Bandits with Multimodal Structure

Hassan Saber, Odalric-Ambrym Maillard

Bandits with Multimodal Structure

Hassan Saber
hassan.saber@inria.fr
Univ. Lille, Inria, CNRS, Centrale Lille,
UMR 9189-CRIStAL, F-59000 Lille, France

Odalric-Ambrym Maillard
odalric.maillard@inria.fr
Univ. Lille, Inria, CNRS, Centrale Lille,
UMR 9189-CRIStAL, F-59000 Lille, France

Abstract

We consider a multi-armed bandit problem specified by a set of one-dimensional exponential family distributions endowed with a multimodal structure. The multimodal structure naturally extends the unimodal structure and appears to be underlying in quite interesting ways popular structures such as linear or Lipschitz bandits. We introduce `IMED-MB`, an algorithm that optimally exploits the multimodal structure, by adapting to this setting the popular Indexed Minimum Empirical Divergence (`IMED`) algorithm. We provide instance-dependent regret analysis of this strategy. Numerical experiments show that `IMED-MB` performs well in practice when assuming unimodal, polynomial or Lipschitz mean function.

1 Introduction

We consider a variant of the stochastic multi-armed bandit problem when reward distributions are single-parameter exponential families parameterized by their mean, and the mean, seen as a function of the arms, is assumed **multimodal** with a bounded number of modes. Multimodality being a qualitative rather than quantitative structural assumption (it involves comparison of arms), its study is of special interest to practitioners, complementing more quantitative assumptions such as Linearity or Lipschitz continuity that are more brittle or hard to check in practice. Multimodality is also appealing from a theoretical standpoint, as such structure presents non trivial challenges. Furthermore, multimodality naturally generalizes the unimodal structure and provides an appealing implicit view on the Lipschitz structure assumption (for which explicit knowledge of the Lipschitz constant is not always available in practice), that both received increasing attention in the recent years. This paper introduces, up to our knowledge, the first theoretical study of stochastic multi-armed bandits with multimodal mean structure, providing a novel algorithm together with both problem-dependent regret lower and upper bounds.

Structured bandits Following the now folklore terminology, by structure we mean that obtaining information about an arm may inform about another arm. This is mainly modeled by assuming the means satisfy very specific properties: for instance the means form a bell curve (unimodal bandits), the means are linearly dependent on a fixed number d of parameters (linear bandits, with d the dimension), the means are continuous and the gap between two consecutive arms is under control (Lipschitz bandits). The study of specific structured configuration sets has received increasing attention over the last few years, motivated by the growing popularity of bandits in a number of industrial and societal application domains. For instance, unimodal structure naturally appears in contexts such as single-peak preference economics, voting theory or wireless communications, and has been first considered in Yu and Mannor (2011) from a bandit perspective, then in Combes and Proutiere (2014); Trinh et al. (2020); Saber et al. (2021)

providing an explicit lower bound and corresponding algorithms. The linear bandit problem is also one typical illustration (Abbasi-Yadkori et al. (2011); Srinivas et al. (2010); Durand et al. (2017); Kveton et al. (2020)), see Lattimore and Szepesvari (2017) for a study of the lower bound (and Degenne et al. (2020a) for the related pure-exploration setup). Lipschitz bandits were studied in Magureanu et al. (2014); Wang et al. (2020); Lu et al. (2019). Bandits with groups of similar arms are studied in Pesquerel et al. (2021). On the theoretical side, these specific properties shape the means thus facilitating the location of the best arm, which translates into smaller regret achievable by optimal algorithms.

Multimodal structure In bandit problems where the goal is mainly to focus on the best arm, it is natural to consider a challenging setting with many local maximal means, which provides a natural motivation for the multimodal structure (formally introduced in Section 2). Multimodal structure has been considered in several places in the literature: Multimodal optimization problems (MMOPs) deal with optimatisation tasks that involve finding most of the locally (eventually globally) optimal solutions and possible approaches are approaches based on multi-armed bandits like in Agrawal et al. (2021). In dynamic pricing, when customers' sensitivity to prices varies heterogeneously over different price ranges, multimodality in the reward function is often observed, which is a common situation in practice, as mentionned in Wang et al. (2021) where an algorithm achieving optimal worst-case regret is proposed under multimodal reward function assumption. However, it appears that no instance-dependent bound has been suggested or exploited so far in the literature. Besides, the multimodal structure is underlying several structures of interest (see Section 2 for details), like the linear structure or the Lipschitz structure (defined below for completeness), that are more constraining and yield possibly computationally expensive strategies to be exploited optimally. Hence considering a multimodal structure can be seen as a relaxation of such problems, intermediate between considering no structure and a challenging one, and hence be appealing to the practitioner. We believe this provides a complementary perspective and motivation on exploiting multimodality in stochastic multi-armed bandits.

Structure adaptive strategies In Graves and Lai (1997) a generic algorithm was proposed to solve any structured bandit problems, with however prohibitive computational complexity. In Combes et al. (2017), the generic `OSSB` (Optimal Structured Stochastic Bandit) strategy is introduced, stepping the path towards generic structure-adaptation. Although asymptotically optimal, the algorithm comes with high computational cost. Inspired by combinatorial structures, a relaxation of the generic constrained optimization problem was proposed in Cuvelier et al. (2021), however at the price of trading-off regret optimality for computational efficiency. In Degenne et al. (2020b), the authors explore an adaptation of `KLUCB` algorithm to structured set of configurations. In Van Parys and Golrezaeiand (2020), the authors propose an approach based on convex duality. In Dong and Ma (2023), the authors develop a generic approach for both bandits and Markov Decision Processes. In all cases, the complexity of the lower bounds limit the practical efficiency of structure exploiting algorithms to small number of arms (say $|\mathcal{A}| \leqslant 500$).

In this article, we follow the rich literature focusing on regret minimization strategies targeting instance-dependent optimality in stochastic bandits. Another body or work focuses on proving asymptotic Bayesian optimality or also asymptotic minimax optimality in the worst-case setting rather than instance-dependent performance bounds, targeting order optimal rather than exact optimal regret bounds. This is the case for example in Kleinberg et al. (2008),Bubeck et al. (2008) and Foster et al. (2023) respectively introducing `ZOOMING`, `HOO` and `E2D`. In particular the provided bounds on the regret are not **instance-dependent** and instance-dependent optimality is not established for these algorithms. Such a worst-case setting is out of the scope of this paper.

Outline and contribution After providing the formal setup (Section 2, 3), we derive in Section 4 a regret lower bound for multi-armed bandits endowed with multimodal structure (Corollary 1). We show in particular that, due to the quantitative nature of the structure, the lower bound has an explicit form. This straightforwardly yields an algorithm exploiting this structure, `IMED-MB`, introduced in Section 5.1. We show in Theorem 2 that `IMED-MB` optimally exploits the structure when given the appropriate number of modes. The proof is non trivial and resort to a careful study of boundary crossing probabilities adapted to the small sample regime (Theorem 1), that is of independent interest. In Section 6, we report numerical experiments confirming the practical efficiency of `IMED-MB` even when the number of arms become large and illustrate the theoretical ratios between asymptotic optimal regrets depending on whether the Lipschitz structure or the multimodal one is considered.

2 Setup and notations

Stochastic multi-armed bandits A bandit instance is specified by a set of unknown probability distributions $\nu = (\nu_a)_{a\in\mathcal{A}}$, called a configuration, with means $(\mu_a(\nu))_{a\in\mathcal{A}}$. When there is no possible confusion, the means are simply denoted $(\mu_a)_{a\in\mathcal{A}}$. At each time $t \geqslant 1$, the learner chooses an arm $a_t \in \mathcal{A}$, based only on the past. The learner then receives and observes a reward $X_t \in [b\,;B]$, with $b, B \in \overline{\mathbb{R}}$, conditionally independent, sampled according to ν_{a_t}. The goal of the learner is to maximize the sum of rewards received over time (up to some unknown horizon T), or equivalently minimize the regret with respect to the algorithm constantly receiving the highest mean reward

$$R(\nu, T) = \mathbb{E}_\nu\left[\sum_{t=1}^{T} \mu^\star - X_t\right] \text{ where } \mu^\star = \max_{a\in\mathcal{A}} \mu_a\,.$$

Considering an horizon $T \geqslant 1$, thanks to the tower rule we can rewrite the regret as follows:

$$R(\nu, T) = \sum_{a\in\mathcal{A}} \Delta_a\, \mathbb{E}_\nu[N_a(T)]\,, \text{ with } \Delta_a = \mu^\star - \mu_a, \tag{1}$$

where $N_a(t) = \sum_{s=1}^{t} \mathbb{I}\{a_s = a\}$ is the number of pulls of arm a at time t. This problem received increased attention in the middle of the 20^{th} century, and the seminal paper Lai and Robbins (1985) established the first lower bound on the cumulative regret, showing that designing an algorithm that is optimal uniformly over a given set of configurations comes with a price : A lower bound on the regret can be explicited for consistent algorithms (Definition 1). The study of the lower performance bounds in multi-armed bandits successfully led to the development of asymptotically optimal algorithms for specific configuration sets, such as `KLUCB` algorithm Lai (1987); Cappé et al. (2013); Maillard (2018) for exponential families, or alternatively `DMED` and `IMED` algorithms from Honda and Takemura (2011; 2015). Other main approaches to optimally solve the stochastic bandit problem are Bayesian algorithm Thompson (1933) and algorithms based on re-sampling methods, such as SSMC from Chan (2020) or `RB-SDA` introduced in Baudry et al. (2020). Following e.g. Degenne et al. (2020b), we make the following simple parametric assumption on the reward distributions.

Assumption 1 (One-dimensional exponential family distributions)**.** *For all $\nu \in \mathcal{D}$, $\nu \subset \mathcal{P} := \{p(\mu), \mu \in \mathrm{I}\}$, where $p(\mu)$ is a regular canonical exponential-family distribution probability with parameter $\eta(\mu)$ and density $f(\cdot, \mu)$ with respect to some positive measure λ on $\mathbb{R}$ and mean $\mu \in \mathrm{I} \subset \mathbb{R}$. $f(\cdot, \mu)$ has the following shape:*

$$f(\cdot, \mu) : \quad x \mapsto h(x) \exp(\eta(\mu)\, T(x) - A(\mu))\,,$$

where $h \in \mathbb{R}_+^{\mathbb{R}}$, $T \in \mathbb{R}^{\mathbb{R}}$ and $A(\mu) = \log \int h(x) \exp(\eta(\mu)\, T(x))\, \lambda(\mathrm{d}x)$ are such that $|A(\mu)| < \infty$.

Remark 1. *Assumption 1 allows us to benefit from the pleasant monotonic properties of the Kullback-Leibler divergence for 1-dimensional exponential family distributions. Indeed, the lower bound on the regret (Section 4) shows that the Kullback-Leibler divergence plays a central role.*

Multimodal setting We assume there exists an undirected graph $G=(\mathcal{A},E)$ whose vertices are arms $\mathcal{A}$, and whose edges E modelize a proximity between the arms. G is assumed to be known to the learner. We denote by $\mathcal{V}_a=\{a'\neq a\colon (a,a')\in E\}$ the neighbours of arm $a\in\mathcal{A}$ in graph $G=(\mathcal{A},E)$ and by $\mathcal{A}_\nu^+=\{a\in\mathcal{A}\colon \forall a'\in\mathcal{V}_a, \mu'_a<\mu_a\}$ the set of arms with locally maximal means. When there is no possible confusion $\mathcal{A}_\nu^+$ is simply denoted $\mathcal{A}^+$. Intuitively, this graph-theoretic definition enables to capture not only multimodal functions on $\mathbb{R}$, for which $\mathcal{A}$ is totally ordered and E contains arms and their successor, but also on $\mathbb{R}^d$. We assume that $\nu\subset\mathcal{P}:=\{p(\mu),\mu\in\mathrm{I}\}$, where $p(\mu)$ is an exponential-family distribution probability with density $f(\cdot,\mu)$ with respect to some positive measure λ on $\mathbb{R}$ and mean $\mu\in\mathrm{I}\subset\mathbb{R}$. $\mathcal{P}$ is assumed to be known to the learner (Assumption 1). Thus, for all $a\in\mathcal{A}$ we have $\nu_a=p(\mu_a)$. We denote by $M^+=|\mathcal{A}_\nu^+|$,the size of subset $\mathcal{A}_\nu^+$. Importantly, we assume M^+ is unknown to the learner. For $\nu\subset\mathcal{P}$, we denote by $\mathcal{A}^\star(\nu)=\arg\max_{a\in\mathcal{A}}\mu_a$ the set of optimal arms of ν. When there is no possible confusion $\mathcal{A}^\star(\nu)$ is simply denoted $\mathcal{A}^\star$. We assume there exists $a^\star\in\mathcal{A}$ such that $\mathcal{A}^\star=\{a^\star\}$ (Assumption 2). In particular, we have

$$\{a^\star\}=\mathcal{A}^\star\subset\mathcal{A}^+\,. \tag{2}$$

Finally, we denote by $\mathcal{D}_{(\mathcal{P},G)}$ or $\mathcal{D}_{M^+}$ (or simply $\mathcal{D}$ when there is no confusion) the structured set of such multimodal-bandit distributions, and then $\mathcal{D}_{\leqslant M^+}=\bigcup_{M=1}^{M^+}\mathcal{D}_M$.

Assumption 2 (Unique maximums)**.** *We assume there exists $a_1,\dots,a_{M^+}\in\mathcal{A}$ such that $\mathcal{A}^+=\{a_1,\dots,a_{M^+}\}$ and $B>\mu_{a_1}>\ldots>\mu_{a_{M^+}}>\min_{a\notin\mathcal{A}^+}\mu_a>b$. In particular, $a_1=a^\star$ and $\mathcal{A}^\star=\{a^\star\}$.*

3 Multimodal and other structures

In this short section, we highlight some links between the multimodal structure and other well-studied structures. We show especially that several classical structures induce a multimodal structure with a natural control on M^+. Hence in such cases, exploiting multimodality can yield a reduced regret, intermediate between that of the unstructured and fully structured case.

Unimodal Structure The unimodal structure imposes by construction that $M^+=1$, then $\mathcal{A}^+=\mathcal{A}^\star=\{a^\star\}$. Hence the multimodal structure generalizes the unimodal structure from Combes and Proutiere (2014). Let us remind that the graph-theoretic definition enables to capture not only unimodality in dimension 1 (say $\mathcal{A}=\{1,\dots,\ell\}$ and $\mathcal{V}_{a^\star}=\{a^\star-1,a^\star+1\}$), but in higher dimension d as well, say $\mathcal{A}=\{1,\dots,\ell^d\}$, and $\mathcal{V}_{a^\star}=\{a^\star-\ell^k,a^\star+\ell^k\}_{k=0,\dots,d-1}$, which represents a discrete hypercube of width ℓ, with $E=\{(a,a')\colon |a-a'|\in\{1,\ell,\dots,\ell^{d-1}\}\}$.

Discretized linear Structure For $A\geqslant 1$, let $\mathcal{A}=[\![0,A-1]\!]$ index the discretisation of the space $\mathcal{X}=\left\{x_a=a/A\ ,a\in\mathcal{A}\right\}\subset[0,1]$, and $E=\{(a,a')\colon |a-a'|=1,\ a,a'\in\mathcal{A}\}$. Let us consider the linear function space $\mathcal{F}_\Theta=\left\{f_\theta : x\in\mathcal{X}\mapsto\theta^{\mathrm{T}}\varphi(x)\ ,\theta\in\Theta\right\}$ with parameter space $\Theta=\mathcal{B}(0,1)\subset\mathbb{R}^d$ of known dimension d and feature function $\varphi:\mathcal{X}\to\mathbb{R}^d$. The linear structure further assumes that there exists a parameter $\theta\in\Theta$ such that for all arm $a\in\mathcal{A}$, the mean of ν_a is $\mu_a=f_\theta(x_a)$. Now, considering e.g. the trigonometric polynomial feature function $\forall x\in\mathcal{X},\varphi(x)=(1,\cos(2\pi x),\sin(2\pi x),\dots,\cos(2\pi px),\sin(2\pi px))$, where $d=2p+1$, it can be

shown that ν belongs to a multinomial structured set $\mathcal{D}_{M^+}$, with $M^+ \leqslant p+1$ modes. Hence, the multimodal structure can be used to approximate a trigonometric polynomial structure.

Lipschitz Structure The multimodal structure can also be used to approximate a Lipschitz structure when $\mathcal{A} = [\![0, A-1]\!]$, $A > 1$, and $\mu : a \in \mathcal{A} \mapsto \mu_a$ is k-Lipschitz, where k is usually assumed to be known. In the following, we focus on the case when Lipschitz constant k is unknown and characterize the multimodality properties of an arbitrary Lipschitz configuration. We refer to Bubeck et al. (2011) for a study in the worse case scenario of Lipschitz bandits without the Lipschitz constant. For all $a, a' \in \mathcal{A}$, $|\mu_a - \mu_{a'}| \leqslant k\,|a-a'|$. In other words, there exists $(U_a)_{a\in\mathcal{A}} \subset [-1,1]$ such that for all $a \geqslant 1$, $\mu_a = \mu_0 + k\sum_{i=1}^{a} U_i$. To give an illustrative example, let's assume that $(U_a)_{a\in\mathcal{A}}$ are sampled from independent uniform distributions on $[-1,1]$. Then, μ can be seen as uniformly sampled in the set of k-Lipschitz functions on $\mathcal{A}$ with first term equal to μ_0. Considering neighbourhoods of the form $\mathcal{V}_a = \{a-1\,; a+1\} \cap \mathcal{A}$, the averaged number of arms with locally maximal means for uniformly sampled k-Lipschitz means is

$$E\Big[\,\big|\mathcal{A}_\nu^+\big|\,\Big] = 2 \times 0.5 + 0.25 \times (|\mathcal{A}| - 2) = 0.5 + 0.25\,|\mathcal{A}|\,. \tag{3}$$

Indeed, the probability of arm 0 and arm $A-1$ being local maximums is $P(A-1 \in \mathcal{A}_\nu^+) = P(0 \in \mathcal{A}_\nu^+) = P(\mu_0 \geqslant \mu_1) = P(U_1 \leqslant 0) = 0.5$, and for an arm $a \in \mathcal{A}$ such that $0 < a < A-1$, this probability is $P(a \in \mathcal{A}_\nu^+) = P(\mu_a \geqslant \mu_{a-1} \cap \mu_a \geqslant \mu_{a+1}) = P(U_a \geqslant 0 \cap U_{a+1} \leqslant 0) = 0.25$. Equation (3) then suggests the choice $M = \lceil 0.5 + 0.25\,|\mathcal{A}| \rceil$ as an estimation of M^+ for uniformly sampled k-Lipschitz means. We note that M does not depend on Lipschitz constant k but only on the number of arms.

4 Regret lower bound

In this section, we now introduce the instance-dependent lower bound on the regret of an algorithm. In order to obtain non trivial lower bound on the regret we consider algorithms that are consistent, in the classical sense (Hannan consistency), see e.g. Lai (1987):

Definition 1 (Consistent algorithm)**.** *An algorithm is consistent on the set $\mathcal{D}_{\leqslant M^+}$ of multimodal bandit configurations with at most M^+ local maximums if for all configuration $\nu \in \mathcal{D}_{\leqslant M^+}$, for all sub-optimal arm $a \notin \mathcal{A}^\star := \arg\max_{a\in\mathcal{A}} \mu_a$, for all $\alpha > 0$,* $\lim_{T\to 0} \mathbb{E}_\nu\left[\dfrac{N_a(T)}{T^\alpha}\right] = 0\,.$

In particular, for $\alpha = 1$, the number of pulls of a sub-optimal arm by a consistent algorithm is at most sub-linear in T, and actually polylogarithmic in T, considering $\alpha \to 0$.

We define for an arm $a \in \mathcal{A}$ its sub-optimality gap $\Delta_a = \mu^\star - \mu_a$ and denote by $\mathcal{V}_a$ its neighbourhood. We derive from the notion of consistency an asymptotic lower bound on the regret for multi-armed bandits endowed with a multimodal structure. Hereafter, we denote by $\mathrm{KL}(\mu|\mu') = \displaystyle\int_{\mathbb{R}} \log(f(x,\mu)/f(x,\mu'))f(x,\mu)\lambda(\mathrm{d}x)$ the Kullback-Leibler divergence between probability distribution $\nu = p(\mu)$ and $\nu' = p(\mu')$, for $\mu, \mu' \in \mathrm{I}$. The first key result is the following.

Proposition 1 (Lower bounds on the numbers of pulls)**.** *Let us consider a consistent algorithm on $\mathcal{D}_{\leqslant M^+}$ and a configuration $\nu \in \mathcal{D}_{M^+}$. Then it must be that for all arm $a \in \mathcal{A}_\nu^+ \cup \mathcal{V}_{\mathcal{A}_\nu^+}$,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\log(T)} \geqslant \frac{1}{\mathrm{KL}(\mu_a|\mu^\star)}\,.$$

Now for a configuration $\nu \in \mathcal{D}_{\leqslant M^+-1}$, it must be that for all arm $a \neq a^\star$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\log(T)} \geqslant \frac{1}{\mathrm{KL}(\mu_a|\mu^\star)}\,.$$

Corollary 1 (Lower bound on the regret). *Let us consider a consistent algorithm on $\mathcal{D}_{\leqslant M^+}$. Let $\nu \in \mathcal{D}_{\leqslant M^+}$. Then it must be that*

$$\liminf_{T\to\infty} \frac{R(\nu,T)}{\log(T)} \geqslant \begin{cases} \mathfrak{C}(\mu) \coloneqq \sum_{a^+\in\mathcal{A}^+} \sum_{\substack{a\in\{a^+\}\cup\mathcal{V}_{a^+} \\ \mu_a\neq\mu^\star}} \frac{\Delta_a}{\mathrm{KL}(\mu_a|\mu^\star)}\,. & \text{if } \nu\in\mathcal{D}_{M^+}\,, \\ \mathfrak{C}_0(\mu) \coloneqq \sum_{a\neq a^\star} \frac{\Delta_a}{\mathrm{KL}(\mu_a|\mu^\star)} & \text{if } \nu\in\mathcal{D}_{\leqslant M^+-1}\,. \end{cases}$$

The proof of Proposition 1 is provided in Appendix B. It is obtained by classical arguments for structured bandits, resorting to a change of measure argument and appropriate identification of a confusing bandit configuration in the multimodal structure. We refer the reader to Combes et al. (2017) for generic lower bounds on the regret that are explicited for several structures other than the multimodal one.

From this lower bound on the regret, an algorithm is considered (asymptotically) optimal on $\mathcal{D}_{M^+}$, if for all configuration $\nu\in\mathcal{D}_{M^+}$ with means μ, $\limsup_{T\to\infty} \frac{R(\nu,T)}{\log(T)} \leqslant \mathfrak{C}(\mu)$.

Remark 2 (Explicit complexities). *We note that the quantity $\mathfrak{C}(\mu)$ and $\mathfrak{C}_0(\mu)$ are fully explicit functions of μ (it does not require solving any optimization problem) for single-parameter exponential families. This useful property may not longer hold in general for arbitrary structures. Further, for Bernoulli distributions, a possible setting is to assume $\lambda = \delta_0 + \delta_1$ (with δ_0, δ_1 Dirac measures), $\mathrm{I} = (0,1)$ and for $\mu\in(0,1)$, $f(\cdot,\mu) =: x\in\{0,1\} \mapsto \mu^x(1-\mu)^{1-x}$. Then for all $\mu,\mu'\in(0,1)$, $\mathrm{KL}(\mu|\mu') = \mu\log(\mu/\mu') + (1-\mu)\log((1-\mu)/(1-\mu'))$. For Gaussian distributions (variance $\sigma^2=1$), we assume λ to be the Lebesgue measure, $\mathrm{I}=\mathbb{R}$, and for $\mu\in\mathbb{R}$, $f(\cdot,\mu) =: x\in\mathbb{R} \mapsto (\sqrt{2\pi})^{-1}e^{-(x-\mu)^2/2}$. Then for all $\mu,\mu'\in\mathbb{R}$, $\mathrm{KL}(\mu|\mu') = (\mu'-\mu)^2/2$. For Exponential distributions, we assume λ to be the Lebesgue measure, $\mathrm{I}=]0\,;+\infty[$, and for $\mu>0$, $f(\cdot,\mu) =: x>0 \mapsto e^{-x/\mu}/\mu$. Then for all $\mu,\mu'>0$, $\mathrm{KL}(\mu|\mu') = \log(\mu'/\mu)+\mu/\mu'-1$.*

Remark 3 (Tight lower bound). *Corollary 1 does not ensure that the stated lower bound on the regret is tight. This is a consequence of Theorem 2 which ensures that there exists an algorithm (`IMED-MB`) able to reach this lower bound. It is noticeable that $\mathfrak{C}(\mu)$ does not involve all the sub-optimal arms but only the ones in $\cup_{a^+\in\mathcal{A}^+}\{a^+\}\cup\mathcal{V}_{a^+}$. This indicates that sub-optimal arms outside of this set are sampled $o(\log(T))$ times, which contrasts with the unstructured stochastic multi-armed bandits.*

5 Optimal algorithm for multimodal bandits

We start this section by introducing some convenient notations and discussing what can be suitable for an optimal algorithm before introducing and defining the `IMED-MB` strategy.

Notations The empirical mean of the rewards from the arm a is denoted by $\widehat{\mu}_a(t) = \sum_{s=1}^t \mathbb{I}_{\{a_s=a\}} X_s / N_a(t)$ if $N_a(t)>0$, 0 otherwise. We also denote by $\widehat{\mu}^\star(t) = \max_{a\in\mathcal{A}} \widehat{\mu}_a(t)$ and $\widehat{\mathcal{A}}^\star(t) = \arg\max_{a\in\mathcal{A}} \widehat{\mu}_a(t)$ respectively the current best mean and the current set of optimal arms. We denote by $\widehat{a}^\star_t$ an arm arbitrarily chosen in $\widehat{\mathcal{A}}^\star(t)$. We denote by $\widehat{\mathcal{A}}^+(t) \coloneqq \{a\in\mathcal{A}\colon \forall a'\in\mathcal{V}_a, \widehat{\mu}_{a'}(t)\leqslant\widehat{\mu}_a(t)\}$ the set of arms with locally maximal empirical means. For all subset of arms $\mathcal{A}'\subset\mathcal{A}$, we denote by $\mathcal{V}_{\mathcal{A}'} \coloneqq \cup_{a\in\mathcal{A}'}\mathcal{V}_a$ the set of neighbours of arms in $\mathcal{A}'$. We recall that $M^+ = |\mathcal{A}^+|$ is not assumed to be known by the learner. In practice, the learner considers a positive integer $M\geqslant 1$ playing the role of M^+. We will see the situation differs when $M\geqslant M^+$ and $M\leqslant M^+$.

5.1 The IMED-MB algorithm

Let us consider a non-decreasing function $\Phi\colon n\in\mathbb{N}\mapsto\Phi(n)\in[0,\ \infty]$. When $\Phi(0)=\infty$, we simply write $\Phi\equiv\infty$. For all arms $a,a'\in\mathcal{A}$ at time step $t\geqslant 1$, in order to test the inequality $\mu_a<\widehat{\mu}_{a'}(t)$, we first introduce the dynamic quantity

$$I^{\Phi}_{a,a'}(t)=\begin{cases} N_a(t)\left(\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}_{a'}(t))\wedge\Phi(N_a(t))\right)+\log(N_a(t)) & \text{, if } \widehat{\mu}_a(t)<\widehat{\mu}_{a'}(t)\\ \log(N_a(t)) & \text{, otherwise,}\end{cases} \tag{4}$$

where $\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}_{a'}(t))\wedge\Phi(N_a(t))=\min\left\{\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}_{a'}(t))\ ;\Phi(N_a(t))\right\}$, and with the convention $0\times\infty=0$ and $\log(0)=-\infty$. We note that this quantity potentially increases when we pull arm a. In our understanding, the greater this quantity, the more plausible the inequality $\mu_a<\widehat{\mu}_{a'}(t)$ is. This understanding is mainly base on Theorem 1 and the well-known monotonic properties of the Kullback-Leibler divergence when assuming one-dimensional exponential family distributions. The term $\Phi(N_a(t))$ is introduced to control the term $\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}_{a'}(t))$ when current mean $\widehat{\mu}_a(t)$ is much smaller that μ_a (which may occur when $N_a(t)$ is small). Furthermore, for a current optimal arm $\widehat{a}^\star\in\widehat{\mathcal{A}}(t)$, we simply have $I^{\Phi}_{\widehat{a}^\star,a'}(t)=\log\big(N_{\widehat{a}^\star}(t)\big)$ and $I^{\Phi}_{a,\widehat{a}^\star}(t)=I^{\Phi}_a(t)$, with

$$I^{\Phi}_a(t)=N_a(t)\ \min\{\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t))\,,\Phi(N_a(t))\}+\log(N_a(t))\,. \tag{5}$$

Note that $(I^{\infty}_a(t))$ are the IMED index from Honda and Takemura (2015). Thus, we abusively refer to $\big(I^{\Phi}_a(t)\big)$ as IMED indexes and simply denote $I^{\infty}_{a,a'}(t),I^{\infty}_a(t)$ as $I_{a,a'}(t),I_a(t)$. We have in particular,

$$I^{\Phi}_{a,a'}(t)\leqslant I_{a,a'}(t)\,,\qquad I^{\Phi}_a(t)\leqslant I_a(t)\,.$$

We remind Indexed Minimum Empirical Divergence (IMED) is a bandit algorithm that has been proven optimal for both the unstructured case (Honda and Takemura (2015)) and the unimodal structure (Saber et al. (2021)).

No structure exploitation Following IMED algorithm (for unstructured bandits), one would naturally pull, at time step t, arm $a_{t+1}=\overline{a}_t$, the arm with minimal IMED index

$$\overline{a}_t\in\arg\min\{I^{\Phi}_a(t):a\in\mathcal{A}\}\quad\text{(arbitrarily chosen)}. \tag{6}$$

The shape of IMED indexes ensures that $\log(N_{\overline{a}_t}(t))\leqslant I^{\Phi}_{\overline{a}_t}(t)\leqslant I^{\Phi}_{\widehat{a}^\star}(t)=\log\big(N_{\widehat{a}^\star}(t)\big)$, which implies

$$N_{\overline{a}_t}(t)\leqslant N_{\widehat{a}^\star}(t),\qquad\forall\widehat{a}^\star\in\widehat{\mathcal{A}}^\star(t)\,. \tag{7}$$

Given Proposition 2, by pulling arm $\overline{a}_t$ at each time step t, IMED ensures that the current optimal arms in $\widehat{\mathcal{A}}^\star(t)$ are generally well estimated. Thus, IMED can be interpreted as firstly, properly estimating the mean of current optimal arm $\widehat{a}^\star_t$ (in other words, making sure that $\widehat{\mu}^\star(t)=\widehat{\mu}_{\widehat{a}^\star_t}(t)$ gets closer to $\mu_{\widehat{a}^\star_t}$), secondly, efficiently testing the inequalities $\mu_a<\widehat{\mu}^\star(t)$. Interestingly, a similar approach could be used to test $\mu_a<\widehat{\mu}_{\widehat{a}^+}(t)$ by using $I^{\Phi}_{a,\widehat{a}^+}(t)$ quantities for $\widehat{a}^+\in\widehat{\mathcal{A}}^+(t),a\in\mathcal{V}_{\widehat{a}^+}$.

Structure exploitation If the multimodal structure is not considered, arm $\overline{a}_t$ with minimal IMED index may be seen as the current most informative arm. However, regarding the lower bound on the regret for multimodal structure (Corollary 1), the current most informative arm should rather be

$$\overline{\overline{a}}_t\in\arg\min\{I^{\Phi}_a(t):a\in\widehat{\mathcal{A}}^+(t)\cup\mathcal{V}_{\widehat{\mathcal{A}}^+(t)}\}\quad\text{(arbitrarily chosen)}\,, \tag{8}$$

where $\widehat{\mathcal{A}}^+(t)$ is the set of arms with locally maximal empirical means, truncated at the M largest locally maximal empirical means (In particular $|\widehat{\mathcal{A}}^+(t)|\leqslant M$). For convenience, we introduce

$$\widehat{\mathcal{A}}^M(t)=\widehat{\mathcal{A}}^+(t)\cup\mathcal{V}_{\widehat{\mathcal{A}}^+(t)}\,. \tag{9}$$

Structure exploitation plus second-order exploration In order to minimize the unwanted effects from a bad identification of locally optimal arms (when $\widehat{\mathcal{A}}^+(t) \neq \mathcal{A}^+$), we allow a second-order exploration outside of $\widehat{\mathcal{A}}^M(t)$. This motivates the introduction of the following structured indexes for arm $a \in \mathcal{A}$,

$$I_a^M(t) = \begin{cases} I_a^\Phi(t) & \text{, if } a \in \widehat{\mathcal{A}}^M(t) \\ \Psi\big(I_a^\Phi(t)\big) & \text{, otherwise,} \end{cases} \tag{10}$$

where Ψ is an increasing function such that $x \leqslant \Psi(x)$ for $x \in \mathbb{R}$, and the associated arm with minimum index,

$$a_t^M \in \arg\min\{I_a^M(t) : a \in \mathcal{A}\}\,. \tag{11}$$

In particular, $a_t^M = \overline{\overline{a}}_t$ if $I^M_{a_t^M}(t) = I^\Phi_{\overline{\overline{a}}_t}(t)$, $a_t^M = \overline{a}_t$ otherwise[1].

Algorithm 1 `IMED-MB`

Input graph G, positive integer M, functions Φ, Ψ
Pull arbitrarily $a_1 \in \mathcal{A}$
for $t = 1 \dots T-1$ **do**
  if $\left|\widehat{\mathcal{A}}^+(t)\right| < M$ **then** ▷▷▷ `NO STRUCTURE EXPLOITATION`
    Pull $a_{t+1} = \overline{a}_t$ (Eq. (6))
  else ▷▷▷ `STRUCTURE EXPLOITATION`
    Pull $a_{t+1} = a_t^M$ (Eq. (11))
  end if
end for

The `IMED-MB` algorithm We finally define `IMED-MB` as follows: if $\left|\widehat{\mathcal{A}}^+(t)\right| = M$, it exploits the multimodal structure while allowing second-order exploration outside $\widehat{\mathcal{A}}^M(t)$, that is, pulling arm a_t^M with minimum **structured** index. Otherwise, $\left|\widehat{\mathcal{A}}^+(t)\right| < M$ and `IMED-MB` simply pulls arm $\overline{a}_t$ with minimal `IMED` index. `IMED-MB` algorithm is summarized in Algorithm 1.

5.2 Well-designed concentration of measurement

In order to provide a regret analysis, one challenge is to ensure that `IMED-MB` does not confused a sub-optimal but locally optimal arm with the best arm during exploitation phases. Intuitively, such challenge does not appear when $M^+ = 1$ because the structure is then unimodal and the best arm is the unique arm with both globally and locally maximal mean. We solve this challenge by proposing a regret analysis in two distinct stages. We first provide (in Appendix C.3) upper bounds on the numbers of pulls of sub-optimal arms that are not locally optimal. Then, we benefit from the following inequalities, $I_a^\Phi(t) \leqslant N_a(t)\,\Phi(N_a(t)) + \log(N_a(t))$ for $a \in \mathcal{A}$, to upper bound (in Appendix C.4) the numbers of pulls of locally optimal arms.

Furthermore, this proof process in two stages requires refined concentration of the empirical means to ensure `IMED-MB` is asymptotically optimal. Interestingly enough, the introduction of function Φ also guaranties stronger control of the ε-deviation from below of empirical mean $\widehat{\mu}_a(t)$ when $N_a(t)$, the number of pulls arm $a \in \mathcal{A}$, is relatively small. This is explained by additional term $\exp(-m_n\,\mathrm{KL}(\mu_a - \varepsilon|\mu_a))$ in the right side of the concentration inequality of Theorem 1

[1] Indeed, if $a_t^M \notin \widehat{\mathcal{A}}^M(t)$ then for $a \notin \widehat{\mathcal{A}}^M(t)$, $\Psi\left(I^\Phi_{a^M_t}(t)\right) \leqslant \Psi\big(I_a^\Phi(t)\big)$ and $I^\Phi_{a^M_t}(t) \leqslant I_a^\Phi(t)$, while for $a \in \widehat{\mathcal{A}}^M(t)$, $I^\Phi_{a^M_t}(t) \leqslant \Psi\left(I^\Phi_{a^M_t}\right) \leqslant I_a^\Phi(t)$. This implies that $\arg\min_{a\in\mathcal{A}} I_a^M(t) \cap \arg\min_{a\in\mathcal{A}} I_a^\Phi(t) \neq \emptyset$ when $\arg\min_{a\in\mathcal{A}} I_a^M(t) \cap \widehat{\mathcal{A}}^M(t) = \emptyset$.

below, where $m_n = 1 \wedge \dfrac{\log(n) - \log\log(n)}{\Phi(\log(n))}$ crucially depends on Φ: Without function Φ, which is equivalent to $\Phi \equiv \infty$, one would get $m_n = 1$ hence no refined concentration. Now by classic time-uniform concentration (Proposition 2 in Appendix), the ε-deviation from below of empirical mean $\widehat{\mu}_a(t)$ is under control when the number of pulls of arm a is greater than $f_{a,\varepsilon}(n) \coloneqq (\log(n) + 2\log\log(n))\,/\mathrm{KL}(\mu_a - \varepsilon|\mu_a)$, for $n \geqslant 3$. More precisely,

$$\mathbb{P}_\nu\Big(\exists t \geqslant 1, \{N_a(t) \geqslant f_{a,\varepsilon}(n)\} \cap \{\widehat{\mu}_a(t) < \mu_a - \varepsilon\}\Big) \leqslant \frac{1}{n\log^2(n)}\,,$$

where $\sum\limits_{n\geqslant 3} \frac{1}{n\log^2(n)} < \infty$. This is the reason why Theorem 1 focuses on the regime $N_a(t) \leqslant f_{a,\varepsilon}(n)$ that corresponds to the case when estimation of means is little accurate.

Theorem 1 (Boundary crossing probabilities)**.** *Let Φ be non-negative non-decreasing function such that $\Phi(\log(n)) \geqslant 1$ for $n \geqslant 18$. For all arm $a \in \mathcal{A}$, $\varepsilon > 0$, $n \geqslant 18$ such that $n \geqslant e\,M_n$, we have*

$$\mathbb{P}_\nu\left(\begin{array}{l}\exists t \geqslant 1, \{\widehat{\mu}_a(t) < \mu_a - \varepsilon\} \cap \{1 \leqslant N_a(t) \leqslant M_n\} \cap \\ \quad \{N_a(t)(\mathrm{KL}(\widehat{\mu}_a(t)|\mu_a - \varepsilon) \wedge \Phi(N_a(t))) + \log(N_a(t)) \geqslant \log(n)\}\end{array}\right)$$

$$\leqslant \quad \mathbb{I}_{\{m_n \leqslant M_n\}} e\,(1 + \log(M_n/m_n)\log(n/M_n))\,M_n\,n^{-1}\exp(-m_n\,\mathrm{KL}(\mu_a - \varepsilon|\mu_a))\,,$$

where $m_n = 1 \wedge \dfrac{\log(n) - \log\log(n)}{\Phi(\log(n))}$ and $M_n = f_{a,\varepsilon}(n) \coloneqq \dfrac{\log(n) + 2\log\log(n)}{\mathrm{KL}(\mu_a - \varepsilon|\mu_a)}$.

A proof of Theorem 1 is provided in Appendix F.

Remark 4. *Stronger control of the deviations of the empirical means are generally obtain by considering, for $\xi > 0$, $\log(\cdot) + \xi\log\log(\cdot)$ exploration terms in the indexes instead of more intuitive $\log(\cdot)$ exploration terms (the latter being known for providing better performance in practice), where ξ can be large to provide theoretical guaranties for structured bandit algorithms (for instance in Magureanu et al. (2014), ξ is set equal to $3\,|\mathcal{A}| + 1$). Thus, Theorem 1 provides an interesting alternative to (at least theoretically) speed up the concentration of empirical means without additional $\log\log(\cdot)$ exploration terms.*

5.3 Asymptotic optimality of `IMED-MB` algorithm

We precise the conditions of asymptotic optimality under `IMED-MB` algorithm in Theorem 2. We show that the lower bound on the regret from Corollary 1 is reached under `IMED-MB` algorithm, which proves both this lower bound is tight and `IMED-MB` is asymptotically optimal.

Theorem 2 (Asymptotic optimality)**.** *Let us consider a configuration $\nu \in \mathcal{D}_{M^+}$ such that $|\mathcal{A}_\nu^+| = M^+$ with means μ. Let us consider functions Φ and Ψ such that $1 \leqslant \Phi(\log(n)) \leqslant \log\log(n)$, for $n \geqslant 18$, and $\Psi(x) \geqslant \max\{x\,; \exp(x^\alpha)\}$, for $x \geqslant 0$ and some fixed constant $\alpha > 1$. Then, for any $M \geqslant 1$ (even if $M \neq M^+$), under `IMED-MB` algorithm,*

$\star$ if $M \geqslant M+$,

$$\forall a \neq a^\star, \quad \limsup_{T\to\infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\log(T)} \leqslant \frac{1}{\mathrm{KL}(\mu_a|\mu^\star)}\,,$$

$\star$ if $M \leqslant M+$,

$$\forall a \notin \mathcal{A}_\nu^+ \cup \mathcal{V}_{\mathcal{A}_\nu^+},\ \limsup_{T\to\infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\log(T)} \leqslant 0\,, \qquad \forall a \in \mathcal{V}_{\mathcal{A}_\nu^+},\ \limsup_{T\to\infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\log(T)} \leqslant \frac{1}{\mathrm{KL}(\mu_a|\mu^\star)}\,.$$

In particular, under `IMED-MB` algorithm,

$$\limsup_{T\to\infty} \frac{R(\nu,T)}{\log(T)} \leqslant \begin{cases} \mathfrak{C}(\mu) = \sum_{a^+\in\mathcal{A}^+} \sum_{\substack{a\in\{a^+\}\cup\mathcal{V}_{a^+}\\ \mu_a\neq\mu^\star}} \frac{\Delta_a}{\mathrm{KL}(\mu_a|\mu^\star)} & \text{if } M = M^+, \\ \mathfrak{C}_0(\mu) = \sum_{a\neq a^\star} \frac{\Delta_a}{\mathrm{KL}(\mu_a|\mu^\star)} & \text{if } M > M^+ . \end{cases}$$

A proof of Theorem 2 is provided in Appendix D, and a more precise finite time analysis is provided in Appendix C.

Handling of structure misidentifications When parameter M is not equal to the number of local maximums M^+, that is the proxy for the number of local maximums is imperfect, Theorem 2 shows that, when $M > M^+$, `IMED-MB` is never worse than the unstructured setting (and it is optimal when $M = M^+$), while when the number of local maximums is under estimated, that is $M < M^+$, the main risk is then to confuse a sub-optimal but locally optimal arm in $\mathcal{A}^+_\nu - \{a^\star\}$ with the best arm. We conjecture that second order exploration is crucial to avoid as best as possible such misidentifications by potentially revealing unexpected local maximums. A precise quantification of this phenomenon would be the subject of future work.

6 Numerical experiments

For all the experiments, we assume that for all arm $a \in \mathcal{A}$, ν_a is a Gaussian distribution with unknown mean $\mu_a \in \mathbb{R}$ and known variance $\sigma^2 = 0.25$. We assume $\mathcal{A} = [\![0\,;499]\!]$ and $\mathcal{V}_a = \{a-1\,;a+1\} \cap \mathcal{A}$, for $a \in \mathcal{A}$. All the regret curves are obtained from 10 runs. The deciles are represented with dotted lines. The horizon time is $T = 10^5$. At each run, each algorithm starts by pulling each arm once. `IMED-MB` is systematically compared to `KLUCB`. `IMED-MB`$(M = \dots)$ is `IMED-MB` algorithm with $\Psi \equiv \infty$ while `IMED-MB`$(M = \dots, \exp)$ is `IMED-MB` algorithm with $\Psi \equiv \exp(\cdot)$. We note that $\Psi \equiv \infty$ and $\Psi \equiv \exp(\cdot)$ are the two extreme functions for which `IMED-MB` algorithm is proven asymptotically optimal (Theorem 2). For these two choices of Ψ, `IMED-MB` seems to perform similarly in practice when the algorithm starts by pulling each arm once. For all the experiments, we set $\Phi \equiv 0 \vee \log(\cdot)$. We illustrate the performance of `IMED-MB` for the unimodal, the polynomial and the Lipschitz structures. We refer to Section 3 where the links with the multimodal structure are established.

Unimodal structure In Figure 1-(a), we compare `IMED-MB` algorithm to `OSUB` from Combes and Proutiere (2014), an optimal algorithm for unimodal bandits. One observes that `IMED-MB` shares similar practical performance with `OSUB`.

Polynomial structure In Figure 1-(b), we randomly generate a configuration μ with polynomial structure (left) such that $|\mathcal{A}^+_\nu| = 3$, that is, dimension $d = 2 \times (|\mathcal{A}^+_\nu| - 1) + 1 = 5$, and compare (right) `IMED-MB` algorithm to `LinUCB`, the popular algorithm for linear bandits.

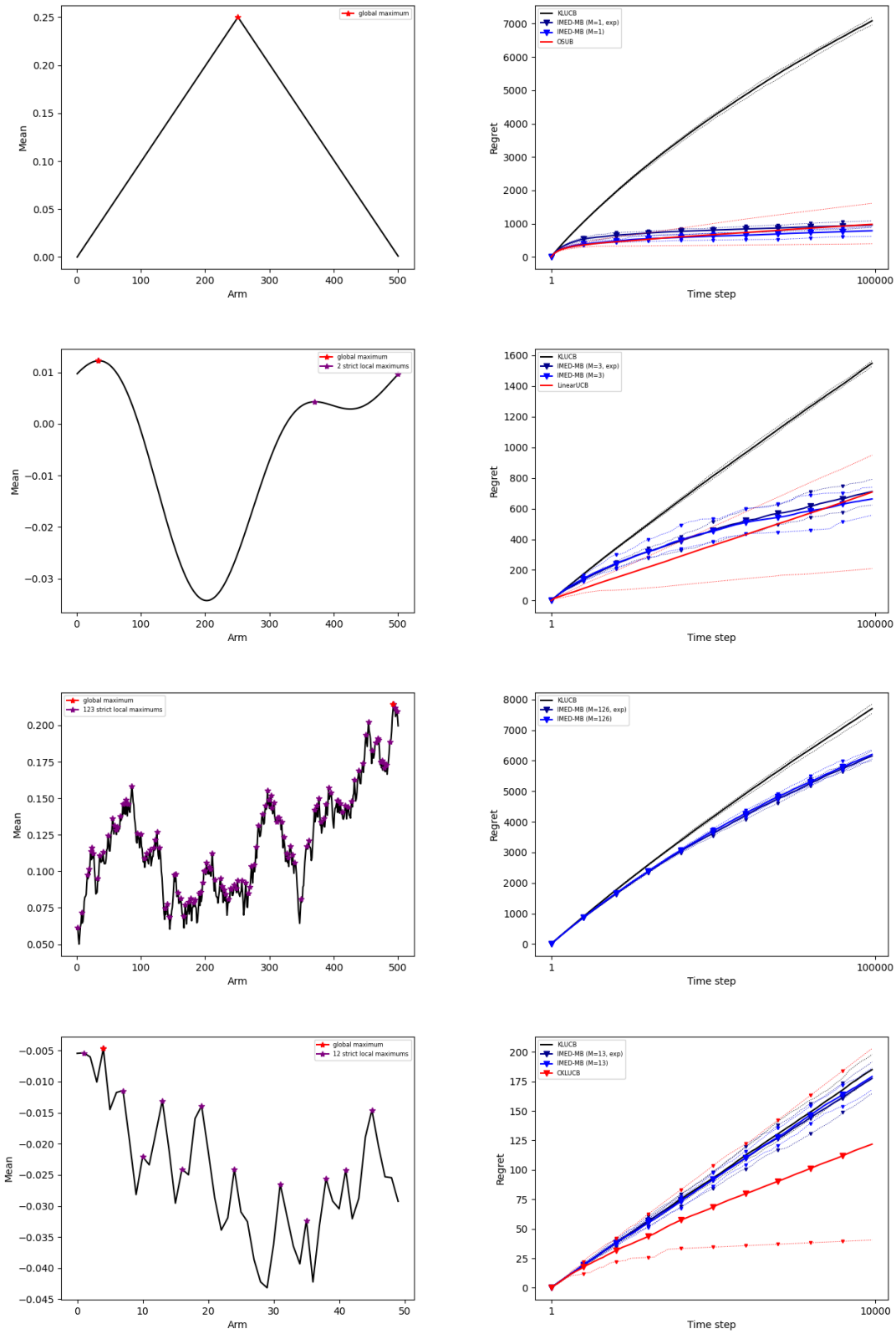

Figure 1: a) (Top) Unimodal structure, b) (Middle top) Polynomial structure, c) (Middle bottom) Lipschitz structure not knowing k nor M^+, d) (Bottom) Lipschitz structure knowing both k and M^+. Plot of cumulative regrets averaged over 10 runs, with mean and deciles.

Lipschitz structure We assume that $k = 0.01$. We sample $(U_a)_{a\in\mathcal{A}}$ from independent uniform distributions on $[-1\,;1]$, then set $\mu_0 = 0.1 \times U_0$ and, for $a \geqslant 1$, $\mu_a = \mu_{a-1} + k \times U_a$. In Figure 2, we represent with box-plots the number of local maximums for 1000 random configurations (left) and the corresponding ratios between asymptotic optimal regrets depending on the structure that is considered (right). We show in particular that, for such configurations, the ratio between the asymptotic optimal multimodal and Lipschitz regrets is, in average, approximately equal to 1.8. This is intuitive, since multimodal structure is less constraining than Lipschitz structure. These asymptotic optimal regrets are computed assuming both perfect knowledge of $\mathcal{A}_\nu^+$ and Lipschitz constant k. In Figure 1-(c), we illustrate the practical performance of `IMED-MB` for a particular random configuration. Its parameter M is set equal to $\lceil 0.5 + 0.25\,|\mathcal{A}| \rceil = 126$. In Figure 1-(d), we compare `IMED-MB` algorithm to `CKL-UCB` for smaller number of arms and smaller horizon (to limit calculation times). `CKL-UCB` is a bandit algorithm specific to the Lipschitz structure introduced in Magureanu et al. (2014). In this experiment (Figure 1-(d)), `IMED-MB` perfectly knows the numbers of local maximums and `CKL-UCB` perfectly knows the Lipschitz constant k.

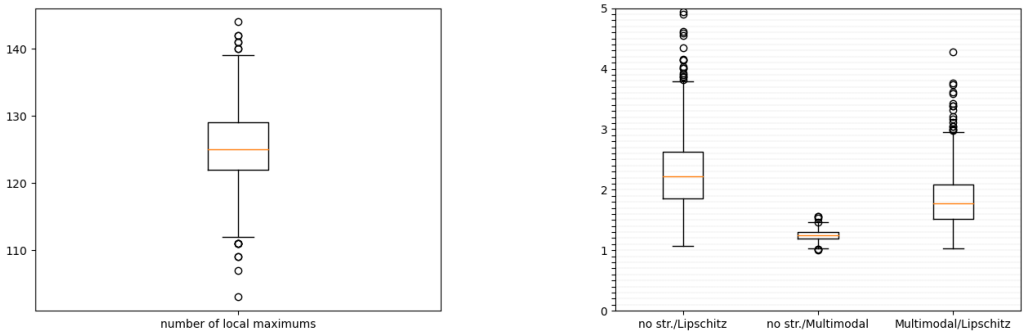

Figure 2: Number of local maximums for 1000 random Lipschitz configurations (left) with Lipschitz constant $k = 0.01$ and the corresponding ratios between asymptotic optimal regrets depending on the structure that is considered (right).

Conclusion We have considered the multimodal structure for stochastic multi-armed bandits and introduced `IMED-MB` algorithm, an adaptation of the `IMED` algorithm for the considered structure. We naturally discuss several situations depending on the knowledge on M^+, the number of modes of the means. When M^+ is assumed to be known to the learner, `IMED-MB` is proven to be asymptotically optimal according to the lower bound on the regret (Theorem 2). When M^+ is unknown, `IMED-MB` still seems to perform well in practice even only partial guaranties are provided in this case : `IMED-MB` algorithm may confuse a local maximum with the best arm when $M < M^+$ and interpolates with the unstructured setup when $M > M^+$. Our experiments show that an appropriate estimation M of M^+ can yield significantly better performance in finite time e.g. for the Lipschitz structure (Figure 1-(c)). The quantitative analysis of the phenomenon is the subject of future work. Finally, we point out that `IMED-MB` is a relatively simple algorithm, easy to implement[2] for common distributions (Remark 2), whose analysis in finite time is mainly based on simple algorithm-based empirical bounds (Appendix C.1) and a carefully-designed concentration tool (Theorem 1) of independent interest.

Acknowledgements This work has been supported by the French Ministry of Higher Education and Research, the Hauts-de-France region, Inria, the MEL, the I-Site ULNE regarding project RPILOTE-19-004-APPRENF, the Inria A.Ex. SR4SG project, and the Inria-Kyoto University Associate Team "RELIANT".

[2]All the code is made available here.

References

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

Agrawal, S., Tiwari, A., Naik, P., and Srivastava, A. (2021). Improved differential evolution based on multi-armed bandit for multimodal optimization problems. *Applied Intelligence*, pages 1–22.

Baudry, D., Kaufmann, E., and Maillard, O.-A. (2020). Sub-sampling for Efficient Non-Parametric Bandit Exploration. In *NeurIPS 2020*, Vancouver, Canada.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. (2008). Online optimization of X-armed bandits. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Proceedings of the 22nd conference on advances in Neural Information Processing Systems*, NIPS '08, Vancouver, British Columbia, Canada. MIT Press.

Bubeck, S., Stoltz, G., and Yu, J. Y. (2011). Lipschitz bandits without the lipschitz constant. In *Algorithmic Learning Theory: 22nd International Conference, ALT 2011, Espoo, Finland, October 5-7, 2011. Proceedings 22*, pages 144–158. Springer.

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

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

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

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

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

Degenne, R., Menard, P., Shang, X., and Valko, M. (2020a). Gamification of pure exploration for linear bandits. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2432–2442. PMLR.

Degenne, R., Shao, H., and Koolen, W. (2020b). Structure adaptive algorithms for stochastic bandits. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2443–2452. PMLR.

Dong, K. and Ma, T. (2023). Asymptotic instance-optimal algorithms for interactive decision making. *International Conference on Learning Representations (ICLR)*.

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

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. (2023). The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*.

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

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

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

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

Kleinberg, R. D., Slivkins, A., and Upfal, E. (2008). Multi-armed bandit problems in metric spaces. In *Proceedings of the 40th ACM symposium on Theory Of Computing*, TOC '08, pages 681–690.

Kveton, B., Zaheer, M., Szepesvari, C., Li, L., Ghavamzadeh, M., and Boutilier, C. (2020). Randomized exploration in generalized linear bandits. In Chiappa, S. and Calandra, R., 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.

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

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

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

Lu, S., Wang, G., Hu, Y., and Zhang, L. (2019). Optimal algorithms for Lipschitz bandits with heavy-tailed rewards. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4154–4163. PMLR.

Magureanu, S., Combes, R., and Proutière, A. (2014). Lipschitz bandits: Regret lower bounds and optimal algorithms. In *COLT 2014*.

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

Pesquerel, F., Saber, H., and Maillard, O.-A. (2021). Stochastic bandits with groups of similar arms. *International Conference on Neural Information Processing Systems (NeurIPS)*.

Saber, H., Ménard, P., and Maillard, O.-A. (2021). Indexed minimum empirical divergence for unimodal bandits. *International Conference on Neural Information Processing Systems (NeurIPS)*.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). 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.

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

Trinh, C., Kaufmann, E., Vernade, C., and Combes, R. (2020). Solving bernoulli rank-one bandits with unimodal thompson sampling. In *International Conference on Algorithmic Learning Theory*.

Van Parys, B. and Golrezaeiand, N. (2020). Optimal learning for structured bandits. *Sloan School of Management, MIT*.

Wang, T., Ye, W., Geng, D., and Rudin, C. (2020). Towards practical lipschitz bandits. *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*.

Wang, Y., Chen, B., and Simchi-Levi, D. (2021). Multimodal dynamic pricing. *Management Science*, 67(10):6136–6152.

Yu, J. Y. and Mannor, S. (2011). Unimodal bandits. In *ICML*, pages 41–48. Citeseer.

A Table of notation

T is the horizon time

$\mathcal{A}$ is the set of arms

$\mathcal{V}_a$ is the neighbourhood of arm a

ν is a configuration $(\nu_a)_{a\in\mathcal{A}}$ of one-dimensional exponential family distributions

$\mathcal{D}$ is the set of configurations ν, known to the learner

μ_a is the mean of distribution ν_a, unknown to the learner

$\mathcal{A}^+$ is the of arms with locally maximal means, unknown to the learner

M^+ is the number of maximums, possibly unknown to the learner

$\mathcal{D}_M$ is the set of configurations ν with M maximums

$\mathcal{D}_{\leqslant M}$ is the set of configurations ν with at most M maximums

$a^\star$ is the best arm, that is, the arm with maximal mean

$\mu^\star$ is the mean of distribution $\nu_{a^\star}$

Δ_a is the gap between the means of arm a and the best arm

ε_μ is a minimal gap defined in Equation (32)

k_μ is a minimal KL-gap defined in Equation (33)

$\mathrm{KL}(\mu|\mu')$ is the Kullback-Leibler divergence between configurations ν, ν' with means μ, μ'.

a_t is the arm pulled at time step t

X_t is the reward at time step t sampled from ν_{a_t}

b is a lower bound on the rewards (X_t)

B is an upper bound on the rewards (X_t)

$N_a(t)$ is the number of pulls of arm a at time step t

$\widehat{\mu}_a(t)$ is the empirical mean of arm a at time step t

$\widehat{\mu}^\star(t)$ is the maximal empirical mean at time step t

$\widehat{\mathcal{A}}^\star(t)$ is the set of arms with maximal empirical mean at time step t

$\widehat{a}_t^\star$ is an arm in $\widehat{\mathcal{A}}^\star(t)$ with maximal empirical mean at time step t

$\widehat{\mathcal{A}}^+(t)$ is the set of arms with locally maximal empirical means at time step t, truncated at the M largest locally maximal empirical means ($|\widehat{\mathcal{A}}^+(t)|\leqslant M$)

$\widehat{\mathcal{A}}^M(t)$ is the set of arms in $\widehat{\mathcal{A}}^+(t)$ or in their neighbourhoods $\mathcal{V}_{\widehat{\mathcal{A}}^+(t)}$

$x \wedge y$ is the minimum between x and y.

$x \vee y$ is the maximum between x and y.

Φ is a non-decreasing non-negative function such that $I^{\Phi}_{a,a'}(t) \leqslant N_a(t)\Phi(N_a(t)) + \log(N_a(t))$

Ψ is a non-decreasing non-negative function such that $x \leqslant \Psi(x)$

φ is the function : $n \geqslant 0 \mapsto \min\left\{n\,; n\,\dfrac{k_\mu \wedge \Phi(n)}{\mathrm{KL}(b|\mu^\star + \varepsilon_\mu) \wedge \Phi(n)}\right\}$, where for all $\mu' \in (b\,; B)$, $\mathrm{KL}(b|\mu') = \lim\limits_{\mu \to b} \mathrm{KL}(\mu|\mu')$

F is the function : $n \geqslant 0 \mapsto e^{n\Phi(n)+\log(n)}$

$I^{\Phi}_{a,a'}(t)$ is a dynamic quantity introduced in order to tests the inequality $\mu_a < \widehat{\mu}_{a'}(t)$

$I^{\Phi}_{a}(t)$ is equal to $I^{\Phi}_{a,\widehat{a}^\star_t}(t)$ and tests the inequality $\mu_a < \widehat{\mu}^\star(t)$

$I^{\infty}_{a}(t)$ denotes $I^{\Phi}_{a}(t)$ when $\Phi \equiv \infty$ and is equal to `IMED` index $I_a(t)$

$I_a(t)$ is the `IMED` index of arm a at time step t

$\overline{a}_t$ is an arm with minimal index $I^{\Phi}_{\overline{a}_t}(t)$ on $\mathcal{A}$

$\overline{\overline{a}}_t$ is an arm with minimal index $I^{\Phi}_{\overline{\overline{a}}_t}(t)$ on $\widehat{\mathcal{A}}^M(t)$

a^M_t is an arm with minimal structured index $I^{M}_{a^M_t}(t)$ on $\mathcal{A}$

B Proof of Proposition 1

Proof. Let us consider a sub-optimal arm $a \neq a^\star$. If $\nu \in \mathcal{D}_{M^+}$, we further assume that $a \in \mathcal{A}^+ \cup \mathcal{V}_{\mathcal{A}^+}$. The proof consists in used Lemma 1 below from Garivier et al. (2016) with configuration ν and the most confusing configuration $\nu^{(a)}(\varepsilon)$ for $\varepsilon > 0$, with means $\mu^{(a)}(\varepsilon)$, where

$$\forall a' \in \mathcal{A}, \quad \mu^{(a)}_{a'}(\varepsilon) = \begin{cases} \mu_{a'} & \text{if } a' \neq a \\ \mu^\star + \varepsilon & \text{if } a' = a\,. \end{cases} \tag{12}$$

Note that the set of optimal arms for the most confusing configuration $\nu^{(a)}$ reduces to the singleton $\mathcal{A}^\star\big(\nu^{(a)}\big) = \{a\}$ and that the most confusing configuration $\nu^{(a)}(\varepsilon)$ still belongs to $\bigcup_{M=1}^{M^+} \mathcal{D}_M$, that is $\mu^{(a)}(\varepsilon)$ also has at most M^+ local maximums. An illustration with an example is provided in Figure 3.

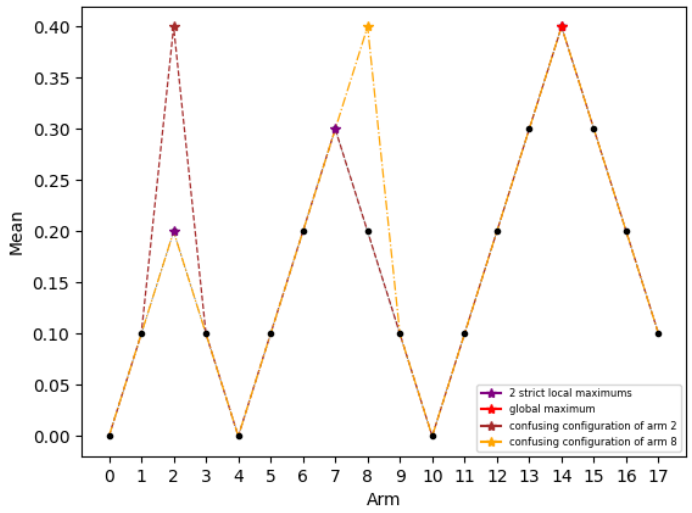

Figure 3: Illustration of confusing configurations for arms $2, 8 \in \mathcal{A}^+ \cup \mathcal{V}_{\mathcal{A}^+}$ when $M^+ = 3$, $\mathcal{A}^+ = \{2\,;7\;14\}$, $\mathcal{A} = [\![0\,;17]\!]$, and $\mathcal{V}_a = \{a-1\,;a+1\} \cap \mathcal{A}$, for $a \in \mathcal{A}$.

Let us consider the random variable $Z_T = N_a(T)/T \in [0,1]$. Then Lemma 1 below implies

$$\sum_{a' \in \mathcal{A}} \mathbb{E}_\nu[N_{a'}(T)]\, \mathrm{KL}(\mu_{a'} | \mu^{(a)}_{a'}(\varepsilon)) \geqslant \mathrm{kl}(\mathbb{E}_\nu[Z_T] | \mathbb{E}_{\nu^{(a)}(\varepsilon)}[Z_T])\,. \tag{13}$$

Since for all $a' \neq a$ we have the equality of means $\mu_{a'} = \mu^{(a)}_{a'}(\varepsilon)$ and since $\mu^{(a)}_a(\varepsilon) = \mu^\star + \varepsilon$, previous Equation (13) rewrites

$$\mathbb{E}_\nu[N_a(T)]\, \mathrm{KL}(\mu_a | \mu^\star + \varepsilon) \geqslant \mathrm{kl}(\mathbb{E}_\nu[Z_T] | \mathbb{E}_{\nu^{(a)}(\varepsilon)}[Z_T])\,. \tag{14}$$

From there, what remains of the proof is classic. For instance, the reader can refer to the proof of Theorem 1 in Garivier et al. (2016).

Since we consider a consistent algorithm on $\mathcal{D}_M$ and $\begin{cases} \nu \in \mathcal{D}_M \\ a \notin \mathcal{A}^\star(\nu) \end{cases}$, the averaged number of pulls of arm a for configuration ν is sub-linear and

$$\lim_{T\to\infty} \mathbb{E}_\nu[Z_T] = \lim_{T\to 0} \mathbb{E}_\nu[N_a(T)]/T = 0\,. \tag{15}$$

Since we consider a consistent algorithm on $\mathcal{D}_M$ and $\begin{cases} \nu^{(a)} \in \mathcal{D}_M \\ \{a\} = \mathcal{A}^\star(\nu^{(a)}(\varepsilon)) \end{cases}$, the averaged number of pulls of arm a for configuration $\nu^{(a)}$ is linear and

$$\lim_{T\to\infty} \mathbb{E}_{\nu^{(a)}(\varepsilon)}[Z_T] = \lim_{T\to 0} \mathbb{E}_{\nu^{(a)}(\varepsilon)}[N_a(T)]/T = 1\,. \tag{16}$$

By combining Equation (15) and (16), we have in particular when T tends to ∞ that

$$\mathrm{kl}(\mathbb{E}_\nu[Z_T]|\mathbb{E}_{\nu^{(a)}(\varepsilon)}[Z_T]) \underset{T\to\infty}{\sim} \log\left(\frac{1}{1-\mathbb{E}_{\nu^{(a)}(\varepsilon)}[Z_T]}\right)\,. \tag{17}$$

Note that the right term of the last equation can be rewritten as follows,

$$\log\left(\frac{1}{1-\mathbb{E}_{\nu^{(a)}(\varepsilon)}[Z_T]}\right) = \log\left(\frac{T}{\sum\limits_{a'\notin\mathcal{A}^\star(\nu^{(a)}(\varepsilon))} \mathbb{E}_{\nu^{(a)}(\varepsilon))}[N_{a'}(T)]}\right) = \log\left(\frac{T}{O(T^\alpha)}\right)\,, \quad \forall \alpha > 0\,. \tag{18}$$

In particular, by combining previous Equation (18) and Equation (17) we get the following asymptotic result,

$$\lim_{T\to\infty} \frac{\mathrm{kl}(\mathbb{E}_\nu[Z_T]|\mathbb{E}_{\nu^{(a)}(\varepsilon)}[Z_T])}{\log(T)} = 1\,. \tag{19}$$

We prove Proposition 1 by combining this last Equation (19) with Equation (14). □

Lemma 1 (Fundamental inequality)**.** *Let us consider a consistent algorithm on $\mathcal{D}$. Then for all configurations $\nu, \nu' \in \mathcal{D}$ with means $\mu, \mu' \in \mathrm{I}^{\mathcal{A}}$, for all horizon $T \geqslant 1$, for random variable Z_T with values in $[0,1]$,*

$$\sum_{a\in\mathcal{A}} \mathbb{E}_\nu[N_a(T)]\,\mathrm{KL}(\mu_a|\mu'_a) \geqslant \mathrm{kl}(\mathbb{E}_\nu[Z_T]|\mathbb{E}_{\nu'}[Z_T])\,,$$

where $\mathrm{kl}(p|q) = p\log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q})$ for $p, q \in [0,1]$.

C Finite time analysis

At a high level, the key interesting step of the proof is to realize that the considered algorithm implies empirical lower and empirical upper bounds on the numbers of pulls (Section C.1). Then, based on concentration tools (Theorem 1 and Proposition 2), the algorithm-based empirical lower bounds ensure the reliability of the estimators of interest (Section C.2). Then, combining the reliability of these estimators with the obtained algorithm-based empirical upper bounds, we firstly obtain (Section C.3) upper bounds on the average numbers of pulls for locally sub-optimal arms outside of $\mathcal{A}^+$, the set of local maximums. Then, we use these upper bounds and benefit from the fact that the structure is well-estimated during exploration phases when parameter $M \geqslant M^+$ to secondly obtain (Section C.4) upper bounds on the numbers of pulls of arms in $\mathcal{A}^+$. For clarity, several intermediate lemmas are presented with variants : *no structure exploitation* when $\left|\widehat{\mathcal{A}}^+(t)\right| < M$, *no second-order exploration* when $\left|\widehat{\mathcal{A}}^+(t)\right| = M$ and $a_{t+1} \notin \widehat{\mathcal{A}}^M(t)$, *second-order exploration* when $\left|\widehat{\mathcal{A}}^+(t)\right| = M$ and $a_{t+1} \in \widehat{\mathcal{A}}^M(t)$. This respects the structure of `IMED-MB` algorithm and simplifies its analysis at the price of appearing redundant.

C.1 Algorithm-based empirical bounds

`IMED-MB` algorithm implies inequalities between the indexes that can be rewritten as inequalities on the numbers of pulls. While lower bounds involving $\log(t)$ may be expected in view of the asymptotic regret bounds, we show lower bounds on the numbers of pulls involving instead $\log\big(N_{a_{t+1}}(t)\big)$, the logarithm of the number of pulls of the current chosen arm. We also provide upper bounds on $N_{a_{t+1}}(t)$ involving $\log(t)$.

Lemma 2 (Empirical lower bounds - no structure exploitation)**.** *Under `IMED-MB`, at each step time $t \geqslant 1$ such that $\left|\widehat{\mathcal{A}}^+(t)\right| < M$ (that is, when there is no structure exploitation), for all $a \in \mathcal{A} - \{\widehat{a}_t^\star\}$,*

$$\log\big(N_{a_{t+1}}(t)\big) \leqslant N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) \wedge \Phi(N_a(t)) + \log(N_a(t))\,, \tag{20}$$

$$N_{a_{t+1}}(t) \leqslant N_{\widehat{a}_t^\star}(t)\,, \tag{21}$$

and,

$$\min\left\{ N_{a_{t+1}}(t)\ ;\ N_{a_{t+1}}(t)\,\frac{\mathrm{KL}\big(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^\star(t)\big) \wedge \Phi\big(N_{a_{t+1}}(t)\big)}{\mathrm{KL}\big(b\,|\widehat{\mu}^\star(t)\big) \wedge \Phi\big(N_{a_{t+1}}(t)\big)} \right\} \leqslant N_a(t)\,, \tag{22}$$

where for all $\mu' \in (b\,;B)$, $\mathrm{KL}(b|\mu') = \lim\limits_{\mu \to b} \mathrm{KL}(\mu|\mu')$.

Proof. For $a \in \mathcal{A} - \{\widehat{a}_t^\star\}$, by definition, we have $I_a(t) = N_a(t)\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) + \log(N_a(t))$, hence

$$\log(N_a(t)) \leqslant I_a^\Phi(t) \leqslant I_a(t)\,.$$

This implies, since arm $\overline{a}_t$ with minimum index is pulled when $\left|\widehat{\mathcal{A}}^+(t)\right| < M$,

$$\log\big(N_{a_{t+1}}(t)\big) \leqslant I^\Phi_{a_{t+1}}(t) = \min_{a' \in \mathcal{A}} I^\Phi_{a'}(t) \leqslant I^\Phi_a(t) = N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) \wedge \Phi(N_a(t)) + \log(N_a(t))\,,$$

which proves Equation (20). Similarly, we have

$$\log\big(N_{a_{t+1}}(t)\big) \leqslant I^\Phi_{a_{t+1}}(t) = \min_{a' \in \mathcal{A}} I^\Phi_{a'}(t) \leqslant I^\Phi_{\widehat{a}_t^\star}(t) = \log\Big(N_{\widehat{a}_t^\star}(t)\Big)\,,$$

which implies in particular,

$$\log\bigl(N_{a_{t+1}}(t)\bigr) \leqslant \log\Bigl(N_{\widehat{a}_t^\star}(t)\Bigr)\,.$$

By taking the $\log^{-1}(\cdot)$, we prove Equation (21).

Furthermore, since arm $\overline{a}_t$ with minimum index is pulled when $\bigl|\widehat{\mathcal{A}}^+(t)\bigr| < M$,

$$\begin{aligned} & N_{a_{t+1}}(t)\,\mathrm{KL}\bigl(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^\star(t)\bigr) \wedge \Phi\bigl(N_{a_{t+1}}(t)\bigr) + \log\bigl(N_{a_{t+1}}(t)\bigr) \\ = \;& I^\Phi_{a_{t+1}}(t) \\ \leqslant \;& I^\Phi_a(t) \\ \leqslant \;& N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) \wedge \Phi(N_a(t)) + \log(N_a(t))\,. \end{aligned}$$

Since $\Phi(\cdot)$ and $\log(\cdot)$ are non-decreasing function either $N_{a_{t+1}}(t) \leqslant N_a(t)$, or $\mathrm{KL}\bigl(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^\star(t)\bigr) \leqslant \mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) \leqslant \mathrm{KL}(b\,|\widehat{\mu}^\star(t))$ (that is, $\widehat{\mu}^\star(t) \geqslant \widehat{\mu}_{a_{t+1}}(t) \geqslant \widehat{\mu}_a(t) \geqslant b$) and

$$N_{a_{t+1}}(t)\,\mathrm{KL}\bigl(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^\star(t)\bigr) \wedge \Phi\bigl(N_{a_{t+1}}(t)\bigr) \leqslant N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) \wedge \Phi(N_a(t))\,,$$

which implies

$$N_{a_{t+1}}(t)\,\frac{\mathrm{KL}\bigl(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^\star(t)\bigr) \wedge \Phi\bigl(N_{a_{t+1}}(t)\bigr)}{\mathrm{KL}(b\,|\widehat{\mu}^\star(t)) \wedge \Phi\bigl(N_{a_{t+1}}(t)\bigr)} \leqslant N_{a_{t+1}}(t)\,\frac{\mathrm{KL}\bigl(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^\star(t)\bigr) \wedge \Phi\bigl(N_{a_{t+1}}(t)\bigr)}{\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) \wedge \Phi(N_a(t))} \leqslant N_a(t)\,.$$

□

Lemma 3 (Empirical lower bounds - no second-order exploration)**.** *Under* `IMED-MB`*, at each step time $t \geqslant 1$ such that $\bigl|\widehat{\mathcal{A}}^+(t)\bigr| = M$ (that is, when there is structure exploitation) and $a_{t+1} \in \widehat{\mathcal{A}}^M(t)$ (that is, there is no second-order exploration), for all $a \in \widehat{\mathcal{A}}^M(t) - \{\widehat{a}_t^\star\}$,*

$$\log\bigl(N_{a_{t+1}}(t)\bigr) \leqslant N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) + \log(N_a(t))\,, \tag{23}$$

$$N_{a_{t+1}}(t) \leqslant N_{\widehat{a}_t^\star}(t)\,, \tag{24}$$

and,

$$\min\left\{ N_{a_{t+1}}(t)\;;\; N_{a_{t+1}}(t)\,\frac{\mathrm{KL}\bigl(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^\star(t)\bigr) \wedge \Phi\bigl(N_{a_{t+1}}(t)\bigr)}{\mathrm{KL}(b\,|\widehat{\mu}^\star(t)) \wedge \Phi\bigl(N_{a_{t+1}}(t)\bigr)} \right\} \leqslant N_a(t)\,, \tag{25}$$

where for all $\mu' \in (b\,;B)$, $\mathrm{KL}(b|\mu') = \lim\limits_{\mu \to b} \mathrm{KL}(\mu|\mu')$.

Proof. A proof is obtained from the proof of Lemma 2 by replacing $\mathcal{A}$ by $\widehat{\mathcal{A}}^M(t)$ and $\overline{a}_t$ by $a_t^M = \overline{\overline{a}}_t$ (that is, the arm with minimum index on $\mathcal{A}$ by the arm with minimum index on $\widehat{\mathcal{A}}^M(t)$). □

Lemma 4 (Empirical lower bounds - second-order exploration)**.** *Under* `IMED-MB`*, at each step time $t \geqslant 1$ such that $\bigl|\widehat{\mathcal{A}}^+(t)\bigr| = M$ (that is, when there is structure exploitation) and $a_{t+1} \notin \widehat{\mathcal{A}}^M(t)$ (that is, there is second-order exploration), for all $a \in \mathcal{A} - \{\widehat{a}_t^\star\}$,*

$$\log\bigl(N_{a_{t+1}}(t)\bigr) \leqslant N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) + \log(N_a(t))\,, \tag{26}$$

$$N_{a_{t+1}}(t) \leqslant N_{\widehat{a}_t^\star}(t)\,, \tag{27}$$

and,

$$\min\left\{ N_{a_{t+1}}(t)\ ;\ N_{a_{t+1}}(t)\,\frac{\mathrm{KL}\big(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^\star(t)\big)\wedge\Phi\big(N_{a_{t+1}}(t)\big)}{\mathrm{KL}\big(b\,|\widehat{\mu}^\star(t)\big)\wedge\Phi\big(N_{a_{t+1}}(t)\big)} \right\} \leqslant N_a(t)\,, \tag{28}$$

where for all $\mu'\in(b\,;B)$, $\mathrm{KL}(b|\mu') = \lim\limits_{\mu\to b}\mathrm{KL}(\mu|\mu')$.

Proof. A proof is obtained directly from the proof of Lemma 2 by noting that $a_t^M = \overline{a}_t$. In particular, for all $a\in\mathcal{A}$, $I^\Phi_{a_t^M}(t)\leqslant I^\Phi_a(t)$. □

Lemma 5 (Empirical upper bounds - no structure exploitation)**.** *Under* `IMED-MB` *at each step time* $t\geqslant 1$ *such that* $\left|\widehat{\mathcal{A}}^+(t)\right| < M$ *(that is, when there is no structure exploitation),*

$$N_{a_{t+1}}(t)\,\mathrm{KL}\big(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^\star(t)\big)\wedge\Phi\big(N_{a_{t+1}}(t)\big) \leqslant \log(t)\,. \tag{29}$$

Proof. From the definitions of the indexes, we have

$$I^\Phi_{a_{t+1}}(t) \leqslant I^\Phi_{\widehat{a}_t^\star}(t) \leqslant I_{\widehat{a}_t^\star}(t)\,.$$

It remains, to conclude, to note that

$$N_{a_{t+1}}(t)\,\min\big\{\mathrm{KL}\big(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^\star(t)\big)\,;\Phi\big(N_{a_{t+1}}(t)\big)\big\} \leqslant I^\Phi_{a_{t+1}}(t)\,,$$

and

$$I_{\widehat{a}_t^\star}(t) = \log(N_{\widehat{a}_t^\star}(t)) \leqslant \log(t)\,.$$

□

Lemma 6 (Empirical upper bounds - no second-order exploration)**.** *Under* `IMED-MB` *at each step time* $t\geqslant 1$ *such that* $\left|\widehat{\mathcal{A}}^+(t)\right| = M$ *(that is, when there is structure exploitation) and* $a_{t+1}\in\widehat{\mathcal{A}}^M(t)$ *(that is, there is no second-order exploration),*

$$N_{a_{t+1}}(t)\,\mathrm{KL}\big(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^\star(t)\big)\wedge\Phi\big(N_{a_{t+1}}(t)\big) \leqslant \log(t)\,. \tag{30}$$

Proof. The same proof as that of Lemma 5 holds. □

Lemma 7 (Empirical upper bounds - second-order exploration)**.** *Under* `IMED-MB` *at each step time* $t\geqslant 1$ *such that* $\left|\widehat{\mathcal{A}}^+(t)\right| = M$ *(that is, when there is structure exploitation) and* $a_{t+1}\notin\widehat{\mathcal{A}}^M(t)$ *(that is, there is second-order exploration),*

$$N_{a_{t+1}}(t)\,\mathrm{KL}\big(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^\star(t)\big)\wedge\Phi\big(N_{a_{t+1}}(t)\big) \leqslant \Psi^{-1}(\log(t))\ . \tag{31}$$

Proof. From the definitions of the indexes, we have

$$\Psi\Big(I^\Phi_{a_{t+1}}(t)\Big) \leqslant I^\Phi_{\widehat{a}_t^\star}(t) \leqslant I_{\widehat{a}_t^\star}(t)\,.$$

It remains, to conclude, to note that

$$N_{a_{t+1}}(t)\,\min\left\{\mathrm{KL}\big(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^\star(t)\big)\,;\Phi\big(N_{a_{t+1}}(t)\big)\right\}\leqslant I^{\Phi}_{a_{t+1}}(t)\,,$$

and

$$I_{\widehat{a}^\star_t}(t)=\log(N_{\widehat{a}^\star_t}(t))\leqslant\log(t)\,.$$

□

C.2 Well-estimated means and structure

Before going further in the analysis, we inform the reader that sets $\mathcal{E}_a(\varepsilon)$, $\mathcal{E}^-_a(f,\varepsilon)$, $\mathcal{K}^-_a(\varepsilon)$ for $a\in\mathcal{A}$, f a function, $\varepsilon>0$, used in this subsection are introduced and studied in Section E. We further introduce the following notations before presenting the conditions of reliability of our estimators.

$$\varepsilon_\mu=\frac{1}{3}\left(\min_{a\neq a^\star}\mu^\star-\mu_a\right)\wedge\left(\min_{a^+\in\mathcal{A}^+}\max_{a\in\mathcal{V}_{a^+}}\mu_{a^+}-\mu_a\right)\wedge(B-\mu^\star)\wedge\left(\min_{a\in\mathcal{A}}\mu_a-b\right)\tag{32}$$

$$k_\mu=1\wedge\min_{a\neq a^\star}\mathrm{KL}(\mu_a-\varepsilon_\mu|\mu^\star+\varepsilon_\mu)\wedge\min_{\substack{a^+\in\mathcal{A}^+\\ a\in\mathcal{V}_{a^+}}}\mathrm{KL}(\mu_a-\varepsilon_\mu|\mu_{a^+}\varepsilon_\mu)\tag{33}$$

$$\varphi:x\geqslant0\mapsto\min\left\{x\,;x\,\frac{k_\mu\wedge\Phi(x)}{\mathrm{KL}(b\,|\mu^\star+\varepsilon_\mu)\wedge\Phi(x)}\right\}\tag{34}$$

Lemma 8 (Well-estimated means)**.** *Under* `IMED-MB`*, for all* $0<\varepsilon<\varepsilon_\mu$*, at each time step* $t\notin\mathcal{E}_{a_{t+1}}(\varepsilon)\cup\mathcal{E}_{\widehat{a}^\star_t}(\varepsilon)$*,*

$$\left|\widehat{\mu}_{a_{t+1}}(t)-\mu_{a_{t+1}}\right|<\varepsilon\,,\tag{35}$$

$$\left|\widehat{\mu}_{\widehat{a}^\star_t}(t)-\mu_{\widehat{a}^\star_t}\right|<\varepsilon\,.\tag{36}$$

Proof. These inequalities are derived from the definition of $\mathcal{E}_a(\varepsilon)=\mathcal{E}_a(f,\varepsilon)$ for $a\in\mathcal{A}$ and identity function $f:x\mapsto x$ detailed in Equations (81)-(82)-(83) and the following empirical lower bound from Lemmas 2-3-4,

$$N_{a_{t+1}}(t)\leqslant N_{\widehat{a}^\star_t}(t)\,.$$

□

Lemma 9 (Local maximum)**.** *Under* `IMED-MB`*, for all* $0<\varepsilon<\varepsilon_\mu$*, at each time step* $t\notin\mathcal{E}_{a_{t+1}}(\varepsilon)\cup\mathcal{E}_{\widehat{a}^\star_t}(\varepsilon)\bigcup\limits_{a\in\mathcal{V}_{\widehat{a}^\star_t}}\mathcal{K}^-_a(\Phi,\varepsilon_\mu)$*,*

$$\widehat{a}^\star_t\in\mathcal{A}^+\,.\tag{37}$$

Proof. By contradiction: we assume $\widehat{a}^\star_t\notin\mathcal{A}^+$*.*
Since $t\notin\mathcal{E}_{a_{t+1}}(\varepsilon)\cup\mathcal{E}_{\widehat{a}^\star_t}(\varepsilon)$, Lemma 8 implies the means of the current optimal arm is well estimated, in particular

$$\widehat{\mu}^\star(t)\leqslant\mu_{\widehat{a}^\star_t}+\varepsilon\,.\tag{38}$$

Since $\widehat{a}_t^\star \notin \mathcal{A}^+$ and $\varepsilon < \varepsilon_\nu$, there exist an arm $a \in \underset{a' \in \mathcal{V}_{\widehat{a}_t^\star}}{\arg\max}\, \mu'_a$ such that

$$\widehat{\mu}_a(t) \leqslant \widehat{\mu}^\star(t) \leqslant \mu_{\widehat{a}_t^\star} + \varepsilon < \mu_a - \varepsilon_\nu \,. \tag{39}$$

Furthermore, the empirical lower bounds on the numbers of pulls from Lemmas 2-3-4 imply

$$\log\big(N_{a_{t+1}}(t)\big) \leqslant N_a(t)\, \mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) \wedge \Phi(N_a(t)) + \log(N_a(t)) \,. \tag{40}$$

Noting that : $\mu \geqslant \widehat{\mu}_a(t) \mapsto \mathrm{KL}(\widehat{\mu}_a(t)|\mu)$ is an increasing function, by combining previous Equations (39)-(40) we obtain

$$\log\big(N_{a_{t+1}}(t)\big) \leqslant N_a(t)\, \mathrm{KL}(\widehat{\mu}_a(t)|\mu_a - \varepsilon_\mu) \wedge \Phi(N_a(t)) + \log(N_a(t)) \,. \tag{41}$$

Then, Equation (41) contradicts the assumption that $t \notin \bigcup_{a \in \mathcal{V}_{\widehat{a}_t^\star}} \mathcal{K}_a^-(\Phi, \varepsilon_\mu)$, which ends the proof. $\mathcal{K}_a^-(\Phi, \varepsilon_\mu)$ is defined in Equation (84). □

Lemma 10 (Global maximum - no structure exploitation)**.** *Under* `IMED-MB`*, for all* $0 < \varepsilon < \varepsilon_\mu$*, at each time step* $t \notin \mathcal{E}_{a_{t+1}}(\varepsilon) \cup \mathcal{E}_{\widehat{a}_t^\star}(\varepsilon) \cup \mathcal{K}_{a^\star}^-(\Phi, \varepsilon_\mu)$ *such that* $\left|\widehat{\mathcal{A}}^+(t)\right| < M$ *(that is, when there is no structure exploitation),*

$$\widehat{a}_t^\star = a^\star \,. \tag{42}$$

Proof. By contradiction: we assume $\widehat{a}_t^\star \neq a^\star$.
Since $t \notin \mathcal{E}_{a_{t+1}}(\varepsilon) \cup \mathcal{E}_{\widehat{a}_t^\star}(\varepsilon)$, Lemma 8 implies the means of the current optimal arm is well estimated, in particular

$$\widehat{\mu}^\star(t) \leqslant \mu_{\widehat{a}_t^\star} + \varepsilon \,. \tag{43}$$

Since $\widehat{a}_t^\star \neq a^\star$ and $\varepsilon < \varepsilon_\nu$,

$$\widehat{\mu}_a(t) \leqslant \widehat{\mu}^\star(t) \leqslant \mu_{\widehat{a}_t^\star} + \varepsilon < \mu_{a^\star} - \varepsilon_\nu \,. \tag{44}$$

Furthermore, the empirical lower bounds on the numbers of pulls from Lemmas 2-3-4 imply

$$\log\big(N_{a_{t+1}}(t)\big) \leqslant N_{a^\star}(t)\, \mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) \wedge \Phi(N_a(t)) + \log(N_{a^\star}(t)) \,. \tag{45}$$

Noting that : $\mu \geqslant \widehat{\mu}_{a^\star}(t) \mapsto \mathrm{KL}(\widehat{\mu}_{a^\star}(t)|\mu)$ is an increasing function, by combining previous Equations (44)-(45) we obtain

$$\log\big(N_{a_{t+1}}(t)\big) \leqslant N_{a^\star}(t)\, \mathrm{KL}(\widehat{\mu}_{a^\star}(t)|\mu_a - \varepsilon_\mu) \wedge \Phi(N_a(t)) + \log(N_{a^\star}(t)) \,. \tag{46}$$

Then, Equation (46) contradicts the assumption that $t \notin \mathcal{K}_{a^\star}^-(\Phi, \varepsilon_\mu)$, which ends the proof. $\mathcal{K}_{a^\star}^-(\Phi, \varepsilon_\mu)$ is defined in Equation (84). □

Lemma 11 (Global maximum - second-order exploration)**.** *Under* `IMED-MB`*, for all* $0 < \varepsilon < \varepsilon_\mu$*, at each time step* $t \notin \mathcal{E}_{a_{t+1}}(\varepsilon) \cup \mathcal{E}_{\widehat{a}_t^\star}(\varepsilon) \cup \mathcal{K}_{a^\star}^-(\Phi, \varepsilon_\mu)$ *such that* $\left|\widehat{\mathcal{A}}^+(t)\right| = M$ *(that is, when there is structure exploitation) and* $a_{t+1} \notin \widehat{\mathcal{A}}^M(t)$ *(that is, there is second-order exploration),*

$$\widehat{a}_t^\star = a^\star \,. \tag{47}$$

Proof. The same proof as that of Lemma 10 holds. □

Lemma 12 (Well-estimated means - no structure exploitation)**.** *Under* `IMED-MB`*, for all* $0<\varepsilon<\varepsilon_\mu$*, at each time step* $t \notin \bigcup_{a\in\mathcal{A}} \mathcal{E}_a(\varepsilon)\cup\mathcal{E}_a(\varphi,\varepsilon)\cup\mathcal{K}_a^-(\Phi,\varepsilon_\mu)$ *such that* $\left|\widehat{\mathcal{A}}^+(t)\right|<M$ *(that is, when there is no structure exploitation) and* $a_{t+1}\neq\widehat{a}_t^\star$*, for all* $a\in\mathcal{A}$*,*

$$|\widehat{\mu}_a(t)-\mu_a|<\varepsilon\,. \tag{48}$$

Proof. From Lemma 8 and since $t\notin\mathcal{E}_{a_{t+1}}(\varepsilon)\cup\mathcal{E}_{\widehat{a}_t^\star,a_{t+1}}$, the means of the current pulled arm and the current optimal arm are well-estimated,

$$\mu_{a_{t+1}}-\varepsilon_\mu<\widehat{\mu}_{a_{t+1}}(t)<\mu_{a_{t+1}}+\varepsilon_\mu\,, \tag{49}$$

$$\mu_{\widehat{a}_t^\star}-\varepsilon_\mu<\widehat{\mu}_{\widehat{a}_t^\star}(t)<\mu_{\widehat{a}_t^\star}+\varepsilon_\mu\,. \tag{50}$$

From Lemma 10 and since $t\notin\mathcal{E}_{a_{t+1}}(\varepsilon)\cup\mathcal{E}_{\widehat{a}_t^\star}(\varepsilon)\cup\mathcal{K}_{a^\star}^-(\Phi,\varepsilon_\mu)$, the current best arm is the best arm, that is, $\widehat{a}_t^\star=a^\star$. Since $a_{t+1}\neq\widehat{a}_t^\star$, this implies

$$\mu_{a_{t+1}}-\varepsilon_\mu<\widehat{\mu}_{a_{t+1}}(t)<\mu_{a_{t+1}}+\varepsilon_\mu<\mu_{\widehat{a}_t^\star}-\varepsilon_\mu<\widehat{\mu}_{\widehat{a}_t^\star}(t)<\mu_{\widehat{a}_t^\star}+\varepsilon_\mu\,. \tag{51}$$

By combining previous Equation (51) and the monotonic properties of $\mathrm{KL}(\cdot|\cdot)$, we get

$$k_\mu\wedge\Phi\big(N_{a_{t+1}}(t)\big)\leqslant\mathrm{KL}\Big(\widehat{\mu}_{a_{t+1}}(t)\Big|\widehat{\mu}_{\widehat{a}_t^\star}(t)\Big)\wedge\Phi\big(N_{a_{t+1}}(t)\big)=\mathrm{KL}\big(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^\star(t)\big)\wedge\Phi\big(N_{a_{t+1}}(t)\big)\,, \tag{52}$$

where k_μ is defined in Equation (33). From Lemmas 2- 3-4, we have the following empirical lower bound on $N_a(t)$,

$$\min\left\{N_{a_{t+1}}(t)\;;\;N_{a_{t+1}}(t)\,\frac{\mathrm{KL}\big(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^\star(t)\big)\wedge\Phi\big(N_{a_{t+1}}(t)\big)}{\mathrm{KL}\big(b\,|\widehat{\mu}^\star(t)\big)\wedge\Phi\big(N_{a_{t+1}}(t)\big)}\right\}\leqslant N_a(t)\,. \tag{53}$$

We note that $\mathrm{KL}(b|\cdot)$ is a non-decreasing function on $[b\,;B[$. Then, Equation (51) also implies

$$\mathrm{KL}(b|\widehat{\mu}^\star(t))\leqslant\mathrm{KL}(b|\mu^\star+\varepsilon_\mu)\;. \tag{54}$$

By combining previous Equations (52)-(53)-(54), we have

$$\varphi\big(N_{a_{t+1}}(t)\big)=\min\left\{N_{a_{t+1}}(t)\;;\;N_{a_{t+1}}(t)\,\frac{k_\mu\wedge\Phi\big(N_{a_{t+1}}(t)\big)}{\mathrm{KL}\big(b\,|\mu^\star+\varepsilon_\mu\big)\wedge\Phi\big(N_{a_{t+1}}(t)\big)}\right\}\leqslant N_a(t)\,. \tag{55}$$

Since $t\notin\mathcal{E}_a(\varphi,\varepsilon)$ defined in Equations (81)-(82)-(83), previous Equation (55) implies

$$|\widehat{\mu}_a(t)-\mu_a|<\varepsilon\,,$$

which ends the proof. □

Lemma 13 (Well-estimated means - no second-order exploration)**.** *Under* `IMED-MB`*, for all* $0<\varepsilon<\varepsilon_\mu$*, at each time step* $t\notin\bigcup_{a\in\mathcal{A}}\mathcal{E}_a(\varepsilon)\cup\mathcal{E}_a(\varphi,\varepsilon)\cup\mathcal{K}_a^-(\Phi,\varepsilon_\mu)$ *such that* $\left|\widehat{\mathcal{A}}^+(t)\right|=M$ *(that is, when there is structure exploitation),* $a_{t+1}\in\widehat{\mathcal{A}}^M(t)$ *(that is, there no is second-order*

exploration), and $a_{t+1} \neq \widehat{a}_t^\star$*, for all* $a \in \widehat{\mathcal{A}}^M(t)$*,*

$$|\widehat{\mu}_a(t) - \mu_a| < \varepsilon\,. \tag{56}$$

Proof. From Lemma 8 and since $t \notin \mathcal{E}_{a_{t+1}}(\varepsilon) \cup \mathcal{E}_{\widehat{a}_t^\star, a_{t+1}}$, the means of the current pulled arm and the current optimal arm are well-estimated,

$$\mu_{a_{t+1}} - \varepsilon_\mu < \widehat{\mu}_{a_{t+1}}(t) < \mu_{a_{t+1}} + \varepsilon_\mu\,, \tag{57}$$

$$\mu_{\widehat{a}_t^\star} - \varepsilon_\mu < \widehat{\mu}_{\widehat{a}_t^\star}(t) < \mu_{\widehat{a}_t^\star} + \varepsilon_\mu\,. \tag{58}$$

From Lemma 9 and since $t \notin \mathcal{E}_{a_{t+1}}(\varepsilon) \cup \mathcal{E}_{\widehat{a}_t^\star}(\varepsilon) \bigcup_{a \in \mathcal{V}_{\widehat{a}_t^\star}} \mathcal{K}_a^-(\Phi, \varepsilon_\mu)$, the current best arm is a locally optimal arm, that is, $\widehat{a}_t^\star \in \mathcal{A}^+$. Since $a_{t+1} \in \mathcal{V}_{\widehat{a}_t^\star}$, this implies

$$\mu_{a_{t+1}} - \varepsilon_\mu < \widehat{\mu}_{a_{t+1}}(t) < \mu_{a_{t+1}} + \varepsilon_\mu < \mu_{\widehat{a}_t^\star} - \varepsilon_\mu < \widehat{\mu}_{\widehat{a}_t^\star}(t) < \mu_{\widehat{a}_t^\star} + \varepsilon_\mu\,. \tag{59}$$

By combining previous Equation (59) and the monotonic properties of $\mathrm{KL}(\cdot|\cdot)$, we get

$$k_\mu \wedge \Phi\big(N_{a_{t+1}}(t)\big) \leqslant \mathrm{KL}\Big(\widehat{\mu}_{a_{t+1}}(t)\Big|\widehat{\mu}_{\widehat{a}_t^\star}(t)\Big) \wedge \Phi\big(N_{a_{t+1}}(t)\big) = \mathrm{KL}\big(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^\star(t)\big) \wedge \Phi\big(N_{a_{t+1}}(t)\big)\,, \tag{60}$$

where k_μ is defined in Equation (33). From Lemmas 2-3-4, we have the following empirical lower bound on $N_a(t)$,

$$\min\left\{N_{a_{t+1}}(t)\ ;\ N_{a_{t+1}}(t)\,\frac{\mathrm{KL}\big(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^\star(t)\big) \wedge \Phi\big(N_{a_{t+1}}(t)\big)}{\mathrm{KL}\big(b\,|\widehat{\mu}^\star(t)\big) \wedge \Phi\big(N_{a_{t+1}}(t)\big)}\right\} \leqslant N_a(t)\,. \tag{61}$$

We note that $\mathrm{KL}(b|\cdot)$ is a non-decreasing function on $[b\,; B[$. Then, Equation (59) also implies

$$\mathrm{KL}(b|\widehat{\mu}^\star(t)) \leqslant \mathrm{KL}(b|\mu^\star + \varepsilon_\mu)\,. \tag{62}$$

By combining previous Equations (60)-(61)-(62), we have

$$\varphi\big(N_{a_{t+1}}(t)\big) = \min\left\{N_{a_{t+1}}(t)\ ;\ N_{a_{t+1}}(t)\,\frac{k_\mu \wedge \Phi\big(N_{a_{t+1}}(t)\big)}{\mathrm{KL}\big(b\,|\mu^\star + \varepsilon_\mu\big) \wedge \Phi\big(N_{a_{t+1}}(t)\big)}\right\} \leqslant N_a(t)\,. \tag{63}$$

Since $t \notin \mathcal{E}_a(\varphi, \varepsilon)$ defined in Equations (81)-(82)-(83), previous Equation (63) implies

$$|\widehat{\mu}_a(t) - \mu_a| < \varepsilon\,,$$

which ends the proof. □

Lemma 14 (Well-estimated means - second-order exploration)**.** *Under* `IMED-MB`*, for all* $0 < \varepsilon < \varepsilon_\mu$*, at each time step* $t \notin \bigcup_{a \in \mathcal{A}} \mathcal{E}_a(\varepsilon) \cup \mathcal{E}_a(\varphi, \varepsilon) \cup \mathcal{K}_a^-(\Phi, \varepsilon_\mu)$ *such that* $\left|\widehat{\mathcal{A}}^+(t)\right| = M$ *(that is, when there is structure exploitation),* $a_{t+1} \notin \widehat{\mathcal{A}}^M(t)$ *(that is, there is second-order exploration), and* $a_{t+1} \neq \widehat{a}_t^\star$*, for all* $a \in \mathcal{A}$*,*

$$|\widehat{\mu}_a(t) - \mu_a| < \varepsilon\,. \tag{64}$$

Proof. The same proof as that of Lemma 12 holds. □

Lemma 15 (Structure estimation - exploration)**.** *Under* `IMED-MB`*, for all* $0<\varepsilon<\varepsilon_\mu$*, at each time step* $t \notin \bigcup_{a\in\mathcal{A}} \mathcal{E}_a(\varepsilon) \cup \mathcal{E}_a(\varphi,\varepsilon) \cup \mathcal{K}_a^-(\Phi,\varepsilon_\mu)$ *such that* $a_{t+1} \neq \widehat{a}_t^\star$*,*

$$\widehat{a}_t^\star = a^\star\,,$$

$$\widehat{\mathcal{A}}^+(t) = \mathcal{A}^+(M)\,,$$

where $\mathcal{A}^+(M) = \left\{a_1,\dots,a_{\min(M,M^+)}\right\} \subset \mathcal{A}^+$*. We refer to Assumption 2 for the definition of* $a_1,\dots,a_{M^+}$*.*

Proof. It is a direct consequence of Lemmas 12-13-14. □

Lemma 16 (Structure estimation - exploitation)**.** *Under* `IMED-MB`*, for all* $0<\varepsilon<\varepsilon_\mu$*, at each time step* $t \notin \bigcup_{a\in\mathcal{A}} \mathcal{E}_a(\varepsilon) \cup \mathcal{E}_a(\varphi,\varepsilon) \cup \mathcal{K}_a^-(\Phi,\varepsilon_\mu)$ *such that* $a_{t+1} = \widehat{a}_t^\star \neq a^\star$*,*

$$a_{t+1} = \widehat{a}_t^\star \in \mathcal{A}^+\,,$$

$$\left|\widehat{\mathcal{A}}^+(t)\right| = M\,,$$

$$a^\star \notin \widehat{\mathcal{A}}^M(t)\,,$$

$$\widehat{\mathcal{A}}^+(t) - \mathcal{A}^+(M) \neq \emptyset\,,$$

where $\mathcal{A}^+(M) = \left\{a_1,\dots,a_{\min(M,M^+)}\right\} \subset \mathcal{A}^+$*. We refer to Assumption 2 for the definition of* $a_1,\dots,a_{M^+}$*.*

Proof. Since $t \notin \mathcal{E}_{a_{t+1}}(\varepsilon) \cup \mathcal{E}_{\widehat{a}_t^\star}(\varepsilon) \bigcup_{a\in\mathcal{V}_{\widehat{a}_t^\star}} \mathcal{K}_a^-(\Phi,\varepsilon_\mu)$, from Lemma 9 we directly have that $\widehat{a}_t^\star \in \mathcal{A}^+$.

Since $t \notin \mathcal{E}_{a_{t+1}}(\varepsilon) \cup \mathcal{E}_{\widehat{a}_t^\star}(\varepsilon) \cup \mathcal{K}_{a^\star}^-(\Phi,\varepsilon_\mu)$ and $\widehat{a}_t^\star \neq a^\star$, from Lemma 10 we directly have that $\left|\widehat{\mathcal{A}}^+(t)\right| = M$. Furthermore, since $t \notin \mathcal{E}_{a_{t+1}}(\varepsilon) \cup \mathcal{E}_{\widehat{a}_t^\star,a_{t+1}}$, Lemma 8 implies $\widehat{\mu}_{\widehat{a}_t^\star}(t)$ is well-estimated, which implies

$$\widehat{\mu}_{a^\star}(t) \leqslant \widehat{\mu}^\star(t) = \widehat{\mu}_{\widehat{a}_t^\star}(t) < \mu_{\widehat{a}_t^\star} + \varepsilon < \mu_{a^\star} - \varepsilon_\nu\,. \tag{65}$$

Since : $\mu \geqslant \widehat{\mu}_{a^\star}(t) \mapsto \mathrm{KL}(\widehat{\mu}_{a^\star}(t)|\mu)$ is increasing function, previous Equation (65) implies

$$\begin{aligned} I_{a^\star}^\Phi(t) &= N_{a^\star}(t)\,\mathrm{KL}(\widehat{\mu}_{a^\star}(t)|\widehat{\mu}^\star(t)) \wedge \Phi(N_{a^\star}(t)) + \log(N_{a^\star}(t)) \\ &\leqslant N_{a^\star}(t)\,\mathrm{KL}(\widehat{\mu}_{a^\star}(t)|\mu_{a^\star} - \varepsilon_\mu) \wedge \Phi(N_{a^\star}(t)) + \log(N_{a^\star}(t))\,. \end{aligned} \tag{66}$$

Since $t \notin \mathcal{K}_{a^\star}^-(\Phi,\varepsilon_\mu)$ defined in Equation (84), previous Equation (66) implies

$$I_{a^\star}^\Phi(t) < \log\left(N_{a_{t+1}}\right) \leqslant \operatorname*{arg\,min}_{a\in\widehat{\mathcal{A}}^M(t)} I_a^\Phi(t)\,, \tag{67}$$

which naturally implies $a^\star \notin \widehat{\mathcal{A}}^M(t)$. This implies that $a^\star \notin \widehat{\mathcal{A}}^+(t)$ and $\widehat{\mathcal{A}}^+(t) - \mathcal{A}^+(M) \neq \emptyset$, since $\left|\widehat{\mathcal{A}}^+(t)\right| = M$ and $|\mathcal{A}^+(M) - \{a^\star\}| \leqslant M-1$.

□

C.3 Upper bounds on the numbers of pulls of locally sub-optimal arms

In this subsection, we mainly combine the results from Lemmas 15-16 of the previous section and empirical upper bounds of section C.1 in order to obtain randomized upper bounds on the numbers of pulls of arms outside of $\mathcal{A}^+$. We first introduce following subsets of time steps

$$\begin{aligned} \mathcal{U}(\varepsilon) &= \bigcup_{a\in\mathcal{A}} \mathcal{E}_a(\varepsilon) \cup \mathcal{E}_a(\varphi,\varepsilon) \cup \mathcal{K}_a^-(\Phi,\varepsilon_\mu)\,, \\ \mathcal{U}_a(\varepsilon) &= \{t \in \mathcal{U}(\varepsilon) :\ a_{t+1} = a\}\,, \quad \forall a \in \mathcal{A}\,. \end{aligned} \tag{68}$$

From the definition of φ in Equation (34), we have $\varphi(x) \leqslant x$ for $x \geqslant 0$. Then, $\mathcal{U}(\varepsilon)$ can be simplify as follows

$$\mathcal{U}(\varepsilon) = \bigcup_{a\in\mathcal{A}} \mathcal{E}_a(\varphi,\varepsilon) \cup \mathcal{K}_a^-(\Phi,\varepsilon_\mu)\,.$$

The next lemma borrows elements of proof from Combes and Proutiere (2014) in the way of providing upper bounds on the numbers of pulls involving the sizes of these (bad events) sets $(\mathcal{U}_a)$, see Equation (69).

Lemma 17. *Under* `IMED-MB`*, for all* $0<\varepsilon<\varepsilon_\mu$*, for all arm* $a \notin \mathcal{A}^+$*, for all time step* $t \geqslant 1$*,*
$\star$ *if* $|\mathcal{A}^+|<M$*,*

$$N_a(t) \leqslant \left(\Phi^{-1}(\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)) \wedge \frac{\log(t)}{\Phi(\log(t))}\right) \vee \frac{\log(t)}{\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)} + |\mathcal{U}_a(\varepsilon)| + 1\,,$$

$$\mathbb{E}_\nu[N_a(t)] \leqslant \left(\Phi^{-1}(\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)) \wedge \frac{\log(t)}{\Phi(\log(t))}\right) \vee \frac{\log(t)}{\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)} + \mathbb{E}[|\mathcal{U}_a(\varepsilon)|] + 1\,,$$

$\star$ *if* $|\mathcal{A}^+|\geqslant M$ *and* $a\in\mathcal{V}_{\mathcal{A}^+(M)}$*,*

$$N_a(t) \leqslant \left(\Phi^{-1}(\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)) \wedge \frac{\log(t)}{\Phi(\log(t))}\right) \vee \frac{\log(t)}{\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)} + |\mathcal{U}_a(\varepsilon)| + 1\,,$$

$$\mathbb{E}_\nu[N_a(t)] \leqslant \left(\Phi^{-1}(\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)) \wedge \frac{\log(t)}{\Phi(\log(t))}\right) \vee \frac{\log(t)}{\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)} + \mathbb{E}[|\mathcal{U}_a(\varepsilon)|] + 1\,,$$

$\star$ *otherwise,* $|\mathcal{A}^+|\geqslant M$ *and* $a\notin\mathcal{A}^+\cup\mathcal{V}_{\mathcal{A}^+(M)}$ *and*

$$N_a(t) \leqslant \left(\Phi^{-1}(\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)) \wedge \frac{\Psi^{-1}(\log(t))}{\Phi(\Psi^{-1}(\log(t)))}\right) \vee \frac{\Psi^{-1}(\log(t))}{\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)} + |\mathcal{U}_a(\varepsilon)| + 1\,,$$

$$\mathbb{E}_\nu[N_a(t)] \leqslant \left(\Phi^{-1}(\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)) \wedge \frac{\Psi^{-1}(\log(t))}{\Phi(\Psi^{-1}(\log(t)))}\right) \vee \frac{\Psi^{-1}(\log(t))}{\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)} + \mathbb{E}[|\mathcal{U}_a(\varepsilon)|] + 1\,,$$

where $\Phi^{-1} : y \geqslant 0 \mapsto \max\{x\geqslant 0 : \Phi(x) \leqslant y\}$ *(with the convention* $\max\emptyset = 0$*) and* $\mathcal{A}^+(M) = \{a_1,\ldots,a_{\min(M,M^+)}\}$ *(see Assumption 2).*

Furthermore, when $1 \leqslant \Phi(\log(n)) \leqslant \log\log(n)$ *for* $n \geqslant 18$, *an upper bound on* $\mathbb{E}[|\mathcal{U}(\varepsilon)|]$ *is*

$$\mathbb{E}[|\mathcal{U}(\varepsilon)|] \leqslant \left(1 + 18 \vee \frac{e}{\min_{a \in \mathcal{A}} \mathrm{KL}(\mu_a - \varepsilon|\mu_a)}\right)|\mathcal{A}| + \sum_{a \in \mathcal{A}} 2 + \sum_{n \geqslant 1} e^{-\varphi(n)\mathrm{KL}(\mu_a - \varepsilon|\mu_a)} + \frac{1}{n \log^2(n)} + \sum_{a \in \mathcal{A}} \frac{e^{1+\mathrm{KL}(\mu_a - \varepsilon|\mu_a)}}{\mathrm{KL}(\mu_a - \varepsilon|\mu_a)} \sum_{n \geqslant 18} \frac{\big(1+\log^2(n)\big)\big(\log(n)+2\log\log(n)\big)}{n^{1+\mathrm{KL}(\mu_a - \varepsilon|\mu_a)/\log\log(n)}},$$

where φ *and* $\mathcal{U}(\varepsilon) = \bigcup_{a \in \mathcal{A}} \mathcal{U}_a(\varepsilon)$ *are respectively defined in Equation (34) and Equation (68). In particular,* $\mathbb{E}[|\mathcal{U}(\varepsilon)|] < \infty$, *which implies* $\mathcal{U}_a(\varepsilon) < \infty$ *almost surely.*

Proof. We note that for an arm $a \in \mathcal{A}$, its number of pulls up to time step $t \geqslant 1$ can be broken down as follows,

$$N_a(t) = N_a(\tau_a) + N_a(t) - N_a(\tau_a)\,,$$

where $\tau_a = \max\{s \leqslant t-1 : a_{s+1} = a,\ t \notin \mathcal{U}_a(\varepsilon)\}$ and $N_a(t) - N_a(\tau_a) \leqslant |\mathcal{U}_a(\varepsilon)| + 1$. This implies

$$N_a(t) \leqslant N_a(\tau_a) + |\mathcal{U}_a(\varepsilon)| + 1\,, \tag{69}$$

and the upper bounds on the numbers of pulls are a direct consequence of Lemma 19.

We now prove the upper bound on $\mathbb{E}[|\mathcal{U}(\varepsilon)|]$ as a consequence of Lemma 24. Indeed, we just note that

$$\mathcal{U}(\varepsilon) \subset \mathcal{I}\left(e^{-1} \min_{a \in \mathcal{A}} \mathrm{KL}(\mu_a - \varepsilon|\mu_a)\right) \bigcup_{a \in \mathcal{A}} \mathcal{E}_a(\varphi, \varepsilon) \cup \mathcal{E}_a(f_{a,\varepsilon}, \varepsilon) \bigcup_{a \in \mathcal{A}} \mathcal{K}_a^-(\Phi, \varepsilon) - \mathcal{E}_a(f_{a,\varepsilon}, \varepsilon) - \mathcal{I}_a(\varepsilon)\,,$$

where $\mathcal{I}(K)$ for $K > 0$ is defined in Equation 85 while $f_{a,\varepsilon}(\cdot)$ and $\mathcal{I}_a(\varepsilon)$ are defined in Lemma 24. This implies

$$|\mathcal{U}(\varepsilon)| \leqslant \left|\mathcal{I}\left(e^{-1} \min_{a \in \mathcal{A}} \mathrm{KL}(\mu_a - \varepsilon|\mu_a)\right)\right| + \sum_{a \in \mathcal{A}} |\mathcal{E}_a(\varphi, \varepsilon)| + |\mathcal{E}_a(f_{a,\varepsilon}, \varepsilon)| + \left|\mathcal{K}_a^-(\Phi, \varepsilon) - \mathcal{E}_a(f_{a,\varepsilon}, \varepsilon) - \mathcal{I}_a(\varepsilon)\right|. \tag{70}$$

The upper bound on $\mathbb{E}[|\mathcal{U}(\varepsilon)|]$ is then proved by taking the expectation on both sides of previous Equation (70) and applying Lemma 24. □

Lemma 18. *We assume* $1 \leqslant \Phi(\log(n)) \leqslant \log\log(n)$ *for* $n \geqslant 18$ *and* $\Psi^{-1}(\log(t)) = \underset{t \to \infty}{o}(\log(t))$. *Then, under* `IMED-MB`, *for all arm* $a \notin \mathcal{A}^+$,

$\star$ *if* $|\mathcal{A}^+| < M$,

$$a.s. \quad \limsup_{T \to \infty} \frac{N_a(T)}{\log(T)} \leqslant \frac{1}{\mathrm{KL}(\mu_a|\mu^\star)}\,,$$

which implies

$$\limsup_{T \to \infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\log(T)} \leqslant \mathbb{E}_\nu\left[\limsup_{T \to \infty} \frac{N_a(T)}{\log(T)}\right] \leqslant \frac{1}{\mathrm{KL}(\mu_a|\mu^\star)}\,,$$

⋆ *if* $|\mathcal{A}^+| \geqslant M$ *and* $a \in \mathcal{V}_{\mathcal{A}^+(M)}$,

$$a.s. \quad \limsup_{T\to\infty} \frac{N_a(T)}{\log(T)} \leqslant \frac{1}{\mathrm{KL}(\mu_a|\mu^\star)}\,,$$

which implies

$$\limsup_{T\to\infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\log(T)} \leqslant \mathbb{E}_\nu\left[\limsup_{T\to\infty} \frac{N_a(T)}{\log(T)}\right] \leqslant \frac{1}{\mathrm{KL}(\mu_a|\mu^\star)}\,,$$

⋆ *otherwise,* $|\mathcal{A}^+| \geqslant M$ *and* $a \notin \mathcal{A}^+ \cup \mathcal{V}_{\mathcal{A}^+(M)}$ *and*

$$a.s. \quad \limsup_{T\to\infty} \frac{N_a(T)}{\log(T)} \leqslant 0\,,$$

which implies

$$\limsup_{T\to\infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\log(T)} \leqslant \mathbb{E}_\nu\left[\limsup_{T\to\infty} \frac{N_a(T)}{\log(T)}\right] \leqslant 0\,,$$

where $\mathcal{A}^+(M) = \left\{a_1, \dots, a_{\min(M,M^+)}\right\}$ *(see Assumption 2).*

Proof. It is a direct consequence of Lemma 17. □

Lemma 19. *Under* `IMED-MB`*, for all* $0 < \varepsilon < \varepsilon_\mu$*, at each time step* $t \notin \mathcal{U}(\varepsilon)$ *such that* $a_{t+1} = a \notin \mathcal{A}^+$,
⋆ *if* $|\mathcal{A}^+| < M$,

$$N_a(t) \leqslant \left(\Phi^{-1}(\mathrm{KL}(\mu_a + \varepsilon|\mu^\star - \varepsilon)) \wedge \frac{\log(t)}{\Phi(\log(t))}\right) \vee \frac{\log(t)}{\mathrm{KL}(\mu_a + \varepsilon|\mu^\star - \varepsilon)}\,,$$

⋆ *if* $|\mathcal{A}^+| \geqslant M$ *and* $a \in \mathcal{V}_{\mathcal{A}^+(M)}$,

$$N_a(t) \leqslant \left(\Phi^{-1}(\mathrm{KL}(\mu_a + \varepsilon|\mu^\star - \varepsilon)) \wedge \frac{\log(t)}{\Phi(\log(t))}\right) \vee \frac{\log(t)}{\mathrm{KL}(\mu_a + \varepsilon|\mu^\star - \varepsilon)}\,,$$

⋆ *otherwise,* $|\mathcal{A}^+| \geqslant M$ *and* $a \notin \mathcal{A}^+ \cup \mathcal{V}_{\mathcal{A}^+(M)}$ *and*

$$N_a(t) \leqslant \left(\Phi^{-1}(\mathrm{KL}(\mu_a + \varepsilon|\mu^\star - \varepsilon)) \wedge \frac{\Psi^{-1}(\log(t))}{\Phi(\Psi^{-1}(\log(t)))}\right) \vee \frac{\Psi^{-1}(\log(t))}{\mathrm{KL}(\mu_a + \varepsilon|\mu^\star - \varepsilon)}\,,$$

where $\Phi^{-1} : y \geqslant 0 \mapsto \max\{x \geqslant 0 : \Phi(x) \leqslant y\}$ *(with the convention* $\max \emptyset = 0$*) and* $\mathcal{A}^+(M) = \left\{a_1, \dots, a_{\min(M,M^+)}\right\}$ *(see Assumption 2).*

Proof. Since $t \notin \mathcal{U}(\varepsilon)$ defined in Equation (68), $t \notin \mathcal{E}_{a_{t+1}}(\varepsilon) \cup \mathcal{E}_{\widehat{a}_t^\star}(\varepsilon) \bigcup_{a \in \mathcal{V}_{\widehat{a}_t^\star}} \mathcal{K}_a^-(\Phi, \varepsilon_\mu)$ and Lemma 9 implies $\widehat{a}_t^\star \in \mathcal{A}^+$. Since $a_{t+1} = a \notin \mathcal{A}^+$, this implies in particular $a_{t+1} \neq \widehat{a}_t^\star$. Then, since $t \notin \mathcal{U}(\varepsilon) = \bigcup_{a \in \mathcal{A}} \mathcal{E}_a(\varepsilon) \cup \mathcal{E}_a(\varphi, \varepsilon) \cup \mathcal{K}_a^-(\Phi, \varepsilon_\mu)$ and $a_{t+1} \neq \widehat{a}_t^\star$, Lemma 15 implies

$$\widehat{a}_t^\star = a^\star, \qquad \widehat{\mathcal{A}}^+(t) = \mathcal{A}^+(M)\,.$$

Before going any further, we note that since $t \notin \mathcal{U}(\varepsilon)$ and $a_{t+1} = a, \widehat{a}_t^\star = a^\star$, $t \notin \mathcal{E}_{a_{t+1}}(\varepsilon) \cup \mathcal{E}_{\widehat{a}_t^\star}(\varepsilon)$ and Lemma 8 implies

$$\widehat{\mu}_a(t) < \mu_a + \varepsilon < \mu^\star - \varepsilon < \widehat{\mu}^\star(t)\,. \tag{71}$$

$\star$ *Case 1:* $|\mathcal{A}^+| < M$
Then according to `IMED-MB` algorithm, $\left|\widehat{\mathcal{A}}^+(t)\right| = |\mathcal{A}^+| < M$, $a = a_{t+1} \in \mathcal{A}$ and the empirical upper bound from Lemma 5 is satisfied,

$$N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) \wedge \Phi(N_a(t)) \leqslant \log(t)\,. \tag{72}$$

From Equation (71) and the monotonic properties of $\mathrm{KL}(\cdot|\cdot)$, we have $\mathrm{KL}(\mu_a + \varepsilon|\mu^\star - \varepsilon) \leqslant \mathrm{KL}(\widehat{\mu}_a|\widehat{\mu}^\star(t))$. Then previous Equation (72) implies

$$N_a(t)\,\mathrm{KL}(\mu_a + \varepsilon|\mu^\star - \varepsilon) \wedge \Phi(N_a(t)) \leqslant N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) \wedge \Phi(N_a(t)) \leqslant \log(t)\,.$$

This implies

$$N_a(t) \leqslant \left(\Phi^{-1}(\mathrm{KL}(\mu_a + \varepsilon|\mu^\star - \varepsilon)) \wedge \frac{\log(t)}{\Phi(\log(t))}\right) \vee \frac{\log(t)}{\mathrm{KL}(\mu_a + \varepsilon|\mu^\star - \varepsilon)}\,.$$

$\star$ *Case 2:* $|\mathcal{A}^+| \geqslant M$ *and* $a \in \mathcal{V}_{\mathcal{A}^+}(M)$
Then according to `IMED-MB` algorithm, $\left|\widehat{\mathcal{A}}^+(t)\right| = |\mathcal{A}^+(M)| = M$, $a = a_{t+1} \in \mathcal{V}_{\mathcal{A}^+(M)} = \mathcal{V}_{\widehat{\mathcal{A}}^+(t)} \subset \widehat{\mathcal{A}}^M(t)$ and the empirical upper bound from Lemma 6 is satisfied,

$$N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) \wedge \Phi(N_a(t)) \leqslant \log(t)\,. \tag{73}$$

From Equation (71) and the monotonic properties of $\mathrm{KL}(\cdot|\cdot)$, we have $\mathrm{KL}(\mu_a + \varepsilon|\mu^\star - \varepsilon) \leqslant \mathrm{KL}(\widehat{\mu}_a|\widehat{\mu}^\star(t))$. Then previous Equation (73) implies

$$N_a(t)\,\mathrm{KL}(\mu_a + \varepsilon|\mu^\star - \varepsilon) \wedge \Phi(N_a(t)) \leqslant N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) \wedge \Phi(N_a(t)) \leqslant \log(t)\,.$$

This implies

$$N_a(t) \leqslant \left(\Phi^{-1}(\mathrm{KL}(\mu_a + \varepsilon|\mu^\star - \varepsilon)) \wedge \frac{\log(t)}{\Phi(\log(t))}\right) \vee \frac{\log(t)}{\mathrm{KL}(\mu_a + \varepsilon|\mu^\star - \varepsilon)}\,.$$

$\star$ *Case 3:* $|\mathcal{A}^+| \geqslant M$ *and* $a \notin \mathcal{A}^+ \cup \mathcal{V}_{\mathcal{A}^+(M)}$ *(since it is assumed that* $a \notin \mathcal{A}^+$*)*
Then according to `IMED-MB` algorithm, $\left|\widehat{\mathcal{A}}^+(t)\right| = |\mathcal{A}^+(M)| = M$, $a = a_{t+1} \notin \mathcal{A}^+(M) \cup \mathcal{V}_{\mathcal{A}^+(M)} = \widehat{\mathcal{A}}^M(t)$ and the empirical upper bound from Lemma 7 is satisfied,

$$N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) \wedge \Phi(N_a(t)) \leqslant \Psi^{-1}\left(\log(t)\right)\,. \tag{74}$$

From Equation (71) and the monotonic properties of $\mathrm{KL}(\cdot|\cdot)$, we have $\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon) \leqslant \mathrm{KL}(\widehat{\mu}_a|\widehat{\mu}^\star(t))$. Then previous Equation (74) implies

$$N_a(t)\,\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)\wedge\Phi(N_a(t)) \leqslant N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t))\wedge\Phi(N_a(t)) \leqslant \Psi^{-1}\left(\log(t)\right) .$$

This implies

$$N_a(t) \leqslant \left(\Phi^{-1}(\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)) \wedge \frac{\Psi^{-1}\left(\log(t)\right)}{\Phi(\Psi^{-1}\left(\log(t)\right))}\right) \vee \frac{\Psi^{-1}\left(\log(t)\right)}{\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)} .$$

□

C.4 Upper bounds on the numbers of pulls of arms in $\mathcal{A}^+$ when $M \geqslant |\mathcal{A}^+|$

Interestingly enough, in order to obtain upper bounds on the numbers of pulls of arms in $\mathcal{A}^+$, we use the proven upper bounds on arms outside of $\mathcal{A}^+$. This is key point of our proof technique.

Lemma 20. *We assume $M \geqslant |\mathcal{A}^+|$. Then, under* `IMED-MB`*, for all $0<\varepsilon<\varepsilon_\mu$, for all arm $a \in \mathcal{A}^+ - \{a^\star\}$, for all time step $t \geqslant 1$,*

$$N_a(t) \leqslant \max_{a' \notin \mathcal{A}^+ \cup \mathcal{V}_{\mathcal{A}^+}} F\left(\left(\Phi^{-1}(\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)) \wedge \frac{\Psi^{-1}(\log(t))}{\Phi(\Psi^{-1}(\log(t)))}\right) \vee \frac{\Psi^{-1}(\log(t))}{\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)} + |\mathcal{U}_a(\varepsilon)| + 1\right)$$
$$+ \left(\Phi^{-1}(\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)) \wedge \frac{\log(t)}{\Phi(\log(t))}\right) \vee \frac{\log(t)}{\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)} + |\mathcal{U}_a(\varepsilon)| + 1 ,$$

where where $F : x > 0 \mapsto e^{x\Phi(x)+\log(x)}$, $\Phi^{-1} : y \geqslant 0 \mapsto \max\{x \geqslant 0 : \Phi(x) \leqslant y\}$ (with the convention $\max \emptyset = 0$).

Furthermore, when $1 \leqslant \Phi(\log(n)) \leqslant \log\log(n)$ for $n \geqslant 18$, an upper bound on $\mathbb{E}[|\mathcal{U}(\varepsilon)|]$ is

$$\mathbb{E}[|\mathcal{U}(\varepsilon)|] \leqslant \left(1 + 18 \vee \frac{e}{\min\limits_{a\in\mathcal{A}} \mathrm{KL}(\mu_a-\varepsilon|\mu_a)}\right)|\mathcal{A}| + \sum_{a\in\mathcal{A}} 2 + \sum_{n\geqslant 1} e^{-\varphi(n)\mathrm{KL}(\mu_a-\varepsilon|\mu_a)} + \frac{1}{n\log^2(n)}$$
$$+ \sum_{a\in\mathcal{A}} \frac{e^{1+\mathrm{KL}(\mu_a-\varepsilon|\mu_a)}}{\mathrm{KL}(\mu_a-\varepsilon|\mu_a)} \sum_{n\geqslant 18} \frac{\left(1+\log^2(n)\right)\left(\log(n)+2\log\log(n)\right)}{n^{1+\mathrm{KL}(\mu_a-\varepsilon|\mu_a)/\log\log(n)}} ,$$

where φ and $\mathcal{U}(\varepsilon) = \bigcup\limits_{a\in\mathcal{A}} \mathcal{U}_a(\varepsilon)$ are respectively defined in Equation (34) and Equation (68). In particular, $\mathbb{E}[|\mathcal{U}(\varepsilon)|] < \infty$, which implies $\mathcal{U}_a(\varepsilon) < \infty$ almost surely.

Proof. We note that for an arm $a \in \mathcal{A}$, its number of pulls up to time step $t \geqslant 1$ can be broken down as follows,

$$N_a(t) = N_a\big(\max\left\{\tau_a^{exploitation}\,;\tau_a^{exploration}\right\}\big) + N_a(t) - N_a\big(\max\left\{\tau_a^{exploitation}\,;\tau_a^{exploration}\right\}\big) ,$$

where

$$\tau_a^{exploitation} = \max\left\{s \leqslant t-1 : a_{s+1} = a,\ a_{s+1} = \widehat{a}_t^\star,\ t \notin \mathcal{U}_a(\varepsilon)\right\},$$
$$\tau_a^{exploration} = \max\left\{s \leqslant t-1 : a_{s+1} = a,\ a_{s+1} \neq \widehat{a}_t^\star,\ t \notin \mathcal{U}_a(\varepsilon)\right\},$$
$$N_a(t) - N_a\big(\max\left\{\tau_a^{exploitation}; \tau_a^{exploration}\right\}\big) \leqslant |\mathcal{U}_a(\varepsilon)| + 1.$$

This implies

$$N_a(t) \leqslant N_a\big(\tau_a^{exploitation}\big) + N_a\big(\tau_a^{exploration}\big) + |\mathcal{U}_a(\varepsilon)| + 1\,,$$

and the upper bounds on the numbers of pulls are a direct consequence of Lemma 22 and Lemma 23.

We refer to Lemma 17 for a proof of the upper bound on $\mathbb{E}[|\mathcal{U}(\varepsilon)|]$. □

Lemma 21. *We assume $M \geqslant |\mathcal{A}^+|$, $1 \leqslant \Phi(\log(n)) \leqslant \log\log(n)$ for $n \geqslant 18$, $\Psi^{-1}(\log(t)) = \underset{t\to\infty}{o}(\log(t))$, and for all $U > 0$,*

$$\max_{a' \notin \mathcal{A}^+ \cup \mathcal{V}_{\mathcal{A}^+}} F\left(\left(\Phi^{-1}(\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)) \wedge \frac{\Psi^{-1}(\log(t))}{\Phi(\Psi^{-1}(\log(t)))}\right) \vee \frac{\Psi^{-1}(\log(t))}{\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)} + U\right) = \underset{t\to\infty}{o}(\log(t))$$

Then, under `IMED-MB`*, for all arm $a \in \mathcal{A}^+ - \{a^\star\}$,*

$$a.s. \quad \limsup_{T\to\infty} \frac{N_a(T)}{\log(T)} \leqslant \frac{1}{\mathrm{KL}(\mu_a|\mu^\star)}\,,$$

which implies

$$\limsup_{T\to\infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\log(T)} \leqslant \mathbb{E}_\nu\left[\limsup_{T\to\infty} \frac{N_a(T)}{\log(T)}\right] \leqslant \frac{1}{\mathrm{KL}(\mu_a|\mu^\star)}\,,$$

where where $F : x > 0 \mapsto e^{x\Phi(x)+\log(x)}$, $\Phi^{-1} : y \geqslant 0 \mapsto \max\{x \geqslant 0 : \Phi(x) \leqslant y\}$ (with the convention $\max\emptyset = 0$).

Proof. It is a direct consequence of Lemma 20. □

Lemma 22. *We assume $M \geqslant |\mathcal{A}^+|$. Then, under* `IMED-MB`*, for all $0 < \varepsilon < \varepsilon_\mu$, at each time step $t \notin \mathcal{U}(\varepsilon)$ such that $a_{t+1} = a \in \mathcal{A}^+ - \{a^\star\}$ and $a_{t+1} \neq \widehat{a}_t^\star$, it must be that*

$$N_a(t) \leqslant \left(\Phi^{-1}(\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)) \wedge \frac{\log(t)}{\Phi(\log(t))}\right) \vee \frac{\log(t)}{\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)}\,.$$

Proof. Since $t \notin \mathcal{U}(\varepsilon) = \bigcup_{a\in\mathcal{A}} \mathcal{E}_a(\varepsilon) \cup \mathcal{E}_a(\varphi,\varepsilon) \cup \mathcal{K}_a^-(\Phi,\varepsilon_\mu)$ and $a_{t+1} \neq \widehat{a}_t^\star$, Lemma 15 implies

$$\widehat{a}_t^\star = a^\star, \qquad \widehat{\mathcal{A}}^+(t) = \mathcal{A}^+(M)\,.$$

Before going any further, we note that since $t \notin \mathcal{U}(\varepsilon)$ and $a_{t+1} = a, \widehat{a}_t^\star = a^\star$, then $t \notin \mathcal{E}_{a_{t+1}}(\varepsilon) \cup \mathcal{E}_{\widehat{a}_t^\star}(\varepsilon)$ and Lemma 8 implies

$$\widehat{\mu}_a(t) < \mu_a+\varepsilon < \mu^\star-\varepsilon < \widehat{\mu}^\star(t)\,. \tag{75}$$

⋆ *Case 1:* $M = |\mathcal{A}^+|$
Since $\left|\widehat{\mathcal{A}}^+(t)\right| = |\mathcal{A}^+(M)| = |\mathcal{A}^+| = M$ and according to `IMED-MB` algorithm, $a_{t+1} = a \in \mathcal{A}^+ \subset \widehat{\mathcal{A}}^M(t)$ and the empirical upper bound from Lemma 6 is satisfied,

$$N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t))\wedge\Phi(N_a(t)) \leqslant \log(t)\,. \tag{76}$$

From Equation (75) and the monotonic properties of $\mathrm{KL}(\cdot|\cdot)$, we have $\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon) \leqslant \mathrm{KL}(\widehat{\mu}_a|\widehat{\mu}^\star(t))$. Then previous Equation (76) implies

$$N_a(t)\,\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)\wedge\Phi(N_a(t)) \leqslant N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t))\wedge\Phi(N_a(t)) \leqslant \log(t)\,.$$

This implies

$$N_a(t) \leqslant \left(\Phi^{-1}(\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)) \wedge \frac{\log(t)}{\Phi(\log(t))}\right) \vee \frac{\log(t)}{\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)}\,.$$

⋆ *Case 2:* $M > |\mathcal{A}^+|$
Since $\left|\widehat{\mathcal{A}}^+(t)\right| = |\mathcal{A}^+(M)| < M$ and according to `IMED-MB` algorithm, $a_{t+1} = \overline{a}_t$ and the empirical upper bound from Lemma 5 is satisfied,

$$N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t))\wedge\Phi(N_a(t)) \leqslant \log(t)\,. \tag{77}$$

From Equation (75) and the monotonic properties of $\mathrm{KL}(\cdot|\cdot)$, we have $\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon) \leqslant \mathrm{KL}(\widehat{\mu}_a|\widehat{\mu}^\star(t))$. Then previous Equation (77) implies

$$N_a(t)\,\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)\wedge\Phi(N_a(t)) \leqslant N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t))\wedge\Phi(N_a(t)) \leqslant \log(t)\,.$$

This implies

$$N_a(t) \leqslant \left(\Phi^{-1}(\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)) \wedge \frac{\log(t)}{\Phi(\log(t))}\right) \vee \frac{\log(t)}{\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)}\,.$$

□

Lemma 23. *We assume $M \geqslant |\mathcal{A}^+|$. Then, under `IMED-MB`, for all $0<\varepsilon<\varepsilon_\mu$, at each time step $t \notin \mathcal{U}(\varepsilon)$ such that $a_{t+1} = a \in \mathcal{A}^+ - \{a^\star\}$ and $a_{t+1} = \widehat{a}_t^\star$,*

$$N_a(t) \leqslant \max_{a' \notin \mathcal{A}^+ \cup \mathcal{V}_{\mathcal{A}^+}} F\left(\left(\Phi^{-1}(\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)) \wedge \frac{\Psi^{-1}(\log(t))}{\Phi(\Psi^{-1}(\log(t)))}\right) \vee \frac{\Psi^{-1}(\log(t))}{\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)} + |\mathcal{U}_a(\varepsilon)| + 1\right)$$

where $F : x > 0 \mapsto e^{x\Phi(x)+\log(x)}$ and $\Phi^{-1} : y \geqslant 0 \mapsto \max\{x \geqslant 0 : \Phi(x) \leqslant y\}$ (with the convention $\max\emptyset = 0$).

Proof. Since $t \notin \mathcal{U}(\varepsilon) = \bigcup_{a\in\mathcal{A}} \mathcal{E}_a(\varepsilon) \cup \mathcal{E}_a(\varphi,\varepsilon) \cup \mathcal{K}_a^-(\Phi,\varepsilon_\mu)$ and $a_{t+1} = \widehat{a}_t^\star$, Lemma 16 implies

$$\left|\widehat{\mathcal{A}}^+(t)\right| = M, \qquad \widehat{\mathcal{A}}^+(t) - \mathcal{A}^+(M) = \widehat{\mathcal{A}}^+(t) - \mathcal{A}^+ \neq \emptyset\,.$$

Since $\widehat{\mathcal{A}}^+(t) - \mathcal{A}^+ \neq \emptyset$ is equivalent to $\widehat{\mathcal{A}}^+(t) \cup \mathcal{V}_{\widehat{\mathcal{A}}^+(t)} - \mathcal{A}^+ \cup \mathcal{V}_{\mathcal{A}^+} \neq \emptyset$ and $\widehat{\mathcal{A}}^+(t) \cup \mathcal{V}_{\widehat{\mathcal{A}}^+(t)} = \widehat{\mathcal{A}}^M(t)$, then there exists an arm $a' \in \widehat{\mathcal{A}}^M(t) - \mathcal{A}^+ \cup \mathcal{V}_{\mathcal{A}^+}$. Since $\left|\widehat{\mathcal{A}}^+(t)\right| = M$, $a_{t+1} = \widehat{a}_t^\star = a$ and according to `IMED-MB` algorithm,

$$\log(N_a(t)) = \log\left(N_{\widehat{a}_t^\star}(t)\right) = I^\Phi_{\widehat{a}_t^\star}(t) = I^\Phi_{a_{t+1}}(t) = \min_{a'' \in \widehat{\mathcal{A}}^M(t)} I^\Phi_{a''}(t) \leqslant I^\Phi_{a'}(t)\,. \tag{78}$$

Furthermore, we have from the definition of the indexes

$$I^\Phi_{a'}(t) \leqslant N_{a'}(t)\Phi(N_{a'}(t)) + \log(N_{a'}(t))\ , \tag{79}$$

where $a' \notin \mathcal{A}^+ \cup \mathcal{A}$. Then, previous Equations (78)-(79) imply

$$N_a(t) \leqslant \max_{a' \notin \mathcal{A}^+ \cup \mathcal{V}_{\mathcal{A}^+}} F(N_{a'}(t))\ , \tag{80}$$

where $F : x > 0 \mapsto e^{x\Phi(x)+\log(x)}$. Then, we use the upper bounds from Lemma 17 and obtain

$$N_a(t) \leqslant \max_{a' \notin \mathcal{A}^+ \cup \mathcal{V}_{\mathcal{A}^+}} F\left(\left(\Phi^{-1}(\mathrm{KL}(\mu_a + \varepsilon | \mu^\star - \varepsilon)) \wedge \frac{\Psi^{-1}(\log(t))}{\Phi(\Psi^{-1}(\log(t)))}\right) \vee \frac{\Psi^{-1}(\log(t))}{\mathrm{KL}(\mu_a + \varepsilon | \mu^\star - \varepsilon)} + |\mathcal{U}_a(\varepsilon)| + 1\right).$$

□

D Proofs of Theorem 2

Let us consider functions Φ and Ψ such that $1 \leqslant \Phi(\log(n)) \leqslant \log\log(n)$, for $n \geqslant 18$, and $\Psi(x) \geqslant \max\{x\,; \exp(x^\alpha)\}$, for $x \geqslant 0$ and some fixed constant $\alpha > 1$. Then, $\Psi^{-1}(\log(t)) = \underset{t\to\infty}{o}(\log(t))$, for all $U > 0$,

$$\max_{a' \notin \mathcal{A}^+ \cup \mathcal{V}_{\mathcal{A}^+}} F\left(\left(\Phi^{-1}(\mathrm{KL}(\mu_a + \varepsilon | \mu^\star - \varepsilon)) \wedge \frac{\Psi^{-1}(\log(t))}{\Phi(\Psi^{-1}(\log(t)))}\right) \vee \frac{\Psi^{-1}(\log(t))}{\mathrm{KL}(\mu_a + \varepsilon | \mu^\star - \varepsilon)} + U\right) = \underset{t\to\infty}{o}(\log(t))\,,$$

where $F : x \geqslant 0 \mapsto e^{x\Phi(x)+\log(x)}$, and Theorem 2 is proven by combining Lemma 18 and Lemma 21.

E Non-reliable current means

In this section, we define and study relevant subsets of time steps for which the current mean of a specific arm is not reliable. Note that the definitions and the stated properties of these subsets of time steps are independent from the considered algorithms.

For all arm $a \in \mathcal{A}$ and for all accuracy $\varepsilon > 0$, let $\mathcal{E}^+_a(f, \varepsilon)$ be the set of times where the current mean of arm a ε-deviates from above while arm a has more pulls than any function f of the current pulled arm,

$$\mathcal{E}^+_a(f, \varepsilon) \coloneqq \left\{t \in [\![1, T-1]\!] :\ N_a(t) \geqslant f\left(N_{a_{t+1}}(t)\right),\ \widehat{\mu}_a(t) \geqslant \mu_a + \varepsilon\right\}. \tag{81}$$

We similarly define

$$\mathcal{E}^-_a(f, \varepsilon) \coloneqq \left\{t \in [\![1, T-1]\!] :\ N_a(t) \geqslant f\left(N_{a_{t+1}}(t)\right),\ \widehat{\mu}_a(t) \leqslant \mu_a - \varepsilon\right\}. \tag{82}$$

We also define

$$\mathcal{E}_a(f,\varepsilon) = \mathcal{E}_a^+(f,\varepsilon) \cup \mathcal{E}_a^-(f,\varepsilon)\,. \tag{83}$$

When function f is equal to idendity function, we respectively write $\mathcal{E}_a^+(\varepsilon)$, $\mathcal{E}_a^-(\varepsilon)$, and $\mathcal{E}_a(\varepsilon)$ instead of $\mathcal{E}_a^+(f,\varepsilon)$, $\mathcal{E}_a^-(f,\varepsilon)$, and $\mathcal{E}_a(f,\varepsilon)$.

Definition 2 (KL-log deviation)**.** *Let Φ be a positive non-decreasing function. For $\varepsilon>0$, arm $a\in\mathcal{A}$ shows (Φ,ε^-)-*KL-log *deviation at time step $t\geqslant 1$ if the following conditions are satisfied*

$$\begin{aligned}&(1)\quad \widehat{\mu}_a(t) \leqslant \mu_a - \varepsilon\\ &(2)\quad N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\mu_a-\varepsilon)\wedge\Phi(N_a(t)) + \log(N_a(t)) \geqslant \log\big(N_{a_{t+1}}(t)\big)\,.\end{aligned}$$

For all arm $a\in\mathcal{A}$ and for all accuracy $\varepsilon>0$, let $\mathcal{K}_a^-(\Phi,\varepsilon)$ be the set of times where arm a shows (Φ,ε^-)-KL-log deviation, that is

$$\mathcal{K}_a^-(\Phi,\varepsilon) \coloneqq \left\{ t\in[\![1,T-1]\!] : \begin{array}{ll}(1) & \widehat{\mu}_a(t) \leqslant \mu_a - \varepsilon\\ (2) & N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\mu_a-\varepsilon)\wedge\Phi(N_a(t)) + \log(N_a(t)) \geqslant \log\big(N_{a_{t+1}}(t)\big)\end{array}\right\}. \tag{84}$$

Let us consider for $K>0$,

$$\mathcal{I}(K) = \left\{ t\geqslant 1 :\ N_{a_{t+1}}(t) \leqslant 17 \vee \frac{\log\big(N_{a_{t+1}}(t)\big) + 2\log\log\big(N_{a_{t+1}}(t)\big)}{K} \right\}, \tag{85}$$

the subset of time steps for which the number of pulls of current pulled arm is relatively small. Then, it can be shown that

$$\mathcal{I}(K) \subset \left\{ t\geqslant 1 :\ N_{a_{t+1}}(t) \leqslant 18 \vee \frac{1}{K} \right\}.$$

We can now resort to concentration arguments in order to control the size of these sets, which yields the following upper bounds.

Lemma 24 (Bounded subsets of times)**.** *Let f be a non-negative increasing function and Φ be a non-negative non-decreasing function such that $1\leqslant\Phi(\log(n))\leqslant\log\log(n)$ for $n\geqslant 18$. For $\varepsilon>0$, for $a\in\mathcal{A}$, for $K>0$,*

$$\mathbb{E}_\nu[|\mathcal{I}(K)|] \leqslant \big(1+18\vee K^{-1}\big)\,|\mathcal{A}|\,,$$

$$\mathbb{E}_\nu\big[\big|\mathcal{E}_a^+(f,\varepsilon)\big|\big]\,,\ \mathbb{E}_\nu\big[\big|\mathcal{E}_a^-(f,\varepsilon)\big|\big] \leqslant 1 + \sum_{n\geqslant 1}\exp(-f(n)\mathrm{KL}(\mu_a-\varepsilon|\mu_a))\,,$$

$$\mathbb{E}_\nu\big[\big|\mathcal{E}_a^+(\varepsilon)\big|\big]\,,\ \mathbb{E}_\nu\big[\big|\mathcal{E}_a^-(\varepsilon)\big|\big] \leqslant \frac{1}{1-e^{-\mathrm{KL}(\mu_a-\varepsilon|\mu_a)}},$$

$$\mathbb{E}_\nu\big[\big|\mathcal{K}_a^-(\Phi,\varepsilon)-\mathcal{E}_a^-(f_{a,\varepsilon},\varepsilon)-\mathcal{I}_a(\varepsilon)\big|\big] \leqslant \frac{e^{1+\mathrm{KL}(\mu_a-\varepsilon|\mu_a)}}{\mathrm{KL}(\mu_a-\varepsilon|\mu_a)} \sum_{n\geqslant 18}\frac{\big(1+\log^2(n)\big)\big(\log(n)+2\log\log(n)\big)}{n^{1+\mathrm{KL}(\mu_a-\varepsilon|\mu_a)/\log\log(n)}} < \infty\,,$$

where $f_{a,\varepsilon}(n) = \frac{\log(n)+2\log\log(n)}{\mathrm{KL}(\mu_a-\varepsilon|\mu_a)}$ for $n\geqslant e$ and $\mathcal{I}_a(\varepsilon) = \mathcal{I}\big(e^{-1}\mathrm{KL}(\mu_a-\varepsilon|\mu_a)\big)$.

Proof. We start by proving the upper bound on $\mathbb{E}_\nu[|\mathcal{E}_a^-(f,\varepsilon)|]$. The proof of the upper bound on $\mathbb{E}_\nu[|\mathcal{E}_a^+(f,\varepsilon)|]$ is similar.

For $a \in \mathcal{A}$ and $n \geqslant 0$, we define $\tau_a(n) = \inf\{t \geqslant 0 : \ N_a(t) = n\}$ as the first time step arm a is pulled n times, with the conventions $N_a(0) = 0$, $\widehat{\mu}_a(0) = 0$. Then we write

$$\begin{aligned}
|\mathcal{E}_a^-(f,\varepsilon)| &\leqslant \sum_{t\geqslant 0} \mathbb{I}_{\left\{f\left(N_{a_{t+1}}(t)\right)\leqslant N_a(t),\ \widehat{\mu}_a(t)\leqslant \mu_a-\varepsilon\right\}} \\
&\leqslant \sum_{n\geqslant 0}\sum_{t\geqslant 0} \mathbb{I}_{\left\{N_{a_{t+1}}(t)=n,\ f(n)\leqslant N_a(t),\ \widehat{\mu}_a(t)\leqslant \mu_a-\varepsilon\right\}} \\
&= \sum_{n\geqslant 0}\sum_{t\geqslant 0} \mathbb{I}_{\left\{t+1=\tau_{a_{t+1}}(n+1),\ f(n)\leqslant N_a(t),\ \widehat{\mu}_a(t)\leqslant \mu_a-\varepsilon\right\}} \ \text{(since } N_{a_{t+1}}(t+1)=N_{a_{t+1}}(t)+1\text{)} \\
&= \sum_{n\geqslant 0}\sum_{t\geqslant 0} \mathbb{I}_{\left\{t=\tau_{a_{t+1}}(n+1)-1,\ f(n)\leqslant N_a(t),\ \widehat{\mu}_a(t)\leqslant \mu_a-\varepsilon\right\}}.
\end{aligned}$$

We note that for $n \geqslant 0$, since only one arm is pulled at each time step, the $\tau_a(n+1)$, for $a \in \mathcal{A}$, are all different. Furthermore, if $f(n) > 0$, $N_a(t) \geqslant f(n)$ implies $N_a(t) \geqslant 1$ and $t \geqslant 1$. The last upper bound on $|\mathcal{E}_a^-(f,\varepsilon)|$ then implies

$$\left|\mathcal{E}_a^-(f,\varepsilon)\right| \leqslant 1 + \sum_{n\geqslant 1} \mathbb{I}_{\left\{\exists t\geqslant 1,\ f(n)\leqslant N_a(t),\ \widehat{\mu}_a(t)\leqslant \mu_a-\varepsilon\right\}} \tag{86}$$

Taking the expectation of Equation (86), it comes

$$\mathbb{E}_\nu\left[\left|\mathcal{E}_a^-(f,\varepsilon)\right|\right] \leqslant 1 + \sum_{n\geqslant 1} \mathbb{P}_\nu\left(\bigcup_{\substack{t\geqslant 1\\ N_a(t)\geqslant f(n)}} \widehat{\mu}_a(t) \leqslant \mu_a - \varepsilon\right). \tag{87}$$

From Proposition 2, previous Equation (87) implies

$$\mathbb{E}_\nu\left[\left|\mathcal{E}_a^-(f,\varepsilon)\right|\right] \leqslant 1 + \sum_{n\geqslant 1} \exp(-f(n)\,\mathrm{KL}(\mu_a-\varepsilon|\mu_a)). \tag{88}$$

We now show the upper bound on $\mathbb{E}_\nu[|\mathcal{K}_a^-(\Phi,\varepsilon) - \mathcal{E}_a^-(f_{a,\varepsilon},\varepsilon) - \mathcal{I}_a(\varepsilon)|]$.

Let $t \in \mathcal{K}_a^-(\Phi,\varepsilon) - \mathcal{E}_a^-(f_{a,\varepsilon},\varepsilon) - \mathcal{I}_a(\varepsilon)$. There exists an unique $n \geqslant 0$ such that $t+1 = \tau_{a_{t+1}}(n+1)$. In particular, $N_{a_{t+1}}(t) = n$.
Since $t \notin \mathcal{I}_a(\varepsilon)$, we have $N_{a_{t+1}}(t) \geqslant 18$ and $N_{a_{t+1}}(t) \geqslant e f_{a,\varepsilon}\left(N_{a_{t+1}}(t)\right)$, that is, $n \geqslant 18$ and $n \geqslant e\,M_n$, with $M_n = f_{a,\varepsilon}(n)$.
Since $t \in \mathcal{K}_a^-(\Phi,\varepsilon) - \mathcal{E}_a^-(f_{a,\varepsilon},\varepsilon)$ then $\widehat{\mu}_a(t) \leqslant \mu_a - \varepsilon$, $N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\mu_a-\varepsilon) \wedge \Phi(N_a(t)) + \log(N_a(t)) \geqslant \log\left(N_{a_{t+1}}(t)\right)$ and $N_a(t) \leqslant f_{a,\varepsilon}\left(N_{a_{t+1}}(t)\right)$, that is, $\widehat{\mu}_a(t) \leqslant \mu_a - \varepsilon$, $N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\mu_a-\varepsilon) \wedge \Phi(N_a(t)) + \log(N_a(t)) \geqslant \log(n)$ and $N_a(t) \leqslant M_n = f_{a,\varepsilon}(n)$.

Thus, the following inequality holds

$$\begin{aligned}
&|\mathcal{K}_a^-(\Phi,\varepsilon) - \mathcal{E}_a^-(f_{a,\varepsilon},\varepsilon) - \mathcal{I}_a(\varepsilon)| \\
\leqslant\ &\sum_{n\geqslant 0}\sum_{t\geqslant 0} \mathbb{I}_{\{n\geqslant 18,\ n\geqslant e f_a(n)\}} \mathbb{I}\left\{\begin{array}{l} t = \tau_{a_{t+1}}(n+1) - 1, \\ \widehat{\mu}_a(t) \leqslant \mu_a - \varepsilon, \\ N_a(t) \leqslant M_n, \\ N_a(t)\,\mathrm{KL}\left(\widehat{\mu}_a(t)|\mu_a - \varepsilon\right) \wedge \Phi\left(N_a(t)\right) + \log(N_a(t)) \geqslant \log(n) \end{array}\right\},
\end{aligned}$$

which implies

$$\begin{aligned} &|\mathcal{K}_a^-(\Phi,\varepsilon) - \mathcal{E}_a^-(f_{a,\varepsilon},\varepsilon) - \mathcal{I}_a(\varepsilon)| \\ \leqslant \quad & \sum_{n\geqslant 18} \mathbb{I}_{\{n\geqslant 18,\ n\geqslant e f_a(n)\}} \mathbb{I}\left\{ \begin{array}{l} \exists t\geqslant 1,\ \widehat{\mu}_a(t)\leqslant \mu_a-\varepsilon,\ 1\leqslant N_a(t)\leqslant M_n, \\ N_a(t)\,\mathrm{KL}\big(\widehat{\mu}_a(t)|\mu_a-\varepsilon\big)\wedge \Phi\left(N_a(t)\right)+\log(N_a(t))\geqslant \log(n). \end{array} \right\} \end{aligned} \tag{89}$$

Taking the expectation of Equation (89), it comes

$$\begin{aligned} &\mathbb{E}_\nu[|\mathcal{K}_a^-(\Phi,\varepsilon) - \mathcal{E}_a^-(f_{a,\varepsilon},\varepsilon) - \mathcal{I}_a(\varepsilon)|] \\ \leqslant \quad & \sum_{\substack{n\geqslant 18 \\ n\geqslant e\, f_a(n)}} \mathbb{P}_\nu\left(\bigcup_{\substack{t\geqslant 1 \\ \widehat{\mu}_a(t)\leqslant \mu_a-\varepsilon \\ 1\leqslant N_a(t)\leqslant M_n}} N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\mu_a-\varepsilon)\wedge\Phi\left(N_a(t)\right)+\log(N_a(t))\geqslant\log(n) \right). \end{aligned} \tag{90}$$

From Theorem 1, previous Equation (90) implies

$$\begin{aligned} &\mathbb{E}_\nu[|\mathcal{K}_a^-(\Phi,\varepsilon) - \mathcal{E}_a^-(f_{a,\varepsilon},\varepsilon) - \mathcal{I}_a(\varepsilon)|] \\ \leqslant \quad & \sum_{\substack{n\geqslant 18 \\ n\geqslant e\, f_a(n) \\ m_n\leqslant M_n}} e\,\left(1+\log(M_n/m_n)\log(n/M_n)\right) M_n\, n^{-1}\exp(-m_n\,\mathrm{KL}(\mu_a-\varepsilon|\mu_a))\,, \end{aligned}$$

where $m_n = \dfrac{\log(n)-\log\log(n)}{\Phi(\log(n))}$ and $M_n = f_{a,\varepsilon}(n) \coloneqq \dfrac{\log(n)+2\log\log(n)}{\mathrm{KL}(\mu_a-\varepsilon|\mu_a)}$. Since it is assumed that $\Phi(x)\leqslant\log(x)$ for $x\geqslant 1$, then $m_n\geqslant \log(n)/\log\log(n)-1$ and

$$\begin{aligned} &\mathbb{E}_\nu[|\mathcal{K}_a^-(\Phi,\varepsilon) - \mathcal{E}_a^-(f_{a,\varepsilon},\varepsilon) - \mathcal{I}_a(\varepsilon)|] \\ \leqslant \quad & \frac{e^{1+\mathrm{KL}(\mu_a-\varepsilon|\mu_a)}}{\mathrm{KL}(\mu_a-\varepsilon|\mu_a)} \sum_{n\geqslant 18} \frac{\left(1+\log^2(n)\right)\left(\log(n)+2\log\log(n)\right)}{n}\, n^{-\mathrm{KL}(\mu_a-\varepsilon|\mu_a)/\log\log(n)}\,, \end{aligned}$$

where $\log^5(n)\, n^{-\mathrm{KL}(\mu_a-\varepsilon|\mu_a)/\log\log(n)} = \exp\left(5\log\log(n) - \dfrac{\log(n)}{\log\log(n)}\mathrm{KL}(\mu_a-\varepsilon|\mu_a)\right) = \underset{n\to\infty}{o}(1)$.

□

F Concentration of measurement - Proof of Theorem 1

Proof. Let us consider $t\geqslant 1$ such that $\widehat{\mu}_a(t) < \mu_a-\varepsilon$ and

$$N_a(t)\,\mathrm{KL}(\widehat{\mu}_a|\mu_a-\varepsilon)\wedge\Phi(N_a(t))+\log(N_a(t))\geqslant\log(n)\,. \tag{91}$$

This equation declines in broken down into two. Firstly, this implies

$$N_a(t)\,\Phi(N_a(t))+\log(N_a(t))\geqslant\log(n)\ , \tag{92}$$

which implies in particular

$$N_a(t)\geqslant m_n \coloneqq 1\wedge\frac{\log(n)-\log\log(n)}{\Phi(\log(n))}\,, \tag{93}$$

where $m_n > 0$ since it is assumed that $n \geqslant 18 > e^e$ and $\Phi(\log(n)) \geqslant 1$. Secondly, Equation (91) implies

$$N_a(t)\,\mathrm{KL}(\widehat{\mu}_a|\mu_a-\varepsilon) + \log(N_a(t)) \geqslant \log(n)\,. \tag{94}$$

We note that the Kullback divergence coincides with the Bregman divergence in dimension 1 and apply the generalized Pythagorean theorem with convex compact set $K = [\widehat{\mu}_a(t)\,;\mu_a - \varepsilon]$

$$\mathrm{KL}(\widehat{\mu}_a(t)|\mu_a) \geqslant \mathrm{KL}(\widehat{\mu}_a(t)|\tilde{\mu}) + \mathrm{KL}(\tilde{\mu}|\mu_a)\;, \tag{95}$$

where $\widehat{\mu}_a(t) \in K$ and $\tilde{\mu} \in \arg\min_{\mu\in K} \mathrm{KL}(\mu|\mu_a)$. Since $\mathrm{KL}(\cdot|\mu_a)$ is a decreasing function on K, then $\tilde{\mu} = \mu_a - \varepsilon$ and

$$\mathrm{KL}(\widehat{\mu}_a(t)|\mu_a) \geqslant \mathrm{KL}(\widehat{\mu}_a(t)|\mu_a-\varepsilon) + \mathrm{KL}(\mu_a-\varepsilon|\mu_a)\;. \tag{96}$$

Then, by combining previous Equation (96) and Equation (94), it comes

$$\mathrm{KL}(\widehat{\mu}_a(t)|\mu_a) - \mathrm{KL}(\mu_a-\varepsilon|\mu_a) \geqslant \frac{\log(n/N_a(t))}{N_a(t)}\,, \tag{97}$$

that is,

$$\mathrm{KL}(\widehat{\mu}_a(t)|\mu_a) \geqslant \frac{\log(n/N_a(t))}{N_a(t)} + \mathrm{KL}(\mu_a-\varepsilon|\mu_a)\;. \tag{98}$$

We now resort to peeling by considering the slices $[\![m_n b^k\,; m_n b^{k+1}]\!]$ for $k \in [\![0\,; k_n]\!]$, $k_n = \lfloor \log(M_n/m_n)/\log(b) \rfloor$, $b > 1$, and apply Proposition 2. In particular, previous Equation (98) now implies

$$\mathbb{I}_{\{m_n b^k \leqslant N_a(t) \leqslant m_n b^{k+1}\}} \mathrm{KL}(\widehat{\mu}_a(t)|\mu_a) \geqslant \mathbb{I}_{\{m_n b^k \leqslant N_a(t) \leqslant m_n b^{k+1}\}} \left[\frac{\log(n/M_n)}{m_n b^{k+1}} + \mathrm{KL}(\mu_a-\varepsilon|\mu_a)\;, \right]$$

that is,

$$\mathbb{I}_{\{m_n b^k \leqslant N_a(t) \leqslant m_n b^{k+1}\}} \mathrm{KL}(\widehat{\mu}_a(t)|\mu_a) \geqslant \mathbb{I}_{\{m_n b^k \leqslant N_a(t) \leqslant m_n b^{k+1}\}} \mathrm{KL}(\mu(k)|\mu_a) \tag{99}$$

where $\widehat{\mu}_a(t) \leqslant \mu(k)$ and $\mathrm{KL}(\mu(k)|\mu_a) = \frac{\log(n/M_n)}{m_n b^{k+1}} + \mathrm{KL}(\mu_a-\varepsilon|\mu_a)$. Proposition 2 now implies

$$\begin{aligned}
&\mathbb{P}_\nu \left(\bigcup_{\substack{t\geqslant 1 \\ \widehat{\mu}_a(t) < \mu_a - \varepsilon \\ m_n b^k \leqslant N_a(t) \leqslant m_n b^{k+1}}} \mathrm{KL}(\widehat{\mu}_a(t)|\mu_a) \geqslant \frac{\log(n/N_a(t))}{N_a(t)} + \mathrm{KL}(\mu_a-\varepsilon|\mu_a) \right) \\
\leqslant\;& \mathbb{I}_{\{m_n \leqslant M_n\}} \exp\left(-m_n b^k \left[\frac{\log(n/M_n)}{m_n b^{k+1}} + \mathrm{KL}(\mu_a-\varepsilon|\mu_a) \right] \right) \\
=\;& \mathbb{I}_{\{m_n \leqslant M_n\}} e^{-\log(n/M_n)/b} \exp(-m_n\,\mathrm{KL}(\mu_a-\varepsilon|\mu_a))\,.
\end{aligned}$$

Thus we have shown that, for all $b > 1$,

$$\mathbb{P}_\nu\left(\bigcup_{\substack{t\geqslant 1\\ \widehat{\mu}_a(t)<\mu_a-\varepsilon\\ 1\leqslant N_a(t)\leqslant M_n}} N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\mu_a-\varepsilon)\wedge\Phi(N_a(t))+\log(N_a(t))\geqslant\log(n)\right)$$

$$\leqslant \mathbb{P}_\nu\left(\bigcup_{\substack{t\geqslant 1\\ \widehat{\mu}_a(t)<\mu_a-\varepsilon\\ m_n\leqslant N_a(t)\leqslant M_n}} \mathrm{KL}(\widehat{\mu}_a(t)|\mu_a)\geqslant\frac{\log(n/N_a(t))}{N_a(t)}+\mathrm{KL}(\mu_a-\varepsilon|\mu_a)\right)$$

$$\leqslant \mathbb{I}_{\{m_n\leqslant M_n\}}(1+k_n)e^{-\log(n/M_n)/b}\exp(-m_n\,\mathrm{KL}(\mu_a-\varepsilon|\mu_a))\,.$$

$$\leqslant \mathbb{I}_{\{m_n\leqslant M_n\}}(1+\log(M_n/m_n)/\log(b))e^{-\log(n/M_n)/b}\exp(-m_n\,\mathrm{KL}(\mu_a-\varepsilon|\mu_a))\,.$$

We now set $b = b_n \coloneqq \dfrac{\log(n/M_n)}{\log(n/M_n)-1}$. Then, since it is assumed that $n \geqslant e\,M_n$, we have $b > 1$. Furthermore, $1/\log(b_n) < \log(n/M_n)$, and

$$\mathbb{I}_{\{m_n\leqslant M_n\}}(1+\log(M_n/m_n)/\log(b))\frac{e^{-\log(n)/b}}{\log(b)}\leqslant\mathbb{I}_{\{m_n\leqslant M_n\}}e\,\left(1+\log(M_n/m_n)\log(n/M_n)\right)M_n\,n^{-1}\,.$$

Thus we have shown that,

$$\mathbb{P}_\nu\left(\bigcup_{\substack{t\geqslant 1\\ \widehat{\mu}_a(t)<\mu_a-\varepsilon\\ 1\leqslant N_a(t)\leqslant M_n}} N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\mu_a-\varepsilon)\wedge\Phi(N_a(t))+\log(N_a(t))\geqslant\log(n)\right)$$

$$\leqslant \mathbb{I}_{\{m_n\leqslant M_n\}}e\,\left(1+\log(M_n/m_n)\log(n/M_n)\right)M_n\,n^{-1}\exp(-m_n\,\mathrm{KL}(\mu_a-\varepsilon|\mu_a))\,.$$

□

Proposition 2 (Time-uniform concentration)**.** *For all arm $a\in\mathcal{A}$, for $x<\mu_a$, $m\geqslant 1$, we have*

$$\mathbb{P}_\nu\left(\bigcup_{\substack{t\geqslant 1\\ N_a(t)\geqslant m}}\widehat{\mu}_a(t)<x\right)\leqslant\exp(-m\,\mathrm{KL}(x|\mu_a))\,.$$