\left(p_{\widehat{\theta}_n} \| p_{\theta_*}\right) \\
&\Leftrightarrow \mathrm{KL}\left(p_{\widehat{\theta}_n} \| p_{\theta_*}\right) \geqslant u + \underbrace{\mathrm{KL}\left(p_{\widehat{\theta}_n} \| p_{\theta_*}\right) - \mathrm{KL}\left(p_{\widehat{\theta}_n} \| p_\theta\right)}_{>0} \\
&\Leftrightarrow \mathrm{KL}\left(p_{\widehat{\theta}_n} \| p_{\theta_*}\right) \geqslant u + \mathrm{KL}(p_\theta \| p_{\theta_*}) + \langle \theta - \widehat{\theta}_n | \underbrace{\psi'(\theta_*) - \psi'(\theta)}_{>0 \text{ because } \psi \text{ is strictly convex}} \rangle \\
&\Leftrightarrow \mathrm{KL}\left(p_{\widehat{\theta}_n} \| p_{\theta_*}\right) \geqslant u + \mathrm{KL}(p_\theta \| p_{\theta_*}) + \langle \underbrace{\theta - \widehat{\theta}_n}_{\geqslant \sqrt{2u/M} \text{ by (12)}} | g(\theta_*, \theta) \rangle \\
&\Rightarrow \mathrm{KL}\left(p_{\widehat{\theta}_n} \| p_{\theta_*}\right) \geqslant u + \mathrm{KL}(p_\theta \| p_{\theta_*}) + \langle \sqrt{\frac{2u}{M}} | g(\theta_*, \theta) \rangle \\
&\Rightarrow \mathrm{KL}\left(p_{\widehat{\theta}_n} \| p_{\theta_*}\right) \geqslant u + \alpha_\psi(\theta_*, \theta)\sqrt{u}.
\end{aligned}$$

The upper bound on the probability is a classical result about exponential families. □

B Proof of the regret lower bound

In this section we prove Theorem 1 that we remind below.

Theorem (Asymptotic lower bound)**.** *Let $q \in \mathbb{N}_*$ be a positive integer and $\nu \in \mathcal{B}_q$ be a bandit configuration having the q-equivalence property. Let $c \subset A$ be a suboptimal equivalence class in ν. Assuming uniform consistency, for all suboptimal arm a,*

$$\forall \alpha > 0, \ \lim_{T \to +\infty} \mathbb{E}\left(\frac{N_a(T)}{T^\alpha}\right) = 0,$$

assuming assumption 1, we have the following asymptotic bandit dependent lower bound on the number of pulls of arms in c:

$$\liminf_{T \to \infty} \frac{\min\limits_{c_q \subseteq c} \sum\limits_{a \in c_q} \mathbb{E}_\nu(N_a(T)) \mathcal{K}_{\mathcal{F}}(\nu_a \| \mu_*) + \inf\limits_{\mu' \in \mathcal{B}_q(\mu, c_q, \lambda)} \sum\limits_{a \notin c_q} \mathbb{E}_\nu(N_a(T)) \mathcal{K}_{eq}(\nu_a \| \mu'_a)}{\log T} \geqslant 1,$$

where c_q is any subset of c having q distinct arms within it.

The proof is standard and makes use of the notion of *most confusing instance* specialized for this structure in the main part of this paper.

Proof of Theorem 1. Let $I_T = (a_k, X_{a_k})_{1 \leqslant k \leqslant T}$ denote the history of actions and rewards taken by a sequential decision maker algorithm up to time T. Then, using the data processing inequality, it is proved in Garivier et al. [2016] that

$$\sum_{a \in \mathcal{A}} \mathbb{E}_\nu \left(N_a(T) \right) \mathrm{KL} \left(\nu_a \| \nu'_a \right) \geqslant \mathrm{kl} \left(\mathbb{E}_\nu \left(Z \right) \| \mathbb{E}_{\nu'} \left(Z \right) \right),$$

for $Z \in (0,1)$ a $\sigma(I_T)$-measurable random variable and kl the Kullback-Leibler divergence between Bernoulli distributions. Let c be suboptimal class and $c_q \subseteq c$ be subset of q elements in c. Applying the previous inequality for $Z = N_c(T)/T$, for all $\lambda > \mu_*$ and $\nu' \in \mathcal{B}_q \left(\nu, c_q, \lambda \right)$ we have that:

$$\begin{aligned}
\sum_{a \in c} \mathbb{E}_\nu \left(N_a(T) \right) \mathrm{KL} \left(\nu_a \| \nu'_a \right) &= \sum_{a \in c_q} \mathbb{E}_\nu \left(N_a(T) \right) \mathrm{KL} \left(\nu_a \| \nu'_a \right) + \sum_{a \notin c_q} \mathbb{E}_\nu \left(N_a(T) \right) \mathrm{KL} \left(\nu_a \| \nu'_a \right) \\
&\geqslant kl \left(\mathbb{E}_\nu \left(\frac{N_c(T)}{T} \right) \| \mathbb{E}_{\nu'} \left(\frac{N_c(T)}{T} \right) \right) \\
&\geqslant \left(1 - \mathbb{E}_\nu \left(\frac{N_c(T)}{T} \right) \right) \log \frac{1}{1 - \mathbb{E}_{\nu'} \left(\frac{N_c(T)}{T} \right)} - \log 2 \\
&= \left(1 - \mathbb{E}_\nu \left(\frac{N_c(T)}{T} \right) \right) \log \frac{T}{T - \mathbb{E}_{\nu'} \left(N_c(T) \right)} - \log 2.
\end{aligned}$$

Since all arms that are not in c are suboptimal for ν', the uniform consistency hypothesis implies that

$$\forall 0 < \alpha \leqslant 1, \qquad 0 \leqslant T - \mathbb{E}_{\nu'} \left(N_c(T) \right) = o \left(T^\alpha \right);$$

and therefore, $T - \mathbb{E}_{\nu'} \left(N_c(T) \right) \leqslant T^\alpha$ for T large enough. We deduce that, for all $0 < \alpha \leqslant 1$,

$$\liminf_{T \to +\infty} \frac{1}{\log T} \log \frac{T}{T - \mathbb{E}_{\nu'} \left(N_c(T) \right)} \geqslant \liminf_{T \to +\infty} \frac{1}{\log T} \log \frac{T}{T^\alpha} = 1 - \alpha.$$

Since all arms within the class c are suboptimal for ν and the considered strategy is assumed to satisfy the uniform consistency hypothesis, $\mathbb{E}_\nu \left(\frac{N_c(T)}{T} \right) \to 0$ as $T \to +\infty$. Together, and letting α be arbitrarily close to 0, these facts implies that

$$\liminf_{T \to \infty} \frac{\sum_{a \in c_q} \mathbb{E}_\nu \left(N_a(t) \right) \mathrm{KL} \left(\nu_a \| \nu'_a \right) + \sum_{a \notin c_q} \mathbb{E}_\nu \left(N_a(t) \right) \mathrm{KL} \left(\nu_a \| \nu'_a \right)}{\log T} \geqslant 1.$$

For each c_q, we can minimize this quantity over all confusing instances $\nu' \in \mathcal{B}_q \left(\nu, c_q, \lambda \right)$ (with a 0 lower bound if the set is empty), and use the continuity of the KL (assumption 1) to let $\lambda > \mu_*$ tends toward μ_*,

$$\liminf_{T \to \infty} \frac{\sum\limits_{a \in c_q} \mathbb{E}_\nu \left(N_a(T) \right) \mathcal{K}_{\mathcal{F}} \left(\nu_a \| \mu_* \right) + \inf\limits_{\mu' \in \mathcal{B}_q(\mu, c_q, \lambda)} \sum\limits_{a \notin c_q} \mathbb{E}_\nu \left(N_a(T) \right) \mathcal{K}_{eq} \left(\nu_a \| \mu'_a \right)}{\log T} \geqslant 1,$$

where each KL can be minimized independently once $\mu' \in \mathcal{B}_q \left(\mu, c_q, \lambda \right)$ has been set, owing to the considered structure. Since this lower bound is valid for all $c_q \in c$ suboptimal, it is valid for the minimal quantity over all q partitions,

$$\liminf_{T \to \infty} \frac{\min\limits_{c_q \subseteq c} \sum\limits_{a \in c_q} \mathbb{E}_\nu \left(N_a(T) \right) \mathcal{K}_{\mathcal{F}} \left(\nu_a \| \mu_* \right) + \inf\limits_{\mu' \in \mathcal{B}_q(\mu, c_q, \lambda)} \sum\limits_{a \notin c_q} \mathbb{E}_\nu \left(N_a(T) \right) \mathcal{K}_{eq} \left(\nu_a \| \mu'_a \right)}{\log T} \geqslant 1,$$

which proves the Theorem 1. □

C Proof of the regret upper bound

In this section, we prove the regret upper bound, Theorem 2, incurred by the IMED-EC algorithm.

Theorem (Upper bound on the number of pulls)**.** *Under the IMED-EC algorithms, under assumption 1 and 2, the number of pulls of a suboptimal arm a is upper bounded by:*

$$\mathbb{E}_\nu\left(N_a(T)\right) \leqslant \frac{\log T}{q\mathcal{K}_{\mathcal{F}}\left(\nu_a \| \mu_*\right)}\left(1+\alpha(\varepsilon)\right) + f(\varepsilon)\,, \tag{13}$$

where $0 < \varepsilon < \frac{1}{3}\min_{a\in\mathcal{A}\setminus\mathcal{A}_*}\left(\mu_* - \mu_a\right)$ α and α tends to 0 as ε tends to 0.

The proof proceeds in several steps. We first derive *empirical* bounds on the number of pulls of a suboptimal arm given that this arm is being pulled at time t.

Lemma 4 (Empirical bounds)**.** *Let a_{t+1} be the pulled arm at time $t+1$, and a be any arm belonging to $\mathcal{A}_*(t)$ at some time t. Under the IMED-EC algorithm,*
if $a_{t+1} \in \mathcal{A}_*(t)$,

$$N_{a_{t+1}}(t) \leqslant N_a(t), \tag{14}$$

$$\log\left(N_{a_{t+1}}(t)\right) \leqslant \min_{c_q \subseteq c_*} \sum_{k\in c_q} N_k(t) d\left(\widehat{\mu}_k(t) \| \widehat{\mu}^*\right) + \log N_k(t), \tag{15}$$

and if $a_{t+1} \notin \mathcal{A}_*(t)$,

$$qN_{a_{t+1}}(t) d\left(\widehat{\mu}_{a_{t+1}} \| \widehat{\mu}^*\right) \leqslant \log t, \tag{16}$$

$$q\log\left(N_{a_{t+1}}(t)\right) \leqslant \min_{c_q \subseteq c_*} \sum_{k\in c_q} N_k(t) d\left(\widehat{\mu}_k(t) \| \widehat{\mu}^*\right) + \log N_k(t)\,. \tag{17}$$

Proof. Assume that the chosen arm, a_{t+1}, belongs to $\mathcal{A}_*(t)$, then by definition of $I(t)$ and $I^*(t)$, $I^*(t) \leqslant I(t)$. $I^*(t) = I_{a_{t+1}}$ because a_{t+1} is the chosen arm amongst elements of $\mathcal{A}_*(t)$, hence belongs to $\arg\min_{a\in\mathcal{A}_*(t)} \log N_a(t)$. Equation (14) follows and $N_{a_{t+1}} \leqslant N_a(t)$ for all $a \in \mathcal{A}_*(t)$. Equation (15) then follows from the fact that:

$$\begin{aligned}
\log N_{a_{t+1}}(t) &= I^*(t) \\
&\leqslant I(t) \\
&= \min_{\substack{\mathcal{A}'\subset\mathcal{A} \\ |\mathcal{A}'|=q}} \sum_{a'\in\mathcal{A}'} I_{a'}(t) \\
&\leqslant \min_{c_q\subseteq c_*} \sum_{k\in c_q} N_k(t) d\left(\widehat{\mu}_k(t)\|\widehat{\mu}^*\right) + \log N_k(t).
\end{aligned}$$

Assume that the chosen arm, a_{t+1}, does not belong to $\mathcal{A}_*(t)$. Then $I(t) \leqslant I^*(t)$. The flow of control of the IMED-EC algorithm implies that a_{t+1} is an arm with minimal IMED index. By definition of $I(t)$, q times $I_{a_{t+1}}$ will always be smaller than or equal to $I(t)$:

$$q\left(N_{a_{t+1}}(t)\, d\left(\widehat{\mu}_{a_{t+1}}(t)\,\|\widehat{\mu}^*(t)\right) + \log N_{a_{t+1}}(t)\right) \leqslant I(t). \tag{18}$$

By definition of $I^*(t)$, there exists $a \in \mathcal{A}_*(t)$ such that $I^*(t) = \log N_a(t)$ implying that $I^*(t) \leqslant \log t$. Since $I(t) \leqslant I^*(t)$, it implies that $I(t) \leqslant \log t$. From equation (18) we deduce equation (16),

$$qN_{a_{t+1}}(t) d\left(\widehat{\mu}_{a_{t+1}} \| \widehat{\mu}^*\right) \leqslant \log t.$$

Last, equation (17) can be deduced from the definition of $I(t)$ and equation (18):

$$\begin{aligned}
q\log N_{a_{t+1}}(t) &\leqslant q\left(N_{a_{t+1}}(t)\, d\left(\widehat{\mu}_{a_{t+1}}(t)\,\|\widehat{\mu}^*(t)\right) + \log N_{a_{t+1}}(t)\right) \\
&\leqslant I(t) \\
&= \min_{\substack{\mathcal{A}'\subset\mathcal{A} \\ |\mathcal{A}'|=q}} \sum_{a'\in\mathcal{A}'} I_{a'}(t) \\
&\leqslant \min_{c_q\subseteq c_*} \sum_{k\in c_q} N_k(t) d\left(\widehat{\mu}_k(t)\|\widehat{\mu}^*\right) + \log N_k(t).
\end{aligned}$$

□

If we were to substitute empirical means with real ones, so that $\widehat{\mu}_a(t) = \mu_a$ for all $a \in \mathcal{A}$ and $\mathcal{A}_*(t) = \mathcal{A}_*$, then one can see that equation (16) gives us the desired behavior, *i.e.* if a_{t+1} is a suboptimal arm, $N_{a_{t+1}}(t) \leqslant \frac{\log t}{qd\left(\mu_{a_{t+1}} \| \mu_*\right)}$. For distributions having *concentration around the mean* property we will be able to say that, for large enough t, with high probability, for all arms $a \in \mathcal{A}$, $|\widehat{\mu}_a(t) - \mu_a| \leqslant \varepsilon$. In this case, we shall still have a desired property. This is the statement of the next Lemma. We then proceed to show the concentration properties later on.

Lemma 5. *Let $0 < \varepsilon \leqslant \frac{1}{3} \min_{a \notin \mathcal{A}_*} (\mu_* - \mu_a)$. Assume that a_{t+1} is the pulled arm at time $t+1$ and $a_{t+1} \in \mathcal{A} \setminus \mathcal{A}_*$ is a suboptimal arm. Let's assume that $\widehat{\mu}_{a_{t+1}}(t) \leqslant \mu_a + \varepsilon$. Let's also assume that $\widehat{\mu}^*(t) \geqslant \mu_* - \varepsilon$. Those hypothesis can be read as:*

$$\widehat{\mu}_{a_{t+1}}(t) \leqslant \mu_{a_{t+1}} + \varepsilon < \mu_* - \varepsilon \leqslant \mu^*(t). \tag{19}$$

Lemma 4 and the monotonicity of the KL *divergence implied by assumption 2, imply that:*

$$N_{a_{t+1}}(t) \leqslant \frac{\log t}{qd\left(\mu_{a_{t+1}} + \varepsilon \| \mu_* - \varepsilon\right)}. \tag{20}$$

Property 2 then implies that there exists α_a such that

$$N_{a_{t+1}}(t) \leqslant \frac{\log t}{qd\left(\mu_{a_{t+1}} \| \mu_*\right)} \left(1 + \alpha_a(\varepsilon)\right), \tag{21}$$

with $\alpha_a(\varepsilon) \to 0$ as $\varepsilon \to 0$.

Proof. Equation (19) is a direct consequence of the hypothesis of the Lemma. The strict inequality, $\mu_{a_{t+1}} < \mu_*$, implies that a_{t+1} cannot belong to $\mathcal{A}_*(t)$. Hence, equation (16) from Lemma 4 applies and

$$qN_{a_{t+1}}(t) d\left(\widehat{\mu}_{a_{t+1}}(t) \| \widehat{\mu}^*\right) \leqslant \log t\,.$$

Because $\widehat{\mu}_{a_{t+1}}(t) < \mu_* - \varepsilon \leqslant \widehat{\mu}^*(t)$, using assumption 2 that implies $d\left(\widehat{\mu}_{a_{t+1}} \| \mu_* - \varepsilon\right) \leqslant d\left(\widehat{\mu}_{a_{t+1}} \| \widehat{\mu}^*(t)\right)$, we have that:

$$qN_{a_{t+1}}(t) d\left(\widehat{\mu}_{a_{t+1}} \| \mu_* - \varepsilon\right) \leqslant \log t.$$

Similarly, $\widehat{\mu}_{a_{t+1}}(t) \leqslant \mu_{a_{t+1}}(t) + \varepsilon < \mu_* - \varepsilon$ and using assumption 2 again, $d\left(\mu_a + \varepsilon \| \mu_* - \varepsilon\right) \leqslant d\left(\widehat{\mu}_{a_{t+1}} \| \mu_* - \varepsilon\right)$, we proved equation (20):

$$N_{a_{t+1}}(t) \leqslant \frac{\log t}{qd\left(\mu_{a_{t+1}} + \varepsilon \| \mu_* - \varepsilon\right)},$$

Using equation (11) from property 2, we deduce that there exists α_a as in property 2 such that:

$$N_{a_{t+1}}(t) \leqslant \frac{\log t}{qd\left(\mu_{a_{t+1}} \| \mu_*\right)} \left(1 + \alpha_a(\varepsilon)\right).$$

□

In order to better clarify the proof of Theorem 2, we add two more lemmas that help emphasize where the hypothesis regarding the distribution space $\mathcal{F}$ are used. The first one is about decomposing the event of choosing a suboptimal arm.

Intuition about Lemma 6 The intuition of the next lemma is the following. Let a be a suboptimal arm. We will decompose the event of choosing arm a at time $t+1$, $\{a_{t+1} = a\}$, on three events. We recall that $\widehat{\mu}^*(t) = \max_{a \in \mathcal{A}} \widehat{\mu}_a(t)$. Let ε be such that $0 < \varepsilon \leqslant \frac{1}{3} \min_{a \notin \mathcal{A}_*} (\mu_* - \mu_a)$ as in Lemma 5.

From Lemma 5, we know that under the event $\left\{\widehat{\mu}^*(t) \geqslant \mu_* - \varepsilon, \widehat{\mu}_{a_{t+1}} \leqslant \mu_{a_{t+1}} + \varepsilon\right\}$, the number $N_{a_{t+1}}$ is upper bounded by the desired asymptotic term. The intuition of this term is given in section 3 and made formal in Lemma 5. We want to play an arm as frequently as the likelihood of optimality of the group of size q it may belong to.

Whatever the space of distributions $\mathcal{F}$, we will always assume a *concentration* around the mean property. Therefore, we know that the event $\left\{\widehat{\mu}_{a_{t+1}} > \mu_{a_{t+1}} + \varepsilon\right\}$ can be controlled by concentration

inequality (a_{t+1} is the chosen arm at time t). The intuition for controlling this term is that one cannot play an arm too much while not having a good estimation of its expected reward.

The remainder of the two aforementioned events, $\{a_{t+1} \notin \mathcal{A}_*, \widehat{\mu}^*(t) < \mu_* - \varepsilon\}$, is the most technical and difficult to handle. An intuition may be the following. If $\widehat{\mu}^*(t) \leqslant \mu_* - \varepsilon$, then it is true that for any arm $k \in \mathcal{A}_*$, $\widehat{\mu}_k(t) \leqslant \mu_* - \varepsilon$. In that case, it means that with a *high enough* probability, most of the arm $k \in \mathcal{A}_*$ have not been pulled too much. The `IMED` and `IMED-EC` algorithms try to match likelihood of optimality with frequency of play. Therefore, we cannot keep *not* playing arm in $\mathcal{A}_*$ (*i.e.* we cannot play the suboptimal arm for *too* long before playing k) and when that happens, we will mostly observe a *regression toward the mean* making the event even more unlikely. We will rely on *concentration* tools to make the *regression toward the mean* statement more precise.

We now formalize those intuitions in Lemma 6.

Lemma 6 (Fundamental decomposition)**.** *Let a be a suboptimal arm. Let ε be such that $0 < \varepsilon \leqslant \frac{1}{3}\min_{a \notin \mathcal{A}_*}(\mu_* - \mu_a)$ as in Lemma 5. Under `IMED-EC`, shifting the time index t by $|\mathcal{A}|$ (each arm is pulled once at the beginning), $N_a(t) = \sum_{t=1}^{T} \mathbb{1}\{a_{t+1} = a\}$, the number of time a suboptimal arm a has been pulled after time $|\mathcal{A}|$ can be upper bounded by:*

$$\sum_{t=1}^{T} \mathbb{1}\{a_{t+1} = a\} \leqslant \sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a, N_a(t) \leqslant \frac{\log t}{qd\left(\mu_a + \varepsilon \| \mu_* - \varepsilon\right)}\right\} \tag{22}$$

$$+ \sum_{t=1}^{T} \mathbb{1}\{a_{t+1} = a, \widehat{\mu}^*(t) \geqslant \mu_* - \varepsilon, \widehat{\mu}_a(t) > \mu_a + \varepsilon\} \tag{23}$$

$$+ \sum_{t=1}^{T} \mathbb{1}\{a_{t+1} = a, \widehat{\mu}^*(t) < \mu_* - \varepsilon\}. \tag{24}$$

Proof. Following the aforementioned intuition, we decompose the event $\{a_{t+1} = a\}$:

$$\begin{aligned}
\{a_{t+1} = a\} &= \{a_{t+1} = a, \widehat{\mu}^*(t) \geqslant \mu_* - \varepsilon\} \cup \{a_{t+1} = a, \widehat{\mu}^*(t) < \mu_* - \varepsilon\} \\
&= \{a_{t+1} = a, \widehat{\mu}^*(t) \geqslant \mu_* - \varepsilon, \widehat{\mu}_a(t) \leqslant \mu_a + \varepsilon\} \\
&\quad \cup \{a_{t+1} = a, \widehat{\mu}^*(t) \geqslant \mu_* - \varepsilon, \widehat{\mu}_a(t) > \mu_a + \varepsilon\} \\
&\quad \cup \{a_{t+1} = a, \widehat{\mu}^*(t) < \mu_* - \varepsilon\} \\
&= \{a_{t+1} = a, qN_a(t) d\left(\mu_a + \varepsilon \| \mu_* - \varepsilon\right) \leqslant \log t\} \qquad \textit{Using Lemma 5} \\
&\quad \cup \{a_{t+1} = a, \widehat{\mu}^*(t) \geqslant \mu_* - \varepsilon, \widehat{\mu}_a(t) > \mu_a + \varepsilon\} \\
&\quad \cup \{a_{t+1} = a, \widehat{\mu}^*(t) < \mu_* - \varepsilon\}.
\end{aligned}$$

Using indicators of those events and shifting the time index t by $|\mathcal{A}|$ (each arm is pulled once at the beginning), we can upper bound $N_a(t) = \sum_{t=1}^{T} \mathbb{1}\{a_{t+1} = a\}$, the number of time a suboptimal arm a as been pulled after time $|\mathcal{A}|$:

$$\sum_{t=1}^{T} \mathbb{1}\{a_{t+1} = a\} \leqslant \sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a, N_a(t) \leqslant \frac{\log t}{qd\left(\mu_a + \varepsilon \| \mu_* - \varepsilon\right)}\right\} \tag{22}$$

$$+ \sum_{t=1}^{T} \mathbb{1}\{a_{t+1} = a, \widehat{\mu}^*(t) \geqslant \mu_* - \varepsilon, \widehat{\mu}_a(t) > \mu_a + \varepsilon\} \tag{23}$$

$$+ \sum_{t=1}^{T} \mathbb{1}\{a_{t+1} = a, \widehat{\mu}^*(t) < \mu_* - \varepsilon\}. \tag{24}$$

□

The second lemma is about bounding equation (24) by a quantity that can be controlled in both assumptions 2 and 3. This particular control is very specific to the q-equivalence structure and rely heavily on the fact that there is at least q distributions in the optimal class $\mathcal{A}_*$.

Lemma 7 (q-factorization)**.** *Let a be a suboptimal arm and $c \subseteq \mathcal{A}_*$ be any subset of q optimal arms. Then, under the* `IMED-EC` *algorithm,*

$$\sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a, \widehat{\mu}^*(t) < \mu_* - \varepsilon\right\} \leqslant \prod_{k \in c} \sum_{\substack{m_1 \geqslant 1 \\ \vdots \\ m_q \geqslant 1}} \mathbb{1}\left\{\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon\right\} m_k \exp\left(m_k d\left(\widehat{\mu}_k^{m_k} \| \mu_* - \varepsilon\right)\right) .$$

Proof. We want to control part (24) of the upper bound on the number of pulls of a suboptimal arm a. From Lemma 4, equation (15) and equation (17), we know that $\log\left(N_{a_{t+1}}(t)\right) \leqslant \min_{c_q \subseteq c_*} \sum_{k \in c_q} N_k(t) d\left(\widehat{\mu}_k(t) \| \widehat{\mu}^*\right) + \log N_k(t)$. Let $c \subseteq \mathcal{A}_*$ be any subset of q optimal arms. Since we are studying the event $\{a_{t+1} = a\}$ this inequality becomes

$$\begin{aligned} \log\left(N_a(t)\right) &\leqslant \min_{c_q \subseteq c_*} \sum_{k \in c_q} N_k(t) d\left(\widehat{\mu}_k(t) \| \widehat{\mu}^*\right) + \log N_k(t) \\ &\leqslant \sum_{k \in c} N_k(t) d\left(\widehat{\mu}_k(t) \| \widehat{\mu}^*\right) + \log N_k(t) . \end{aligned}$$

We use this inequality to control the sum (24).

$$\begin{aligned} (24) &= \sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a, \widehat{\mu}^*(t) < \mu_* - \varepsilon\right\} \\ &= \sum_{t=1}^{T} \mathbb{1}\left\{\widehat{\mu}^*(t) < \mu_* - \varepsilon, \log N_a(t) \leqslant \sum_{k \in c} N_k(t) d\left(\widehat{\mu}_k(t) \| \widehat{\mu}^*(t)\right) + \log N_k(t)\right\} \times \\ &\qquad \mathbb{1}\left\{a_{t+1} = a\right\} \\ &= \sum_{t=1}^{T} \sum_{n=1}^{T} \mathbb{1}\left\{\widehat{\mu}^*(t) < \mu_* - \varepsilon, \log(n) \leqslant \sum_{k \in c} N_k(t) d\left(\widehat{\mu}_k(t) \| \widehat{\mu}^*(t)\right) + \log N_k(t)\right\} \times \\ &\qquad \mathbb{1}\left\{a_{t+1} = a, N_a(t) = n\right\} . \end{aligned}$$

Since $\widehat{\mu}^*(t) < \mu_* - \varepsilon$ and $\widehat{\mu}^*(t) = \max_{b \in \mathcal{A}} \widehat{\mu}_b(t)$, we can use the monotonicity of the KL divergence to state that for all $k \in c$, $d\left(\widehat{\mu}_k^{m_k} \| \widehat{\mu}^*(t)\right) \leqslant d\left(\widehat{\mu}_k(t) \| \mu_* - \varepsilon\right)$. This implies the inclusion of events,

$$\begin{aligned} &\left\{\widehat{\mu}^*(t) < \mu_* - \varepsilon, \log(n) \leqslant \sum_{k \in c} N_k(t) d\left(\widehat{\mu}_k(t) \| \widehat{\mu}^*(t)\right) + \log N_k(t)\right\} \subseteq \\ &\qquad \left\{\widehat{\mu}^*(t) < \mu_* - \varepsilon, \log(n) \leqslant \sum_{k \in c} N_k(t) d\left(\widehat{\mu}_k(t) \| \mu_* - \varepsilon\right) + \log N_k(t)\right\} , \end{aligned}$$

which can be used to control the indicators. Furthermore, $\widehat{\mu}^*(t) \leqslant \mu_* - \varepsilon$ implies that $\max_{k \in c} \widehat{\mu}_k(t) \leqslant \mu_* - \varepsilon$ since $\max_{k \in c} \widehat{\mu}_k(t) \leqslant \max_{k \in \mathcal{A}} \widehat{\mu}_k(t)$. Therefore,

$$\begin{aligned} &\left\{\widehat{\mu}^*(t) < \mu_* - \varepsilon, \log(n) \leqslant \sum_{k \in c} N_k(t) d\left(\widehat{\mu}_k(t) \| \widehat{\mu}^*(t)\right) + \log N_k(t)\right\} \subseteq \\ &\qquad \left\{\max_{k \in c} \widehat{\mu}_k(t) < \mu_* - \varepsilon, \log(n) \leqslant \sum_{k \in c} N_k(t) d\left(\widehat{\mu}_k(t) \| \mu_* - \varepsilon\right) + \log N_k(t)\right\} , \end{aligned}$$

which we use to control the indicators. We then obtain

$$
\begin{aligned}
(24) \leqslant \sum_{t=1}^{T}\sum_{n=1}^{T} \mathbb{1}\left\{\max_{k\in c}\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon, \log(n) \leqslant \sum_{k\in c} N_k(t) d\left(\widehat{\mu}_k(t) \| \mu_* - \varepsilon\right) + \log N_k(t)\right\} \times \\
\mathbb{1}\left\{a_{t+1} = a, N_a(t) = n\right\} \\
= \sum_{m_1\geqslant 1}\sum_{n\geqslant 1}\sum_{t=1}^{T} \mathbb{1}\left\{\max_{k\in c}\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon, \log(n) \leqslant \sum_{k\in c} m_k d\left(\widehat{\mu}_k^{m_k} \| \mu_* - \varepsilon\right) + \log m_k\right\} \times \\
\vdots \\
m_q\geqslant 1 \\
\mathbb{1}\left\{a_{t+1} = a, N_a(t) = n\right\} \underbrace{\prod_{k\in c} \mathbb{1}\left\{N_k(t) = m_k\right\}}_{\leqslant 1} \\
\leqslant \sum_{m_1\geqslant 1}\sum_{n\geqslant 1}\sum_{t=1}^{T} \mathbb{1}\left\{\max_{k\in c}\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon, \log(n) \leqslant \sum_{k\in c} m_k d\left(\widehat{\mu}_k^{m_k} \| \mu_* - \varepsilon\right) + \log m_k\right\} \times \\
\vdots \\
m_q\geqslant 1 \\
\mathbb{1}\left\{a_{t+1} = a, N_a(t) = n\right\} \\
= \sum_{m_1\geqslant 1}\sum_{n\geqslant 1} \mathbb{1}\left\{\max_{k\in c}\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon, \log(n) \leqslant \sum_{k\in c} m_k d\left(\widehat{\mu}_k^{m_k} \| \mu_* - \varepsilon\right) + \log m_k\right\} \times \\
\vdots \\
m_q\geqslant 1 \\
\underbrace{\sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a, N_a(t) = n\right\}}_{\leqslant 1} \\
\leqslant \sum_{m_1\geqslant 1}\sum_{n\geqslant 1} \mathbb{1}\left\{\max_{k\in c}\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon, \log(n) \leqslant \sum_{k\in c} m_k d\left(\widehat{\mu}_k^{m_k} \| \mu_* - \varepsilon\right) + \log m_k\right\}. \\
\vdots \\
m_q\geqslant 1
\end{aligned}
$$

We can then factorize the following term

$$
\begin{aligned}
\mathbb{1}\left\{\max_{k\in c}\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon, \log(n) \leqslant \sum_{k\in c} m_k d\left(\widehat{\mu}_k^{m_k} \| \mu_* - \varepsilon\right) + \log m_k\right\} = \\
\mathbb{1}\left\{\max_{k\in c}\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon\right\} \mathbb{1}\left\{\log(n) \leqslant \sum_{k\in c} m_k d\left(\widehat{\mu}_k^{m_k} \| \mu_* - \varepsilon\right) + \log m_k\right\},
\end{aligned}
$$

and remark that $\mathbb{1}\left\{\max_{k\in c}\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon\right\}$ does not depend on n. Hence,

$$
\begin{aligned}
(24) &\leqslant \sum_{m_1\geqslant 1}\sum_{n\geqslant 1}\mathbb{1}\left\{\max_{k\in c}\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon, \log(n) \leqslant \sum_{k\in c} m_k d\left(\widehat{\mu}_k^{m_k}\|\mu_* - \varepsilon\right) + \log m_k\right\} \\
&\quad \vdots \\
&\quad \sum_{m_q\geqslant 1} \\
&= \sum_{m_1\geqslant 1}\mathbb{1}\left\{\max_{k\in c}\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon\right\}\sum_{n\geqslant 1}\mathbb{1}\left\{\log(n) \leqslant \sum_{k\in c} m_k d\left(\widehat{\mu}_k^{m_k}\|\mu_* - \varepsilon\right) + \log m_k\right\} \\
&\quad \vdots \\
&\quad \sum_{m_q\geqslant 1} \\
&\leqslant \sum_{m_1\geqslant 1}\mathbb{1}\left\{\max_{k\in c}\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon\right\}\exp\left(\sum_{k\in c} m_k d\left(\widehat{\mu}_k^{m_k}\|\mu_* - \varepsilon\right) + \log m_k\right). \\
&\quad \vdots \\
&\quad \sum_{m_q\geqslant 1}
\end{aligned}
$$

Since $\exp\left(\sum_{k\in c} m_k d\left(\widehat{\mu}_k^{m_k}\|\mu_* - \varepsilon\right) + \log m_k\right) = \prod_{k\in c} m_k \exp\left(m_k d\left(\widehat{\mu}_k^{m_k}\|\mu_* - \varepsilon\right)\right)$ and $\mathbb{1}\left\{\max_{k\in c}\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon\right\} = \prod_{k\in c}\mathbb{1}\left\{\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon\right\}$, we can rewrite the last bound as

$$
\begin{aligned}
(24) &\leqslant \sum_{m_1\geqslant 1}\prod_{k\in c}\mathbb{1}\left\{\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon\right\} m_k \exp\left(m_k d\left(\widehat{\mu}_k^{m_k}\|\mu_* - \varepsilon\right)\right) \\
&\quad \vdots \\
&\quad \sum_{m_q\geqslant 1} \\
&\leqslant \prod_{k\in c}\sum_{m_1\geqslant 1}\mathbb{1}\left\{\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon\right\} m_k \exp\left(m_k d\left(\widehat{\mu}_k^{m_k}\|\mu_* - \varepsilon\right)\right). \\
&\quad \vdots \\
&\quad \sum_{m_q\geqslant 1}
\end{aligned}
$$

where we used the fact that a sum of product of non-negative terms is not greater than the product of the sum of these terms, since one contains all the terms of the other. □

Thanks to Lemma 7, controlling the Equation (24) amounts to controlling terms like

$$
\mathbb{E}\left(\sum_{m_k\geqslant 1}\mathbb{1}\left\{\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon\right\} m_k \exp\left(m_k d\left(\widehat{\mu}_k^{m_k}\|\mu_* - \varepsilon\right)\right)\right), \tag{25}
$$

which is linked to the upper bound one can have on $\mathbb{P}\left(d\left(\widehat{\mu}_k^{m_k}\|\mu_* - \varepsilon\right) \geqslant u, \widehat{\mu}_k^{m_k} \leqslant \mu_* - \varepsilon\right)$ and whether or not it is better than $\exp(-m_k u)$. As we can see, controlling this terms amounts to a *concentration* hypothesis of the distributions in $\mathcal{F}$. Assumption 2 and assumption 3 are two possible assumptions that make it possible to control the Equation (25).

We are now ready to prove the Theorem 2.

Proof of Theorem 2. Let a be a suboptimal arm and let ε be such that $0 < \varepsilon \leqslant \frac{1}{3}\min_{a\notin\mathcal{A}_*}(\mu_* - \mu_a)$. By Lemma 6 we have the following decomposition:

$$
\begin{aligned}
\sum_{t=1}^{T}\mathbb{1}\left\{a_{t+1} = a\right\} &\leqslant \sum_{t=1}^{T}\mathbb{1}\left\{a_{t+1} = a, N_a(t) \leqslant \frac{\log t}{q d\left(\mu_a + \varepsilon\|\mu_* - \varepsilon\right)}\right\} && (22) \\
&\quad + \sum_{t=1}^{T}\mathbb{1}\left\{a_{t+1} = a, \widehat{\mu}^*(t) \geqslant \mu_* - \varepsilon, \widehat{\mu}_a(t) > \mu_a + \varepsilon\right\} && (23) \\
&\quad + \sum_{t=1}^{T}\mathbb{1}\left\{a_{t+1} = a, \widehat{\mu}^*(t) < \mu_* - \varepsilon\right\}. && (24)
\end{aligned}
$$

Control of equation (22) Equation (22) can be controlled as a random variable without any concentration tools. The following derivation bound the sum (22) by:

$$\begin{aligned}
(22) &= \sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a, N_a(t) \leqslant \frac{\log t}{qd\left(\mu_a + \varepsilon \| \mu_* - \varepsilon\right)}\right\} \\
&= \sum_{n=1}^{T} \sum_{t=1}^{T} \mathbb{1}\left\{n \leqslant \frac{\log t}{qd\left(\mu_a + \varepsilon \| \mu_* - \varepsilon\right)}\right\} \mathbb{1}\left\{a_{t+1} = a, N_a(t) = n\right\} \\
&= \sum_{n=1}^{T} \mathbb{1}\left\{n \leqslant \frac{\log T}{qd\left(\mu_a + \varepsilon \| \mu_* - \varepsilon\right)}\right\} \underbrace{\sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a, N_a(t) = n\right\}}_{\leqslant 1} \\
&\leqslant \sum_{n=1}^{T} \mathbb{1}\left\{n \leqslant \frac{\log T}{qd\left(\mu_a + \varepsilon \| \mu_* - \varepsilon\right)}\right\} \\
&= \left\lfloor \frac{\log T}{qd\left(\mu_a + \varepsilon \| \mu_* - \varepsilon\right)} \right\rfloor \\
&\leqslant \frac{\log T}{qd\left(\mu_a + \varepsilon \| \mu_* - \varepsilon\right)}.
\end{aligned}$$

The derivation relies on the simple fact that an indicator function is upper bounded by 1. This part proved that:

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a\right\} \leqslant \frac{\log T}{qd\left(\mu_a + \varepsilon \| \mu_* - \varepsilon\right)} & \qquad (22') \\
+ \sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a, \widehat{\mu}^*(t) \geqslant \mu_* - \varepsilon, \widehat{\mu}_a(t) > \mu_a + \varepsilon\right\} & \qquad (23) \\
+ \sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a, \widehat{\mu}^*(t) < \mu_* - \varepsilon\right\} . & \qquad (24)
\end{aligned}$$

Control of equation (23) Equation (23) can be controlled using large deviation hypothesis on the set of distributions that are considered. It should be noted that this term is also bounded by $O(1)$ in the paper Honda and Takemura [2015]. Therefore, this term can be also be handled under assumption 3. We give an upper bound under assumption 2. The common fact of those two assumptions is the *light-tail* property of the considered distributions. A distributions is said *light-tailed* if its moment generating function exists in a neighborhood of 0. In that case one can apply a concentration property

since Cramer's theorem applies (see [Dembo and Zeitouni, 1998, Theorem 2.2.3]).

$$
\begin{aligned}
(23) &= \sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a, \widehat{\mu}^*(t) \geqslant \mu_* - \varepsilon, \widehat{\mu}_a(t) > \mu_a + \varepsilon\right\} \\
&\leqslant \sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a, \widehat{\mu}_a(t) > \mu_a + \varepsilon\right\} \\
&= \sum_{n=1}^{T} \sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a, N_a(t) = n, \widehat{\mu}_a(t) > \mu_a + \varepsilon\right\} \\
&= \sum_{n=1}^{T} \sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a, N_a(t) = n\right\} \mathbb{1}\left\{\widehat{\mu}_a^n > \mu_a + \varepsilon\right\} \\
&= \sum_{n=1}^{T} \mathbb{1}\left\{\widehat{\mu}_a^n > \mu_a + \varepsilon\right\} \underbrace{\sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a, N_a(t) = n\right\}}_{\leqslant 1} \\
&\leqslant \sum_{n=1}^{T} \mathbb{1}\left\{\widehat{\mu}_a^n > \mu_a + \varepsilon\right\}.
\end{aligned}
$$

Taking the expectation of both sides,

$$
\mathbb{E}\left((23)\right) \leqslant \sum_{n=1}^{T} \mathbb{P}\left(\widehat{\mu}_a^n > \mu_a + \varepsilon\right),
$$

which is a series of positive real numbers between 0 and 1 if we set $T = +\infty$. The reason we are interested in the limit is because we want $\mathbb{E}\left((23)\right)$ to be upper bounded by a time independent quantity. For this series to be convergent, we need the terms of the series, $\left(\mathbb{P}\left(\widehat{\mu}_a(n) > \mu_a + \varepsilon\right)\right)_n$ to converge fast enough toward 0. Denoting φ_a the moment generating function of arm a and $\psi_a = \log \varphi_a$ the cumulant generating function we derive that for all $\lambda > 0$:

$$
\begin{aligned}
\mathbb{P}\left(\widehat{\mu}_a^n > \mu_a + \varepsilon\right) &= \mathbb{P}\left(\sum_{i=1}^{n} X_i^a > n\left(\mu_a + \varepsilon\right)\right) \\
&\leqslant \frac{\varphi_a\left(\lambda\right)^n}{e^{n\lambda(\mu_a+\varepsilon)}} && \textit{Markov inequality} \\
&= \exp\left(n\left(\psi_a\left(\lambda\right) - \lambda\left(\mu_a + \varepsilon\right)\right)\right) \\
&= \exp\left(-n\left(\lambda\left(\mu_a + \varepsilon\right) - \psi_a\left(\lambda\right)\right)\right).
\end{aligned}
$$

Since this inequality is true for all λ we can minimize the right-hand side expression which features the Legendre-Fenchel transformation of ψ_a (also known as the Cramer transform). We denote $\psi_a^*(\varepsilon) = \sup_\lambda\left(\lambda\left(\mu_a + \varepsilon\right) - \psi_a\left(\lambda\right)\right)$ the Legendre-Fenchel transform of ψ_a that exists thanks to assumption 2.

For distributions in $\mathcal{F}$ this quantity is strictly positive as it is proved in Dembo and Zeitouni [1998]. Therefore, we proved that

$$
\begin{aligned}
\mathbb{P}\left(\widehat{\mu}_a^n > \mu_a + \varepsilon\right) &\leqslant \exp\left(-n\psi_a^*\left(\varepsilon\right)\right) \\
&= \left(\exp\left(-\psi_a^*\left(\varepsilon\right)\right)\right)^n,
\end{aligned}
$$

with $0 \leqslant \exp\left(-\psi_a^*\left(\varepsilon\right)\right) < 1$. This result is enough to bound equation (23) in expectation,

$$\begin{aligned}
\mathbb{E}\left((23)\right) &\leqslant \sum_{n=1}^{T} \mathbb{P}\left(\widehat{\mu}_a^n > \mu_a + \varepsilon\right) \\
&\leqslant \sum_{n=1}^{+\infty} \mathbb{P}\left(\widehat{\mu}_a^n > \mu_a + \varepsilon\right) \\
&\leqslant \sum_{n=1}^{+\infty} \exp\left(-n\psi_a^*\left(\varepsilon\right)\right) \\
&= C_a\left(\varepsilon\right),
\end{aligned}$$

where $C_a(\varepsilon)$ denotes the limit of the series $\sum_{n=1}^{+\infty} \exp\left(-n\psi_a^*\left(\varepsilon\right)\right)$. This part and the previous one proved that

$$\mathbb{E}\left(\sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a\right\}\right) \leqslant \frac{\log T}{qd\left(\mu_a + \varepsilon \| \mu_* - \varepsilon\right)} \tag{22'}$$

$$+ C_a\left(\varepsilon\right) \tag{23'}$$

$$+ \mathbb{E}\left(\sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a, \widehat{\mu}^*(t) < \mu_* - \varepsilon\right\}\right). \tag{24}$$

Control of equation (24) We are left to control part (24) of the upper bound on the number of pulls of a suboptimal arm. Let $c \subseteq \mathcal{A}_*$ be any subset of q optimal arms. From the q-factorization lemma, Lemma 7, we know that

$$\sum_{t=1}^{T} \mathbb{1}\left\{a_{t+1} = a, \widehat{\mu}^*(t) < \mu_* - \varepsilon\right\} \leqslant \prod_{k \in c} \sum_{\substack{m_1 \geqslant 1 \\ \vdots \\ m_q \geqslant 1}} \mathbb{1}\left\{\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon\right\} m_k \exp\left(m_k d\left(\widehat{\mu}_k^{m_k} \| \mu_* - \varepsilon\right)\right).$$

Since samples from the different arms are independents, the inequality holds in expectation:

$$\mathbb{E}\left((24)\right) \leqslant \prod_{k \in c} \mathbb{E}\left(\sum_{\substack{m_1 \geqslant 1 \\ \vdots \\ m_q \geqslant 1}} \mathbb{1}\left\{\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon\right\} m_k \exp\left(m_k d\left(\widehat{\mu}_k^{m_k} \| \mu_* - \varepsilon\right)\right)\right).$$

The proof of Lemma 14 in Honda and Takemura [2015], from equation (26), features a very similar quantity. In particular, it has been proved that for all $k \in c$, under the assumption 3

$$\mathbb{E}\left(\sum_{m_k \geqslant 1} \mathbb{1}\left\{\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon\right\} m_k \exp\left(m_k d\left(\widehat{\mu}_k^{m_k} \| \mu_* - \varepsilon\right)\right)\right) \leqslant D_{\mathit{IMED}}(\varepsilon), \tag{26}$$

where $D_{\mathit{IMED}}(\varepsilon)$ is given by equation (28) of Honda and Takemura [2015]. The proof of Lemma 14 in Honda and Takemura [2015] relies on the Proposition 11 of Honda and Takemura [2015]. Under assumption 2, Proposition 1 allows us to upper bound the left hand side of Equation (27) by a time independent quantity using the same strategy of integration by parts used in Honda and Takemura [2015]:

$$\mathbb{E}\left(\sum_{m_k \geqslant 1} \mathbb{1}\left\{\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon\right\} m_k \exp\left(m_k d\left(\widehat{\mu}_k^{m_k} \| \mu_* - \varepsilon\right)\right)\right) \leqslant D(\varepsilon). \tag{27}$$

This upper bounds amount to the fact that $u \mapsto \exp\left(-\alpha\sqrt{u}\right)$ is integrable on $[0, +\infty[$ for all $\alpha > 0$. Let $P(u) = \mathbb{P}_{\nu_k}\left(d\left(\widehat{\mu}_k^{m_k} \| \mu_* - \varepsilon\right) > u, \widehat{\mu}_k^{\mu_k} < \mu_* - \varepsilon\right)$. Under assumption 2, Proposition 1 upper bounds $P(u)$, $P(u) \leqslant \exp\left(-m_k\left(u + \lambda\left(\mu_*, \mu_* - \varepsilon\right)\sqrt{u}\right)\right)$ with $\lambda\left(\mu_*, \mu_* - \varepsilon\right) > 0$. Therefore, we can integrate against $P(u)$

$$
\begin{aligned}
&\sum_{m_k \geqslant 1} \mathbb{E}\left(\mathbb{1}\left\{\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon\right\} m_k \exp\left(m_k d\left(\widehat{\mu}_k^{m_k} \| \mu_* - \varepsilon\right)\right)\right) \\
&\qquad = \int_0^\infty m_k \exp\left(m_k u\right)\left(-dP(u)\right) \\
&\qquad = \sum_{m_k \geqslant 1}\left[m_k \exp\left(m_k u\right)\left(-P(u)\right)\right]_0^\infty + \int_0^\infty m_k^2 \exp\left(m_k u\right) P(u) du \\
&\qquad \leqslant h\left(\varepsilon\right) + \sum_{m_k \geqslant 1} \int_0^\infty m_k^2 \exp\left(m_k u\right) \exp\left(-m_k\left(u + \lambda\left(\mu_*, \mu_* - \varepsilon\right)\sqrt{u}\right)\right) du \\
&\qquad = h\left(\varepsilon\right) + \sum_{m_k \geqslant 1} \int_0^\infty m_k^2 \exp\left(-m_k \lambda\left(\mu_*, \mu_* - \varepsilon\right)\sqrt{u}\right) du \\
&\qquad \leqslant h\left(\varepsilon\right) + p\left(\varepsilon\right) \\
&\qquad = D\left(\varepsilon\right),
\end{aligned}
$$

which proves our claim. Therefore, under assumption 2 or assumption 3, we can deduce a bound on equation (24),

$$
\begin{aligned}
(24) &\leqslant \prod_{k \in c} \sum_{\substack{m_1 \geqslant 1 \\ \vdots \\ m_q \geqslant 1}} \mathbb{1}\left\{\widehat{\mu}_k^{m_k} < \mu_* - \varepsilon\right\} m_k \exp\left(m_k d\left(\widehat{\mu}_k^{m_k} \| \mu_* - \varepsilon\right)\right) \\
&\leqslant D\left(\varepsilon\right)^q,
\end{aligned}
$$

because c is a set of size q.

All combined, those derivations proved that

$$
\begin{aligned}
\mathbb{E}\left(\sum_{t=1}^T \mathbb{1}\left\{a_{t+1} = a\right\}\right) &\leqslant \frac{\log T}{q d\left(\mu_a + \varepsilon \| \mu_* - \varepsilon\right)} && (22') \\
&\quad + C_a\left(\varepsilon\right) && (23') \\
&\quad + D(\varepsilon)^q, && (24')
\end{aligned}
$$

and can be written as

$$
\mathbb{E}\left(N_a(t)\right) \leqslant \frac{\log T}{q d\left(\mu_a + \varepsilon \| \mu_* - \varepsilon\right)} + C_a\left(\varepsilon\right) + D(\varepsilon)^q.
$$

□

Using property 2 for the KL divergence, we finally get the final expression proving the Theorem 2,

$$
\mathbb{E}\left(N_a(t)\right) \leqslant \frac{\log T}{q d\left(\mu_a \| \mu_*\right)}\left(1 + \alpha_a(\varepsilon)\right) + f(\varepsilon),
$$

with $f(\varepsilon) = C_a\left(\varepsilon\right) + D(\varepsilon)^q$.

Proof of Theorem 3. The Theorem 2 we just proved states that

$$
\mathbb{E}\left(N_a(t)\right) \leqslant \frac{\log T}{q d\left(\mu_a \| \mu_*\right)}\left(1 + \alpha_a(\varepsilon)\right) + f(\varepsilon)
$$

with $f(\varepsilon) = C_a\left(\varepsilon\right) + D(\varepsilon)^q$. Dividing both sides by $\log T$, we get that for all ε small enough

$$
\liminf_{t \to +\infty} \frac{\mathbb{E}_\nu\left(N_a(T)\right)}{\log T} \leqslant \frac{1}{q d\left(\mu_a \| \mu_*\right)}\left(1 + \alpha_a(\varepsilon)\right).
$$

Letting ε tends to 0 proves the Theorem 3. □

D Experiments

In this section, we illustrate the performances of IMED-EC with a few more plots and explore the *dispatching* of this algorithm. By *dispatching*, we mean to compare the discrepancy in the number of pulls within a class. In particular, we are interested in the behavior of the different sampling strategies within the optimal class. The different settings are the same as the one presented in the main part of this paper. We present additional details.

Balanced class, perfect knowledge In this set of experiments, see Figure 5, we focus on the bandit configurations in which all equivalence classes have the same cardinality and assume that we know the number of elements per class. Recall that since OSSB has to solve a combinatorial optimization problem at each time step, we can cannot carry experiments on large set of arms while comparing IMED-EC to it. In this particular setting, we see that while OSSB and IMED-EC are provably

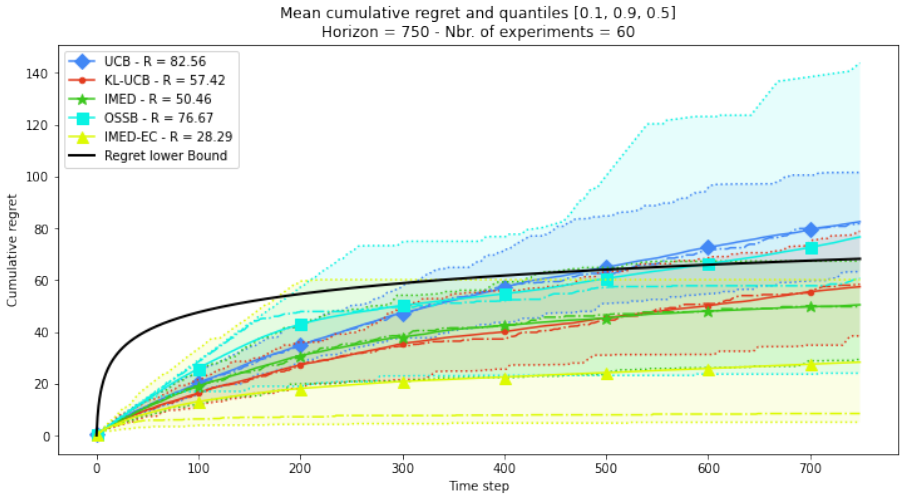

Figure 5: 3 classes, 2 distributions per class - set of means = $\{0.1, 0.3, 0.7\}$

asymptotically optimal, IMED-EC numerically performs better in finite time horizon.

Next, we explore the *dispatching* of IMED-EC and compare it to the one of KLUCB and IMED. We run the three algorithms 1000 times on an bandit problem whose number of class is 4 with means $\{0.1, 0.3, 0.6, 0.9\}$ and 10 Gaussian distributions with unit variance within each class. The chosen horizon is 2000. We assume that IMED-EC has perfect knowledge on the number of elements per class, 10. In the next section, similar plots (see Figure 11, Figure 12, Figure 13 and Figure 14) can be found where IMED-EC only knows a strict lower bound on the number of elements per class. For each of the 4 classes, we report the histogram accounting for the number of times distributions within each class has been pulled. Specifically, we are interested in the statistical order of the number of pulls within each class. After each run, for each class, we sort the number of pulls. The histograms are built using those sorted number of pulls. Error bars corresponds to the standard deviations and have been clipped to not go below the x-axis.

For the most suboptimal class, Figure 6, not much can be said since the number of pulls is very low. Still, one can see that the progression of the order statistics for KLUCB and IMED is somewhat *linear* while it seems more *exponential* for IMED-EC. (Note that we use these terms here informally.)

The same linear versus exponential apparent behavior can be seen on Figure 7.

On Figure 8, one can clearly a difference in the behaviour of KLUCB and IMED, that have *small* error bars, and IMED-EC that have a large error bar for the most pulled arm within the least suboptimal class. We can clearly see how risky it might be to reduce the exploration from this error bar. However, this risk is compensated by the fact that there is at least q similar distributions. This can be read from the fact that the sum of all the number of pulls within this class for KLUCB and IMED is above 200 which is roughly the maximal number of pulls of IMED-EC within this class computed using the upper bounds (given by the maximal value of the standard deviation).

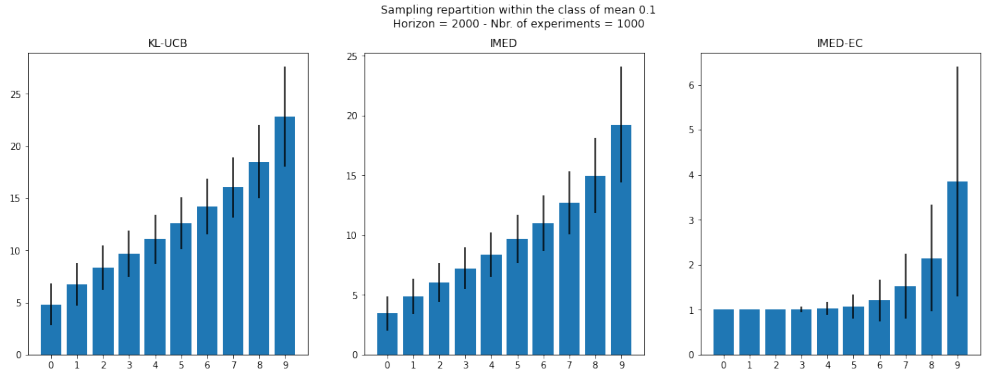

Figure 6: 4 classes, 10 distributions per class - set of means = $\{\mathbf{0.1}, 0.3, 0.6, 0.9\}$ - class of mean 0.1

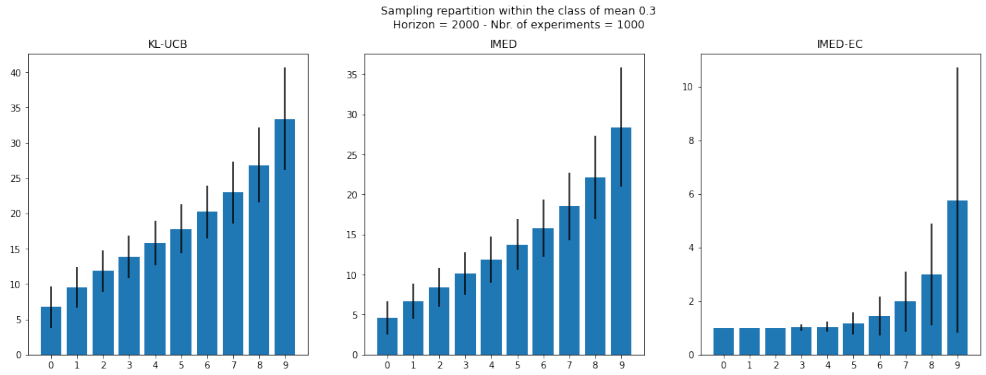

Figure 7: 4 classes, 10 distributions per class - set of means = $\{0.1, \mathbf{0.3}, 0.6, 0.9\}$ - class of mean 0.3

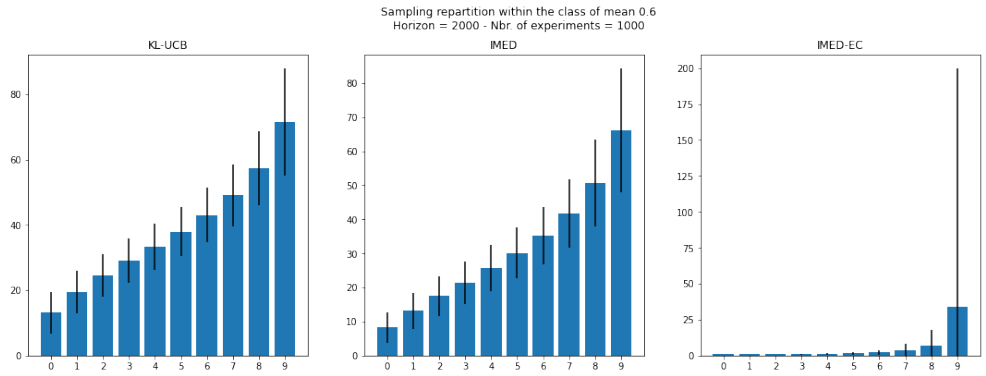

Figure 8: 4 classes, 10 distributions per class - set of means = $\{0.1, 0.3, \mathbf{0.6}, 0.9\}$ - class of mean 0.6

Finally, Figure 9 enables to compare the behaviours of the algorithms within the optimal class. It seems clear that, at least numerically, `IMED-EC` is not a fair algorithm in finite time (in the sense that it does not equally distribute the pulls between arms from the same class) and that it leverages the lower bound on the number of elements per class to play a more risky strategy, and benefits from it. Again, we observe the linear versus exponential progression in the order statistics of the number of pulls.

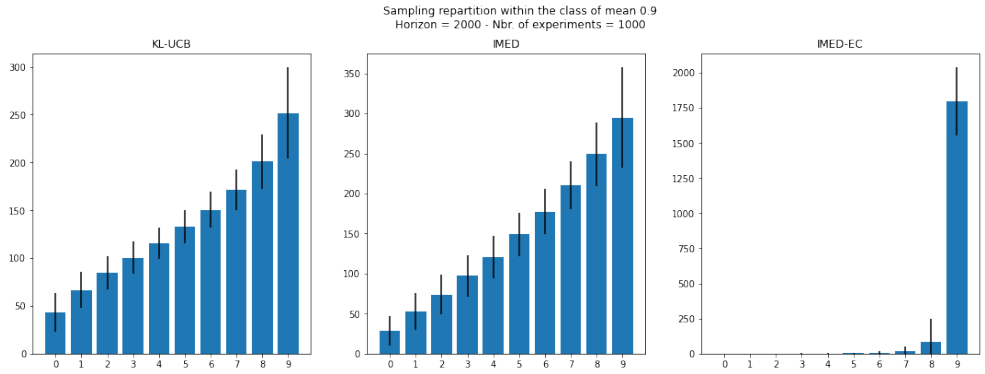

Figure 9: 4 classes, 10 distributions per class - set of means = $\{0.1, 0.3, 0.6, \mathbf{0.9}\}$ - class of mean 0.9

Imperfect knowledge In the experiment plotted Figure 10, we leverage the knowledge hypothesis and assume that we only know a lower bound on the number of elements per class while the classes are still balanced. We compare `IMED-EC` to unspecialized bandit algorithm, `IMED` and `KLUCB`. We can

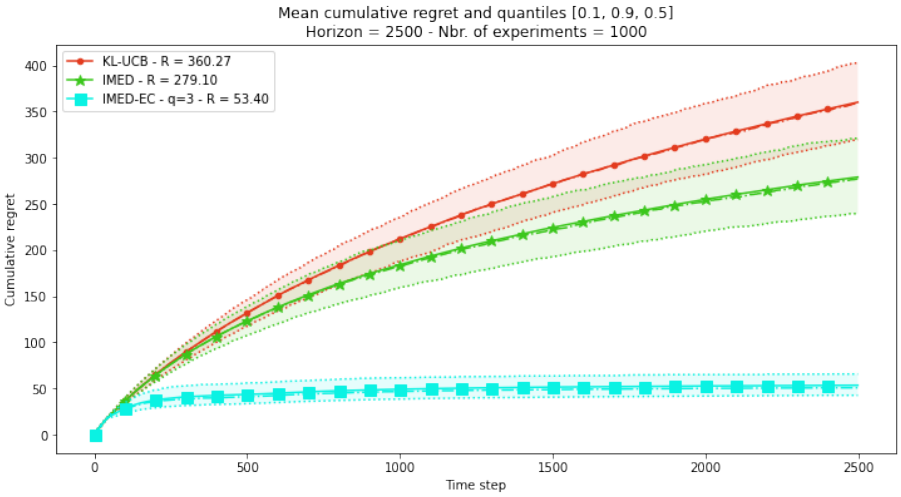

Figure 10: 7 classes, 10 distributions per class - set of means = $\{0.1, 0.3, 0.4, 0.5, 0.6, 0.75, 0.9\}$

see that the finite time cumulative regret of `IMED-EC` indeed is much smaller than the regret of the unspecialized algorithms, showing the ability of `IMED-EC` to effectively exploit this weak knowledge.

For the sake of completeness, we explore the *dispatching* of `IMED-EC` and compare it to the one of `KLUCB` and `IMED` on this setting. We run the three algorithms 1000 times on an bandit problem whose number of class is 4 with means $\{0.1, 0.3, 0.6, 0.9\}$ and 10 Gaussian distributions with unit variance within each class. The chosen horizon is 2000. We assume that `IMED-EC` does not have perfect knowledge on the number of elements per class, and we use 3 as the lower bound parameter of `IMED-EC`. For each of the 4 classes, we report the histogram accounting for the number of times distributions within each class has been pulled.Error bars corresponds to the standard deviations and have been clipped to not go below the x-axis.

Comments that were respectively made for Figure 6, Figure 7, Figure 8, and Figure 9 can similarly made for Figure 11, Figure 12, Figure 13, and Figure 14.

Interestingly, we can tell apart the two settings by looking at the behaviour of the algorithms for the least suboptimal class, *i.e.* comparing Figure 8 and Figure 13. In Figure 8 the error bar for the most pulled elements is much larger than in Figure 13 meaning that the algorithm at stakes explore less.

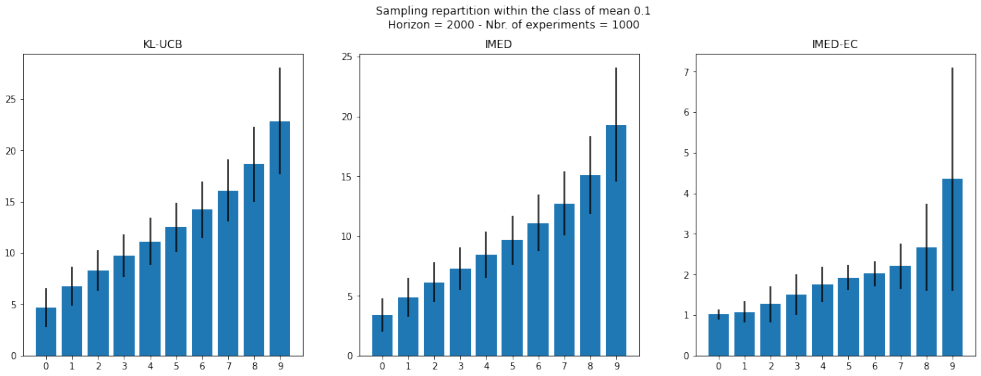

Figure 11: 4 classes, 10 distributions per class - set of means = $\{\mathbf{0.1}, 0.3, 0.6, 0.9\}$ - mean 0.1

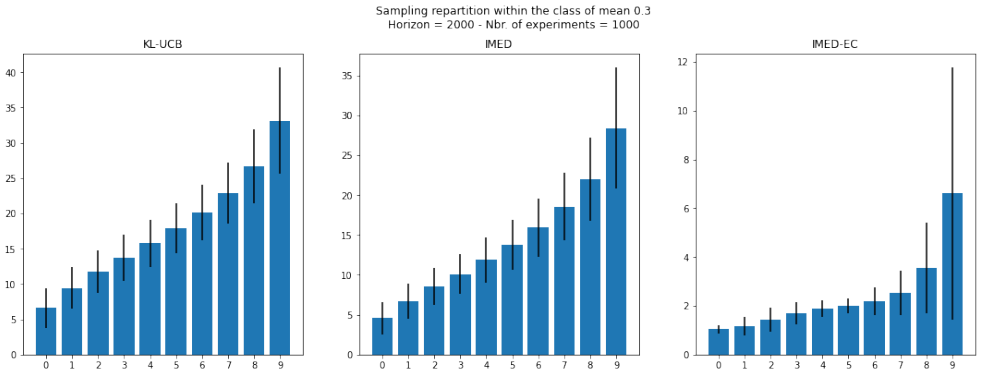

Figure 12: 4 classes, 10 distributions per class - set of means = $\{0.1, \mathbf{0.3}, 0.6, 0.9\}$ - mean 0.3

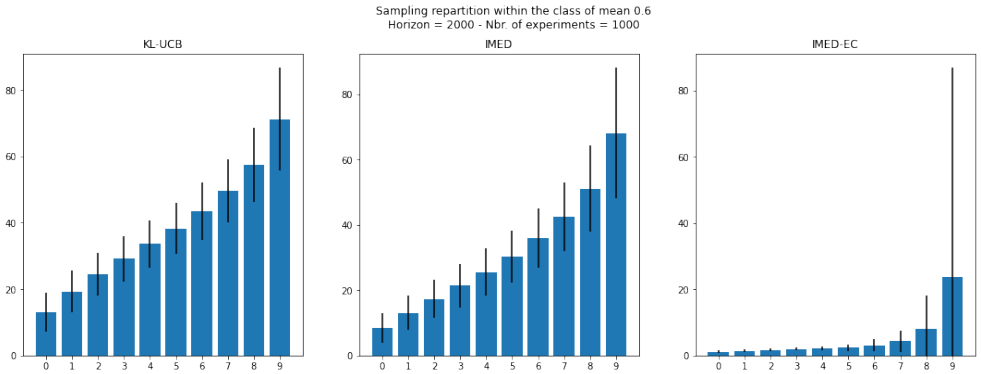

Figure 13: 4 classes, 10 distributions per class - set of means = $\{0.1, 0.3, \mathbf{0.6}, 0.9\}$ - mean 0.6

Influence of the parameter q Here we show the numerical robustness of `IMED-EC` with respect to the lower bound parameter q on the number of elements per classes. On the same bandit problem, we compare different instances of `IMED-EC` where different values of q are used. In the legend, *opt.* stands for optimal and corresponds to the largest valid lower bound on the number of elements per class, *i.e.* the minimal number of elements in a class. The experiments done for Figure 15 are performed on a bandit problem with 4 classes and and 10 distributions per class. While q increases up to the minimum cardinality of a class, we see that the performances of `IMED-EC` increases, which is expected. It is rather remarkable that once we go beyond that theoretical threshold, the performances of `IMED-EC` do not seem to deteriorate.

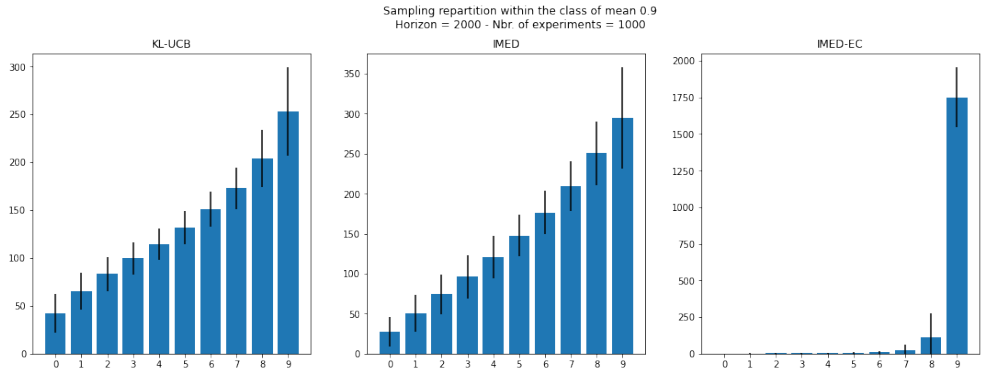

Figure 14: 4 classes, 10 distributions per class - set of means = $\{0.1, 0.3, 0.6, \mathbf{0.9}\}$ - mean 0.9

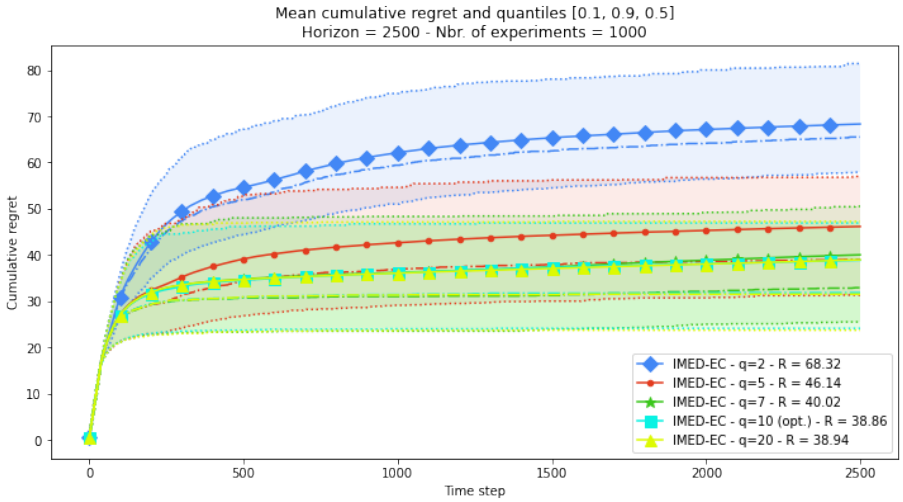

Figure 15: 4 classes, 10 distributions per class - set of means = $\{0.1, 0.3, 0.6, 0.9\}$

Finally, we explore the *dispatching* of `IMED-EC` and compare it to the one of `KLUCB` and `IMED` on this setting. We run the three algorithms 1000 times on a bandit problem whose number of class is 7 with means $\{0.1, 0.3, 0.4, 0.5, 0.6, 0.75, 0.9\}$ and an uneven number of Gaussian distributions with unit variance within each class. The chosen horizon is 2000. We assume that `IMED-EC` does not have perfect knowledge on the number of elements per class, and we use 3 as the lower bound parameter of `IMED-EC`. For some classes, we report the histogram accounting for the number of times distributions within each chosen class has been pulled. Error bars corresponds to the standard deviations and have been clipped to not go below the x-axis.

The comments that can be made about those plots are similar to the one that were already made for similar experiments. We included them to show that the behaviour of `IMED-EC` (and also the behaviour of `IMED` and `KLUCB`) is consistent across multiple settings. In particular, the algorithm `IMED-EC` exhibits the same aforementioned behaviour for the least suboptimal class, as it can be seen by comparing Figure 18 to the corresponding Figure 8 and Figure 13.

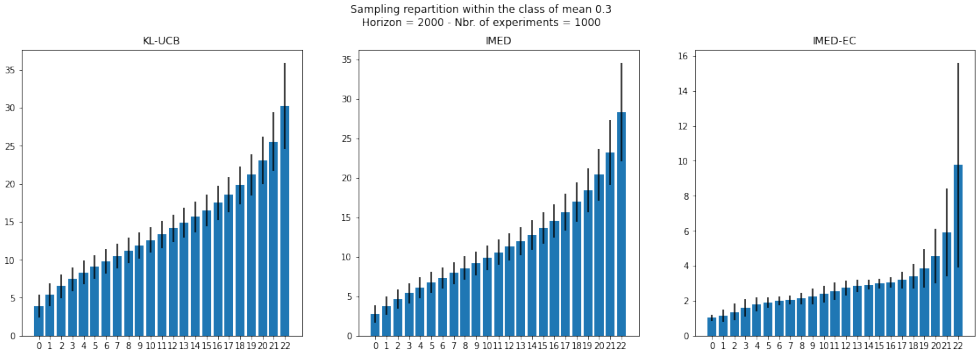

Figure 16: 7 classes - unbalanced - set of means = $\{0.1, \mathbf{0.3}, 0.4, 0.5, 0.6, 0.75, 0.9\}$ - mean 0.3

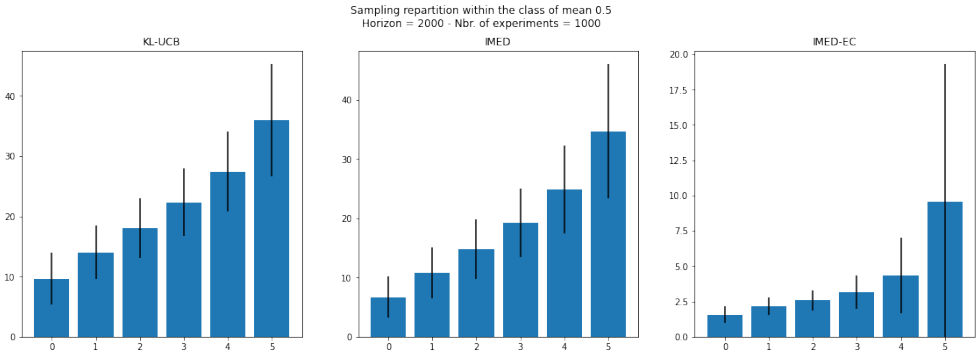

Figure 17: 7 classes - unbalanced - set of means = $\{0.1, 0.3, 0.4, \mathbf{0.5}, 0.6, 0.75, 0.9\}$ - mean 0.5

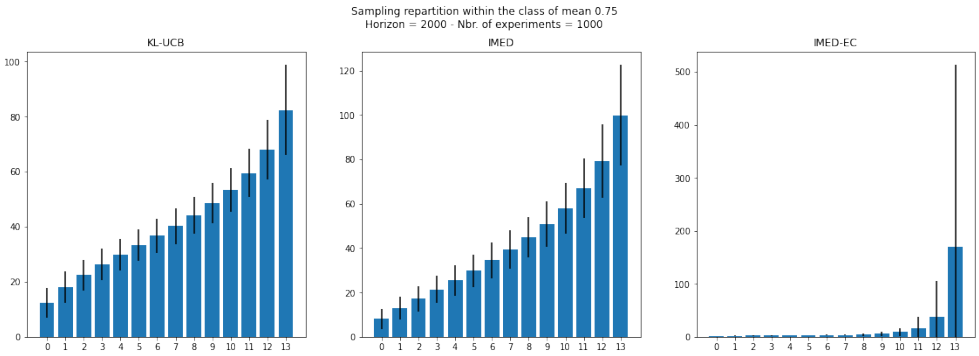

Figure 18: 7 classes - unbalanced - set of means = $\{0.1, 0.3, 0.4, 0.5, 0.6, \mathbf{0.75}, 0.9\}$ - mean 0.75

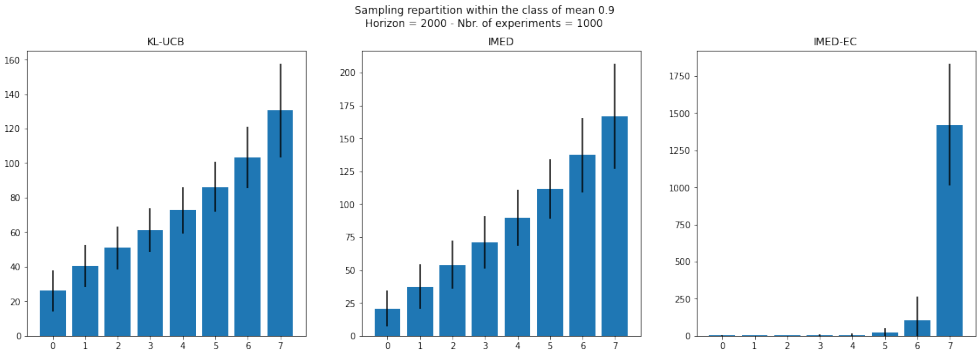

Figure 19: 7 classes - unbalanced - set of means = $\{0.1, 0.3, 0.4, 0.5, 0.6, 0.75, \mathbf{0.9}\}$ - mean 0.9