Structure adaptation in bandit theory

Hassan Saber

Université de Lille
École Doctorale MADIS

Thèse de Doctorat

Spécialité **Informatique**

présentée par

Hassan SABER

Structure Adaptation in Bandit Theory

Adaptation à la Structure en Théorie des Bandits

sous la direction de **Odalric-Ambrym Maillard**

Soutenue publiquement le **19 décembre 2022** à **Villeneuve-d'Ascq**, devant le jury composé de

Gilles **Stoltz**	Directeur de recherche, CNRS	Rapporteur
Junya **Honda**	Professeur associé, Université de Kyoto	Rapporteur
Arnak **Dalalyan**	Professeur affilié, ENSAE/CREST	Examinateur & Président
Alexandre **Proutière**	Professeur, KTH Royal Institute of Technology	Examinateur
Claire **Vernade**	Chargée de recherche, DeepMind London	Examinatrice
Odalric-Ambrym **Maillard**	Chargé de recherche (HDR), Inria	Directeur de thèse
Richard **Combes**	Maître de conférence, CentraleSupélec	Invité

Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL),
UMR 9189 Équipe Scool, 59650, Villeneuve d'Ascq, France

Université de Lille — Inria — CRIStAL Centre de Recherche en Informatique, Signal et Automatique de Lille

Acknowledgements

I thank my advisor Odalric-Ambym Maillard, my colleagues and managers at Inria for allowing and offering me the best conditions to persue my PhD. I thank all those who encouraged me, helped me, motivated me, loved me. Those who knew how to correct me when it was necessary, those who knew how to be patient when it was necessary. I thank the staff of the university of Lille. I thank the organizers of conferences, workshops and summer schools in which I was allowed to participate and share my work. Finally, I thank the members of the jury for offering me the opportunity to defend my thesis and compete for the title of doctor.

Remerciements

Je réserve cette section à mon encadrant et directeur de thèse pour des remerciements que je souhaite en français. Odalric est un homme avec beaucoup d'humanité et dont je pense avoir beaucoup appris. Bien que nos relations soient essentiellement académiques et profesionnelles, il me laisse voir en sa personnalité une personne passionnée et talentueuse, un personnage robuste et raisonné. C'est pour moi un bel exemple d'accomplisssement professionnel. Un accomplissement qui donne sens au métier de chercheur, un accomplissement en devenir. Pour tout cela, Odalric, je te remercie et te souhaite le meilleur des jours à venir.

Abstract

Understanding the dynamics of complex systems, and how to optimally act in them impacts all aspects of human societies where a careful management of natural, energetic, human and computational resources is required. To overcome the limitations of human capabilities to process large amounts of data, researchers from the field of machine learning and mathematical statistics for sequential decision making pursue the long-term goal of developing an optimal and automatic method that can, from partial observations and sequential interactions with a complex system, learn an optimal behavior. While optimal control considers the dynamics of the system is assumed to be known, reinforcement learning is interested in the case when the dynamics is unknown and must be learned from observations only. A key difficulty to design a solution for these problems is that typically, when a decision is made, one only gets to see a noisy effect of that decision, and little about the effect of other alternatives. This gives rise to the study of the fundamental exploration- exploitation trade-off: Shall we follow a algorithm that has been used a lot in the past and has empirically proven good until now (exploitation), or shall we explore a less known but potentially promising algorithm (exploration)? Addressing this trade-off does not yield the same approach depending on the underlying structure of the dynamical system, where structure can be understood in several ways. Since not taking advantage of this (possibly hidden) structure is prone to obtaining loose algorithms, it is crucial to investigate how to build learning algorithms that can be adaptive to it. In this thesis, we want to better understand how the notion of structure modifies the learning guarantees and suggests novel improved algorithms in the context of bandits. We give special attention to the cases of unimodal, multimodal and graph structure. We introduce an algorithm for each of these structures, respectively `IMED-UB`, `IMED-MB` and `IMED-GS`. These algorithms are extensions of the popular Indexed Minimum Empirical Divergence (`IMED`) algorithm from Honda and Takemura (2015) to the considered structures. We provide a finite time analysis for each of them and prove their asymptotic optimality. In particular, we considered new exploring mechanisms (second order exploration allowed by `IMED` approach) and developed new tools (concentrations inequalities) we think that are are of independent interest for the bandit community. Furthermore, these novel algorithms perform well in practice. This is confirmed by numerical illustrations on synthetic data.

Résumé

Dans nos sociétés modernes, une gestion minutieuse des ressources naturelles, énergétiques, humaines et informatiques est nécessaire. Comprendre les dynamiques de systèmes complexes et gérer les interactions de manière optimale constituent un enjeu majeur. Pour surmonter la limite des capacités humaines à traiter de grandes quantités de données, les chercheurs en apprentissage automatique et statistiques mathématiques pour la prise de décision séquentielle ont pour objectif de développer une méthode automatique et optimale pouvant, à partir d'observations partielles et d'interactions séquentielles avec un système complexe, apprendre un comportement optimal. Alors que le contrôle optimal considère que la dynamique du système est supposée connue, l'apprentissage par renforcement s'intéresse au cas où la dynamique est inconnue et doit être apprise uniquement à partir d'observations. Une difficulté clé pour concevoir une solution à ces problèmes est que, lorsqu'une décision est prise, seulement un effet bruité de cette décision est observée. Et cela renseigne peu sur les décisions alternatives. Cela donne lieu à l'étude du compromis fondamental entre exploration et exploitation: doit-on suivre un algorithme déjà beaucoup utilisé par le passé et empiriquement viable (exploitation) ou doit-on explorer un algorithme moins connue mais potentiellement meilleur (exploration)? Aborder ce compromis donne lieu à diverses approches en fonction de la structure sous-jacente au système dynamique, la notion de structure pouvant être comprise de différentes façons. Puisque ne pas tirer avantage de la structure (éventuellement cachée) est susceptible d'occasionner des algorithmes largement sous-optimaux, il est crucial d'examiner la manière de concevoir des algorithmes d'apprentissage qui peuvent s'adapter à celle-ci. Dans cette thèse, nous souhaitons mieux comprendre comment la notion de structure modifie les garanties d'apprentissage et suggère de nouveaux algorithmes plus performants dans le contexte des bandits. Notre attention sera particulièrement portée sur les structures dites unimodal, multimodal et de graphe. Pour chacune de ces structures nous introduisons un algorithme, respectivement `IMED-UB`, `IMED-MB` et `IMED-GS`, dont nous prouvons l'optimalité asymptotique et dont nous étudions les performances à horizon fini. Ces algorithmes vise à adapter l'algorithme `IMED` introduit par Honda and Takemura (2015) pour le cas des bandits non-structurés au cas des bandits unimodaux, multimodaux et structure de graphe. En outre, l'étude de ces algorithmes, nous a amené à développer des outils nouveaux (inégalités de concentration) et considérer de nouveaux méchanismes d'exploration (exploration de second ordre permise par l'approche `IMED`) dont nous pensons qu'ils bénéficieront à la communauté. Enfin, nous illustrons par des simulations l'éfficacité pratique de ces nouveaux algorithmes.

Contents

List of Symbols

$\mathcal{A}$	Set of arms
$\mathcal{D}$	Set of one-dimensional exponential family distributions
ν	Distribution in $\mathcal{D}$
I	Interval in $\mathbb{R}$
μ	Means of distribution ν with values in $\mathrm{I}^{\mathcal{A}}$
$\mu^\star$	Maximum of means μ
Δ_a	Gap between maximal mean $\mu^\star$ and mean μ_a of arm $a \in \mathcal{A}$
$\mathbb{P}_\nu(\cdot)$	Probability measure under distribution ν
$\mathbb{E}_\nu[\cdot]$	Expectation under distribution ν
T	Horizon
$N_a(t)$	Number of pulls of arm $a \in \mathcal{A}$ at time step $t \geqslant 1$
$R(\nu, T)$	Averaged regret
$\mathfrak{C}_{\mathcal{D}}(\mu)$	Constant solution of the constrained linear-optimization problem from the regret lower bound
$\mathrm{KL}(\cdot\|\cdot)$	Kullback-Leibler divergence between one-dimensional exponential family distributions
$\mathrm{kl}(\cdot\|\cdot)$	Kullback-Leibler divergence between Bernoulli distributions
$\widehat{\mu}_a(t)$	Empirical mean of arm $a \in \mathcal{A}$ at time step $t \geqslant 1$
$\widehat{\mu}^\star(t)$	Maximum of empirical means at time step $t \geqslant 1$
$I_a(t)$	`IMED` index of arm $a \in \mathcal{A}$ at time step $t \geqslant 1$
$\mathcal{E}^-_{a,a'}(\varepsilon)$	Set of time steps where the current mean of arm a ε-deviates from below while $N_a(t) \geqslant N_{a'}(t)$
$\mathcal{K}^-_{a,a'}(\varepsilon)$	Set of time steps where couple of arms (a, a') shows ε^--KL-$\log$ deviation
$\mathcal{D}_{\text{uni}}$	Set of distributions for the unimodal structure
$\mathcal{D}_{\text{M-modal}}$	Set of distributions for the multimodal structure
Θ	Set of means in $(0,1)^{\mathcal{A}}$
$\mathcal{D}_\Theta$	Set of Bernoulli distributions for graph structure

Chapter 1

Introduction

1.1 The stochastic multi-armed bandit problem

The multi-armed bandit problem is a popular framework to formalize sequential decision making problems. It was first introduced in the context of medical trials (Thompson, 1933, 1935) and later formalized by Robbins (1952): A bandit instance is specified by a configuration, that is a set of unknown probability distributions, $\nu = (\nu_a)_{a\in\mathcal{A}}$ 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 $t \in \mathbb{N}$, 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, conditionally independent, sampled according to ν_{a_t}. The goal of the learner is to maximize the expected 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

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

Both means and distributions are *unknown*, which makes the problem non trivial, and the learner only knows that $\nu \in \mathcal{D}$ where $\mathcal{D}$ is a given set of bandit configurations. 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 $\mathcal{D}$ comes with a price. The study of the lower performance bounds in multi-armed bandits successfully lead 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). We refer to Lattimore and Szepesvári (2020) for a recent survey. In this regard, it should be highlighted two another main approaches to solve optimally the stochastic bandit problem: Bayesian algorithm (Thompson, 1933) and algorithms based on re-sampling methods, such as `RB-SDA` introduced in a recent work by Baudry et al. (2020).

Both to propose an effective presentation of some relevant algorithms and start introducing the framework within which our work is included, we assume one-dimensional exponential family distributions and then explain the known lower bounds on the regret from Lai and Robbins (1985) under this assumption.

Assumption 1 (One-dimensional exponential family distributions)**.** *For all $\nu \in \mathcal{D}$, $\nu \subset \mathcal{P}_1 := \{p(\mu), \mu \in \mathrm{I}\}$, where $p(\mu)$ is a 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 $|A(\mu)| < \infty$.

We consider algorithms that are *consistent* in order to obtain non trivial lower bound on the regret.

Definition 1 (Consistent algorithm)**.** *An algorithm is consistent on $\mathcal{D}$ if for all configuration $\nu \in \mathcal{D}$, for all sub-optimal arm $a \notin \mathcal{A}^\star \coloneqq \underset{a\in\mathcal{A}}{\operatorname{argmax}}\, \mu_a$, for all $\alpha > 0$,*

$$\lim_{T\to\infty} \mathbb{E}_\nu\left[\frac{N_a(T)}{T^\alpha}\right] = 0\,,$$

where $N_a(t) = \sum_{s=1}^{t} \mathbb{I}_{\{a_s=a\}}$ is the number of pulls of arm a at time $t \geqslant 0$.

In particular, for $\alpha = 1$, the number of pulls of a sub-optimal arm is sub-linear under a consistent algorithm.

Thanks to the *tower rule*, we have

$$R(\nu, T) = T\mu^\star - \mathbb{E}_\nu\left[\sum_{t=1}^{T} X_t\right] = T\mu^\star - \mathbb{E}_\nu\left[\sum_{t=1}^{T} \mu_{a_t}\right].$$

We can then rewrite the regret as follows.

$$R(\nu, T) = \mathbb{E}_\nu\left[\sum_{t=1}^{T} \mu^\star - \mu_{a_t}\right] = \mathbb{E}_\nu\left[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} \mathbb{I}_{\{a_t=a\}}(\mu^\star - \mu_a)\right] = \sum_{a\in\mathcal{A}} \mathbb{E}_\nu[N_a(T)]\,\Delta_a\,. \tag{1.1}$$

When $\mathcal{D} = \mathcal{P}_1^{\mathcal{A}}$ and under a consistent algorithm, from Lai and Robbins (1985) we have the following lower bounds on the numbers of pulls:

$$\forall a \notin \mathcal{A}^\star, \quad \liminf_{T\to\infty} \frac{\mathbb{E}_\nu[N_a(T)]\ \mathrm{KL}(\mu_a|\mu^\star)}{\log(T)} \geqslant 1\,. \tag{1.2}$$

This implies

$$\liminf_{T\to\infty} \frac{R(\nu, T)}{\log(T)} \geqslant \mathfrak{C}_1(\mu) \coloneqq \sum_{a\in\mathcal{A}} \frac{\Delta_a}{\mathrm{KL}(\mu_a|\mu^\star)}\,, \tag{1.3}$$

where $\mathrm{KL}(\mu|\mu') = \int_{\mathbb{R}} \log(f(x,\mu)/f(x,\mu'))f(x,\mu)\lambda(\mathrm{d}x)$ denotes the Kullback-Leibler divergence between $\nu = p(\mu)$ and $\nu' = p(\mu')$, for $\mu, \mu' \in \mathrm{I}$.

We refer to the proof of Theorem 1 from Garivier et al. (2016) for a proof of Equation (1.2) that provides lower bounds on the numbers of pulls when assuming non-structured bandits and consistent algorithms. Note that we adapt this same proof in Proposition 2 to obtain a lower bound on the regret when assuming multimodal structure (see Chapter 3) and consistent algorithms.

Remark 1. *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 2. *We allow and reserve the notation $\mathrm{kl}(\cdot|\cdot)$ for the Kullback-Leibler divergence between Bernoulli distributions.*

1.2 Optimal algorithms for non-structured bandit problems

We briefly present in this section some of the main *(asymptotically) optimal* algorithms (under Assumption 1) for the classical multi-armed bandit problem. Here, optimal means that $\liminf_{T\to\infty} \dfrac{R(\nu,T)}{\log(T)} \leqslant \mathfrak{C}_1(\mu)$ under these algorithms.

Kullback-Leibler Upper Confidence Bounds (`KLUCB`). For an arm $a \in \mathcal{A}$ and a time step $t \geqslant 1$, the upper confidence bound is defined as follows:

$$U_a(t) = \sup\left\{\lambda \geqslant \widehat{\mu}_a(t) : \quad \mathrm{KL}(\widehat{\mu}_a(t)|\lambda) \leqslant \frac{f(t)}{N_a(t)}\right\}$$

where $f(t) = \log(t) + 3\log\log(t)$ if $t \geqslant 3$, 0 otherwise, with the convention $0/0 = 0$. This upper bound is motivated by the following concentration inequality:

$$\mathbb{P}_\nu(U_a(t) < \mu_a) \leqslant e\left\lceil f(t)\log(t)\right\rceil \exp(-f(t))\,.$$

We refer to Theorem 10 of Garivier and Cappé (2011) for a more general formulation and a proof of this concentration inequality. `KLUCB` algorithm then consists in pulling an arm with maximal upper bound at each time step, $a_{t+1} = \operatorname{argmax}_{a\in\mathcal{A}} U_a(t)$. `KLUCB` algorithm is summarized in Algorithm 1.

Algorithm 1 `KLUCB`

for $t = 0 \dots T-1$ **do**
  Pull $a_{t+1} \in \operatorname{argmax}_{a\in\mathcal{A}} U_a(t)$ (chosen arbitrarily)
end for

Indexed Minimum Empirical Divergence (`IMED`). For an arm $a \in \mathcal{A}$ and a time step $t \geqslant 1$, the `IMED` index is defined as follows:

$$I_a(t) = N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t), \widehat{\mu}_\star(t)) + \log(N_a(t))\ .$$

This quantity can be seen as a transportation cost for "moving" a sub-optimal arm to an optimal one, plus exploration terms (the logarithms of the numbers of pulls). When an optimal arm is considered, the transportation cost is null and it remains only the exploration part. Note that, as stated in Honda and Takemura (2011), $I_a(t)$ is an index in the weaker sense since it cannot be determined only by samples from the pair a but also uses empirical means of current optimal arms. `IMED` is the algorithm that consists in pulling an arm with minimal index at each time step, $a_{t+1} = \operatorname{argmin}_{a\in\mathcal{A}} I_a(t)$. Finite time guarantees for `IMED` algorithm and its optimality are established in Corollary 2 from next Chapter 2. `IMED` algorithm is summarized in Algorithm 2.

Algorithm 2 `IMED`

for $t = 0 \dots T-1$ **do**
  Pull $a_{t+1} \in \operatorname{argmin}_{a\in\mathcal{A}} I_a(t)$ (chosen arbitrarily)
end for

Thomson Sampling (TS). After pulling each arm once, we denote by $\pi_a(t)$ the posterior distribution of arm $a \in \mathcal{A}$ at time $t \geqslant |\mathcal{A}|$ and sample $\tilde{\mu}_a(t)$ from $\pi_a(t)$, what we note $\tilde{\mu}_a(t) \sim \pi_a(t)$. TS algorithm consists then in pulling the arm with the maximal sample, $a_{t+1} = \operatorname*{argmax}_{a \in \mathcal{A}} \tilde{\mu}_a(t)$. Note the choice of the posteriors is crucial. TS algorithm is proven to be asymptotically optimal for one-dimensional exponential family distributions in Korda et al. (2013) when the posteriors are obtained from Jeffreys priors. TS algorithm is summarized in Algorithm 3.

Algorithm 3 TS

```
Pull each arm once
for t = |A| ... T − 1 do
  for a ∈ A do
    μ̃_a(t) ∼ π_a(t)
  end for
  Pull a_{t+1} ∈ argmax_{a∈A} μ̃_a(t) (chosen arbitrarily)
end for
```

Best Empirical Sampled Average (BESA). We restrain ourselves to the case of two arms when $\mathcal{A} = \{1, 2\}$. We refer to Baransi et al. (2014) and Baudry et al. (2020) for an optimal generalization beyond the case of two arms of the following approach. After pulling each arm once, we consider at each time $t \geqslant 2$, the sample size $n(t) = \min\{N_1(t), N_2(t)\}$ and choose arbitrarily two samples $\mathcal{S}_1(t)$ and $\mathcal{S}_2(t)$ of rewards of same size $n(t)$, respectively from arm 1 and arm 2: $\mathcal{S}_a(t) \subset \{X_s : a_s = a, 1 \leqslant s \leqslant t\}$, $|\mathcal{S}_a(t)| = n(t)$, $a \in \{1, 2\}$. The empirical sub-sampled averages can be then computed as follows:

$$\widehat{m}_a(t) = \sum_{X \in \mathcal{S}_a(t)} X/n(t)\,.$$

BESA algorithm consists in pulling an arm with maximal empirical sample average, $a_{t+1} = \arg\max_{i \in \{1,2\}} \widehat{m}_i(t)$. BESA algorithm is summarized in Algorithm 4.

Algorithm 4 BESA

```
Pull each arm once
for a ∈ {1, 2} do
  Choose an arbitrarily n(t)-sized sample S_a(t) of rewards from arm a
  Compute the empirical sampled average m̂_a(t)
end for
Pull a_{t+1} ∈ argmax_{a∈{1,2}} m̂_a(t) (chosen arbitrarily)
```

1.3 Structured configurations

The lower bounds from Lai and Robbins (1985), later extended by Burnetas and Katehakis (1997) did not cover all possible configurations, and in particular *structured* configuration sets were not handled until Agrawal et al. (1989) and then Graves and Lai (1997) established generic lower bounds. Here, structure refers to the fact that pulling an arm may reveals information that enables to refine estimation of other arms. Unfortunately, designing numerical efficient algorithms that are provably optimal remains a challenge for many structures.

The study of specific *structured configuration sets* $\mathcal{D}$ 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: The study of Unimodal structure naturally appears in many contexts, e.g. 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 (2014a); Trinh et al. (2020); Saber et al. (2021a) providing an explicit lower bound and corresponding algorithms. Note that in Saber et al. (2021a), we adapt IMED algorithm to the Unimodal structure simply by narrowing on the current best arm and its neighbourhood for pulling an arm at a given time step. See also Kunne et al. (2020); Gao et al. (2019). Lipschitz bandits were studied in Magureanu et al. (2014); Wang et al. (2020); Lu et al. (2019) while combinatorial structures have been studied e.g. in Kveton et al. (2015); Magureanu (2018), and more recently Cuvelier et al. (2021b). 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). Another body or work focuses on proving asymptotic minimax optimality in the worst-case setting rather than instance-dependent performance bounds, also targeting order optimal rather than exact optimal regret bounds. This is the case for example in Kleinberg et al. (2008) and Bubeck et al. (2008), respectively introducing ZOOMING and HOO algorithms. 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 thesis.

For structured bandit problems and one-dimensional exponential family distributions (Assumption 1), the lower bound on the regret takes the generic form

$$\liminf_{T\to\infty} \frac{R(\nu,T)}{\log(T)} \geqslant \mathfrak{C}_{\mathcal{D}}(\mu)\,, \tag{1.4}$$

where $\mathfrak{C}_{\mathcal{D}}(\mu)$ is a constant, solution of a constrained linear-optimization problem. When there is no structure, $\mathcal{D} = \mathcal{P}_1^{\mathcal{A}}$ and then $\mathfrak{C}_{\mathcal{P}_1^{\mathcal{A}}}(\mu) = \mathfrak{C}_1(\mu)$. We note that

$$\begin{aligned} \mathfrak{C}_1(\mu) \;\coloneqq\; \min_{n\in\mathbb{R}_+^{\mathcal{A}}} \quad & \sum_{a\in\mathcal{A}} n_a \left(\max\mu - \mu_a\right) \\ \text{s.t. } \forall a \notin \operatorname{argmax}\mu, \quad & n_a\, \mathrm{KL}(\mu_a|\max\mu) \geqslant 1\,. \end{aligned} \tag{1.5}$$

The $(n_a)_{a\in\mathcal{A}}$ from (1.5) then appear as the normalized numbers of pulls $(\mathbb{E}_\nu[N_a(T)]\,/\,\log(T))_{a\in\mathcal{A}}$.

In Graves and Lai (1997) a generic algorithm was proposed to solve any structured bandit problems, with however *prohibitive* computational complexity, including the requirement to compute a version of $\mathfrak{C}_{\mathcal{D}}(\mu)$ at each time step. In Combes et al. (2017), another generic algorithm called OSSB (Optimal Structured Stochastic Bandit) is introduced, stepping the path towards generic structure-adaptive bandit algorithms. Although asymptotically optimal, the algorithm often suffers from poor finite-time numerical performances (as the asymptotic regime kicks-in possibly late), and still high computational cost. Inspired by combinatorial structures for which computing $\mathfrak{C}_{\mathcal{D}}(\mu)$ is simply not feasible, a relaxation of the generic constrained optimization problem was recently proposed in Cuvelier et al. (2021a), 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 $\mathcal{D}$. In Van Parys and Golrezaeiand (2020), the authors propose an approach base on convex duality. Motivated by these issues, we make specific efforts to build algorithms that are regret efficient in practice.

We will address several structures such as Unimodal and Lipschitz structures which makes relevant to focus on the following generic algorithm: Optimal Sampling for Structured Bandits (OSSB) from Combes et al. (2017). We will naturally compare the algorithms we introduce to this known algorithm.

Optimal Sampling for Structured Bandits (OSSB). When assuming there is a unique optimal arm $a^\star$ (which is a very common assumption), that is $\mathcal{A}^\star = \{a^\star\}$, and under the assumptions of Theorem 1 from Combes et al. (2017), it is shown that $\mathfrak{C}_{\mathcal{D}}(\mu)$ can be chosen so that

$$\begin{aligned} \mathfrak{C}_{\mathcal{D}}(\mu) \; := \; & \min_{n \in \mathbb{R}_+^{\mathcal{A}}} \quad \sum_{a \in \mathcal{A}} n_a \left(\max \mu - \mu_a \right) \\ & \text{s.t. } \forall \lambda \in \Lambda(\mu), \quad \sum_{a \in \mathcal{A}} n_a \, \mathrm{KL}(\mu_a | \lambda_a) \geqslant 1 \, , \end{aligned} \tag{1.6}$$

where

$$\Lambda(\mu) = \left\{ \mu(\nu') \in \mathrm{I}^{\mathcal{A}} : \begin{array}{ll} (1) & \nu' \in \mathcal{D} \\ (2) & \mu_{a^\star}(\nu') = \mu^\star \\ (3) & \max \mu(\nu') > \mu^\star \end{array} \right\} \tag{1.7}$$

is the set of "confusing" means for means μ of configuration $\nu \in \mathcal{D}$.

Let $(n_a(\mu))_{a \in \mathcal{A}}$ be the solution of the previous optimization problem. Then $n_a(\mu) \log(T)$ indicates the asymptotic number of times sub-optimal arm $a \neq a^\star$ should be played under efficient and consistent algorithms.

Note that for all arm $a \in \operatorname{argmax} \mu$, we have $n_a(\mu) = 0$.

The key idea of OSSB algorithm is then to compute *at each time step* $(n_a(\widehat{\mu}(t)))_{a \in \mathcal{A}}$, the current empirical solution of $\mathfrak{C}_{\mathcal{D}}(\widehat{\mu}(t))$, for $t \geqslant 1$, and explore so that the number of pull of sub-optimal arm $a \neq a^\star$, $N_a(t)$ grows at speed $n_a(\widehat{\mu}(t)) \log(T)$, that is so that $N_a(t) \approx n_a(\widehat{\mu}(t)) \log(T)$.

In detail, OSSB algorithm takes theoretically two parameters $\varepsilon, \gamma > 0$ in input. Then, at each time step $t \geqslant 1$ the algorithm alternate between three phases: exploitation, estimation and exploration.

If for all arm $a \in \mathcal{A}$, we have $N_a(t) \geqslant n_a(\widehat{\mu}(t))(1+\gamma) \log(t)$, then we consider there is enough information to consider the structure is well estimated and we enter an exploitation phase: we pull arbitrarily an arm with maximal current mean, that is $a_{t+1} \in \operatorname*{argmax}_{a \in \mathcal{A}} \widehat{\mu}_a(t)$.

Otherwise, we consider more information is needed to identify the optimal arm. Thus, either we enter an estimation phase and pull arbitrarily an arm with a minimal current number of pulls, that is $a_{t+1} \in \operatorname*{argmin}_{a \in \mathcal{A}} N_a(t)$, or we enter an exploration phase and pull arbitrarily an arm with a minimal current ratio, where current ratio of arm $a \in \mathcal{A}$ is $N_a(t)/n_a(\widehat{\mu}(t))$ if $n_a(\widehat{\mu}(t)) > 0$, $+\infty$ otherwise. Note that the current ratios measure how much the desired constraints $N_a(t) \gtrsim n_a(\widehat{\mu}(t)) \log(t)$, for $a \in \mathcal{A}$, are satisfied.

More precisely, we consider a counter $s(t)$ of the number of times we have not entered an exploitation phase. If $\min_{a \in \mathcal{A}} N_a(t) \leqslant \varepsilon s(t)$, we enter an estimation phase, otherwise we enter an exploration phase.

Finally, we note that parameters ε and γ are set equal to 0 in practice for the numerical experiments.

OSSB algorithm is summarized in Algorithm 5.

Algorithm 5 `OSSB`

Input: $\varepsilon, \gamma > 0$
Pull $a_1 \in \mathcal{A}$ (arbitrarily chosen)
$s(1) \leftarrow 0$
for $t = 1 \dots T-1$ **do**
  Compute $(n_a(\widehat{\mu}(t)))_{a\in\mathcal{A}}$, the solution of current optimization problem $\mathfrak{C}_{\mathcal{D}}(\widehat{\mu}(t))$
  if $N_a(t) \geqslant n_a(\widehat{\mu}(t))(1+\gamma)\log(t)$, for all $a \in \mathcal{A}$ **then**
    $s(t+1) \leftarrow s(t)$
    Pull $a_{t+1} \in \underset{a\in\mathcal{A}}{\operatorname{argmax}}\, \widehat{\mu}_a(t)$ (arbitrarily chosen) ▷ `Exploitation`
  else
    $s(t+1) \leftarrow s(t)+1$
    if $\min\limits_{a\in\mathcal{A}} N_a(t) \leqslant \varepsilon s(t)$ **then**
      Pull $a_{t+1} \in \underset{a\in\mathcal{A}}{\operatorname{argmin}}\, N_a(t)$ (arbitrarily chosen) ▷ `Estimation`
    else
      Pull $a_{t+1} \in \underset{a\in\mathcal{A}}{\operatorname{argmin}}\, N_a(t)/n_a(\widehat{\mu}(t))$ (arbitrarily chosen) ▷ `Exploration`
    end if
  end if
end for

Comment. `OSSB` algorithm from Combes et al. (2017) is a quite simple, natural and generic algorithm for structured bandit problems, which makes it a reference in the field. However, `OSSB` presents this major constraint of solving current optimization problem $\mathfrak{C}_{\mathcal{D}}(\widehat{\mu}(t))$ at each time step $t \geqslant 1$ that limits its practical applications. Furthermore, parameters ε and γ are set equal to 0 in practice without analysis of the algorithm for this choice of setting. That being said, we will retain the main result from Combes et al. (2017) about the asymptotic optimality of `OSSB` algorithm that is reproduced in the following theorem. Note that `OSSB` algorithm is the first algorithm that optimally solve genetic structured bandit problems.

Theorem 1 (`OSSB` optimality)**.** *Let us consider a configuration $\nu \in \mathcal{D}$. If the assumptions of Theorem 2 from Combes et al. (2017) hold, then under `OSSB` algorithm, we have for all $0 < \varepsilon < |\mathcal{A}|^{-1}$,*

$$\limsup_{T\to\infty} \frac{R(\nu,T)}{\log(T)} \leqslant \mathfrak{C}_{\mathcal{D}}(\mu) F(\mu,\varepsilon,\gamma)\,,$$

with F a function such that $F(\mu,\varepsilon,\gamma) \to 1$ as $\varepsilon \to 0$ and $\gamma \to 0$.

1.4 Publications

In this section, we detail the papers published in journals and conferences.

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

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., Saci, L., Maillard, O.-A., and Durand, A. (2021b). Routine bandits: Minimizing regret on recurring problems. *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PPKD)*

1.5 Contributions

We summarize our contributions for structured bandit problems.

Unimodal, multimodal and graph structures are respectively the subject of Chapter 2, 3 and 4. Our work on routine bandits is the object of Chapter 6. These chapters have been written to be largely independent and thus facilitate the reading of the manuscript. Our work on bandits with groups of similar arms is included in Appendix D. The complements from Saber et al. (2020) relating to unimodal bandits are not reproduced in the manuscript.

Unimodal bandits. In the unimodal bandit problem, a graph G supports the structure and allows a notion of neighbourhood between the arms. An arm $a \in \mathcal{A}$ is then considered as a local maximum if its mean μ_a is greater than the means of the arms in its neighbourhood $\mathcal{V}_a \subset \mathcal{A} \setminus \{a\}$, that is $\mu_a > \mu_{a'}$, for all $a' \in \mathcal{V}_a$. In the unimodal bandit problem, we only consider distributions ν with a unique local maximum. This is formalized in Chapter 2, where we provide novel regret minimization results related to the unimodal structure. We first revisit Indexed Minimum Empirical Divergence (IMED) algorithm from Honda and Takemura (2015) introduced for unstructured multi-armed bandits, and adapt it to the unimodal setting. We introduce in Section 2.3 IMED-UB algorithm that is limited to the pulling of the current best arm or their no more than d nearest arms at each time step, with d the maximum degree of nodes in G. Being constructed from IMED, IMED-UB does not require any optimization procedure and does not separate exploration from exploitation rounds. IMED-UB appears to be a *local* algorithm. We prove in Theorem 1 that IMED-UB is asymptotically optimal. Furthermore, this novel algorithm competes with the state-of-the-art algorithms in practice. This is confirmed by numerical illustrations on synthetic data. This is the subject of paper Saber et al. (2021a). In Saber et al. (2020), we introduced and studied KLUCB-UB algorithm, a KLUCB version of IMED-UB. Our KLUCB-UB finite time analysis highlighted strong links between KLUCB and IMED approaches. We further developed d-IMED-UB, an algorithm that behaves like IMED-UB while resorting to a dichotomic second order exploration over all nodes of the graph. This helps quickly identify the best arm within a large set of arms $\mathcal{A}$ by empirical considerations. This second order exploration appears as a new mechanism of exploring sub-optimal arms (and the underlying structure) and seems to be specific to IMED approach. This mechanism operates in IMED-MB algorithm introduced in Chapter 3, where we define the multimodal structure that generalises the unimodal bandit problem. Note that our work from Saber et al. (2020) on KLUCB-UB and d-IMED-UB algorithms is not reproduced in the manuscript.

Multimodal bandits. In the multimodal bandit problem, a graph G supports the structure and allows a notion of neighbourhood between the arms. An arm $a \in \mathcal{A}$ is then considered as a local maximum if its mean μ_a is greater than the means of the arms in its neighbourhood $\mathcal{V}_a \subset \mathcal{A} \setminus \{a\}$, that is $\mu_a > \mu_{a'}$, for all $a' \in \mathcal{V}_a$. In the multimodal bandit problem, we only consider distributions ν with a fixed number of local maximums that is known in advance to the learner. This is formalized in Chapter 3, where we generalize the unimodal structure and precisely define the multimodal bandit problem. We provide regret minimization results related to the multimodal structure. We again revisit the Indexed Minimum Empirical Divergence (IMED) algorithm from Honda and Takemura (2015) introduced for unstructured multi-armed bandits, and adapt it to the multimodal-structured setting. We introduce in Section 3.3 IMED-MB algorithm that is limited to the pulling of the arms with locally maximal empirical means or their no more than d nearest arms at each time step, with d the maximum degree of nodes in G. We do not provide finite time analysis of IMED-MB algorithm, but it could be the focus of future work. This novel algorithm performs well in practice and competes with the state-of-the-art algorithms when assuming unimodal structure. This is confirmed by numerical illustrations on synthetic data. We believe that the construction of this algorithm and the use of IMED type indexes to test the local

maximality of arms with locally maximal empirical means (Equation 3.17) are of independent interest for the bandit community.

Bandits with groups of similar arms. We consider a variant of the stochastic multi-armed bandit problem where arms are known to be organized into different groups having the same mean. The groups are unknown but a lower bound q on their size is known. This situation typically appears when each arm can be described with a list of categorical attributes, and the (unknown) mean reward function only depends on a subset of them, the others being redundant. In this case, q is linked naturally to the number of attributes considered redundant, and the number of categories of each attribute. For this structured problem of practical relevance, we first derive the asymptotic regret lower bound and corresponding constrained optimization problem. They reveal the achievable regret can be substantially reduced when compared to the unstructured setup, possibly by a factor q. However, solving exactly the exact constrained optimization problem involves a combinatorial problem. Owing to this key insight, we introduce `IMED-EC`, an adaptation of `IMED` algorithm from Honda and Takemura (2015) to the considered structured set of bandits. One advantage of `IMED` over a `KLUCB` alternative is its reduced complexity, which translates to the equivalence class setup. At each time step, the complexity of computing the next arm to be pulled by `IMED-EC` is no more than the one of sorting a list of $|\mathcal{A}|$ elements once the `IMED` indexes have been computed, which is only $\log(|\mathcal{A}|)$ times larger than looking for the minimal `IMED` index. We prove that `IMED-EC` achieves a controlled asymptotic regret that matches the non-combinatorial part of the lower bound and is at most (less than) a factor of 2 times the optimal regret bound. Last, we illustrate the benefit of the `IMED-EC` over its unstructured version, where it shows a substantial improvement. This work is the subject of Pesquerel et al. (2021) and has been initiated by Fabien Pesquerel.

Graph-structured bandits. In Chapter 4, we consider a structured variant of the multi-armed bandit problem when the difference of means between any pair of arms is constrained not to exceed some value. This graph structure is introduced to encompass some classical structures such as Unimodal and Lipschitz. We assume Bernoulli distributions although most of this work can be generalised to one-dimensional exponential family distributions. In Section 4.2, we first provide an instance-dependent lower bound on the regret for generic graph-structured bandits, see Proposition 3. The proof follows standard change-of-measure techniques and reveals interesting sets of "confusing" bandit instances. In Section 4.3, building on `IMED` algorithm originally introduced for unstructured bandits by Honda and Takemura (2015), we introduce `IMED-GS` algorithm, to which we incorporate some key modifications in order to adapt it to a graph-structure. This `IMED` type approach directly builds on the lower bounds. We detail `IMED-GS` in Section 4.3.1 and Algorithm 9. The main theoretical result is Theorem 3 that provides a *finite-time* upper bound on the regret under `IMED-GS`, which then implies asymptotic optimality as shown in Corollary 3. The proof of this result, detailed in Section 4.5, involves a technique exploiting *empirical lower bounds*. It also involves a novel concentration inequality of independent interest that combines stochastic orderings with approximations. In Section 4.4, we illustrate and discuss the benefit of this approach over the state-of-the-art regarding instance-dependent optimality on some specific structures, including Unimodal and Lipschitz bandits.

Concentration inequalities. In Chapter 5, we highlight new concentration inequalities introduced to provide a finite time analysis of `IMED-GS` algorithm and prove its asymptotic optimality for graph-structured bandits (Chapter 4). These concentration inequalities enable to obtain more precise control of the concentration terms compared to alternative tools present in the literature like in Theorem 2 from Magureanu et al. (2014).

Routine bandits. In the online recommender system problem items are recommended to users. Items can be seen as arms and users as bandit instances. When a recommender system is deployed on multiple users, one does not typically assume that the best recommendation is the same for all users. The naive algorithm in this

situation is to consider each user as being a different bandit instance and learning from scratch for each user. When users can be recognized (e.g., characterized by features), this information can be leveraged to speed up the learning process by sharing observations across users. The resulting setting is known as *contextual bandit* (Langford and Zhang, 2007; Lu et al., 2010). We tackle the case where users cannot be or do not want to be identified (e.g., for privacy reasons), but where we assume that there exists a (unknown) finite set of possible user profiles (bandit instances), such that information may be shared between the current user and some previously encountered users. We call the resulting setting as *routine bandit* (Saber et al., 2021b). We establish lower bounds on the achievable cumulative regret that adapt the bound from Lai and Robbins (1985) to the routine setting. We then extend `KLUCB` algorithm, known to be optimal under the classical stochastic bandit setting, into a new algorithm called `KLUCB-RB` that leverages the information obtained on previously encountered bandits. We provide a theoretical analysis of `KLUCB-RB` and investigate the performance of the algorithm using extensive numerical experiments. These results highlight the empirical conditions required so that past information can be efficiently leveraged to speed up the learning process. The main contributions of this work are 1) the newly proposed routine bandit setting, 2) `KLUCB-RB` algorithm that solves this problem with asymptotically optimal regret minimization guarantees, and 3) an empirical illustration of the conditions for past information to be beneficial to the learning agent. This work is the subject of Saber et al. (2021b) and has been in collaboration with Léo Saci and Audrey Durand. It is the subject of Chapter 6.

Chapter 2

Unimodal Bandits

We consider a multi-armed bandit problem specified by a set of one-dimensional family exponential distributions endowed with a unimodal structure. We introduce `IMED-UB`, an algorithm that optimally exploits the unimodal-structure, by adapting to this setting Indexed Minimum Empirical Divergence (`IMED`) algorithm introduced by Honda and Takemura (2015). Owing to our proof technique, we are able to provide a concise finite-time analysis of `IMED-UB` algorithm. Numerical experiments show that `IMED-UB` competes with the state-of-the-art algorithms. This chapter reproduces the contents from Saber et al. (2021a) published at NeurIPS conference.

2.1 Introduction

We assume a *unimodal* structure similar to that considered in Yu and Mannor (2011) and Combes and Proutiere (2014a). That is, there exists an undirected graph $G=(\mathcal{A},E)$ whose vertices are arms $\mathcal{A}$, and whose edges E characterize a partial order among means $(\mu_a)_{a\in\mathcal{A}}$. This partial order is assumed unknown to the learner. We assume that there exists a unique optimal arm $a^\star=\operatorname{argmax}_{a\in\mathcal{A}}\mu_a$ and that for all sub-optimal arm $a\neq a^\star$, there exists a path $P_a=(a_1=a,\dots,a_{\ell_a}=a^\star)\in\mathcal{A}^{\ell_a}$ of length $\ell_a\geqslant 2$ such that for all $i\in[1,\ell_a-1]$, $(a_i,a_{i+1})\in E$ and $\mu_{a_i}<\mu_{a_{i+1}}$. We denote by $\mathcal{V}_a\coloneqq\{a'\in\mathcal{A}\setminus\{a\}:(a,a')\in E\}$ the neighbourhood of arm $a\in\mathcal{A}$ in graph G. Lastly, we assume that $\nu\subset\mathcal{P}\coloneqq\{p(\mu),\mu\in\mathrm{I}\}$, 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. Thus, for all $a\in\mathcal{A}$ the distribution ν_a is fully specified by its mean μ_a, and we have $\nu_a=p(\mu_a)$. We denote by $\mathcal{D}_{(\mathcal{P},G)}$ or $\mathcal{D}_{\text{uni}}$ (or simply $\mathcal{D}$ when there is no confusion) the structured set of such unimodal-bandit distributions characterized by $(\mathcal{P},G)$. In the following, we assume that $\mathcal{P}$ is a set of one-dimensional exponential family distributions.

Notations. Let $\nu\in\mathcal{D}$. Let $\mu^\star=\max_{a\in\mathcal{A}}\mu_a$ be the optimal mean and $a^\star=\operatorname{argmax}_{a\in\mathcal{A}}\mu_a$ be the optimal arm of ν. We define for an arm $a\in\mathcal{A}$ its sub-optimality gap $\Delta_a=\mu^\star-\mu_a$. Considering an horizon $T\geqslant 1$, thanks to the tower rule we can rewrite the regret as follows:

$$R(\nu,T)=\sum_{a\in\mathcal{A}}\Delta_a\,\mathbb{E}_\nu\big[N_a(T)\big]\,, \tag{2.1}$$

where $N_a(t)=\sum_{s=1}^t\mathbb{I}_{\{a_s=a\}}$ is the number of pulls of arm a at time t.

2.2 Regret lower bound

In this section, we recall for completeness the known lower bound on the regret when we assume a unimodal structure. In order to obtain non trivial lower bound we consider algorithms that are *consistent* (Definition 1). We can derive from the notion of consistency an asymptotic lower bound on the regret, see Combes and Proutiere (2014a).

Proposition 1 (Lower bounds on the regret)**.** *Let us consider a consistent algorithm. Then, for all configuration $\nu \in \mathcal{D}_{uni}$ with means $\mu \in \mathrm{I}^{\mathcal{A}}$, it must be that*

$$\liminf_{T\to\infty} \frac{R(\nu,T)}{\log(T)} \geqslant \mathfrak{C}_{uni}(\mu) := \sum_{a\in\mathcal{V}_{a^\star}} \frac{\Delta_a}{\mathrm{KL}(\mu_a|\mu^\star)} ,$$

where $\mathrm{KL}(\mu|\mu') = \int_{\mathbb{R}} \log(f(x,\mu)/f(x,\mu'))f(x,\mu)\lambda(\mathrm{d}x)$ denotes the Kullback-Leibler divergence between $\nu = p(\mu)$ and $\nu' = p(\mu')$, for $\mu, \mu' \in \mathrm{I}$, and where $\mathcal{V}_{a^\star} \subset \mathcal{A}\setminus\{a^\star\}$ is the neighbourhood of optimal arm $a^\star$.

Remark 3. *$\mathfrak{C}_{\mathcal{D}_{uni}}(\mu)$ is simply denoted $\mathfrak{C}_{uni}(\mu)$.*

Remark 4. *The quantity $\mathfrak{C}_{uni}(\mu)$ is a fully explicit function of μ (it does not require solving any optimization problem) for some set of distributions ν (see Remark 1). This useful property no longer holds in general for arbitrary structures. Also, it is noticeable that $\mathfrak{C}_{uni}(\mu)$ does not involve all the sub-optimal arms but only the ones in $\mathcal{V}_{a^\star}$. This indicates that sub-optimal arms outside $\mathcal{V}_{a^\star}$ are sampled $o(\log(T))$, which contrasts with the unstructured stochastic multi-armed bandits. See Combes and Proutiere (2014a) for further insights.*

2.3 Optimal algorithm for unimodal bandits

We present in this section a novel algorithm that matches the asymptotic lower bound of Proposition 1. This algorithm is inspired by the Indexed Minimum Empirical Divergence (`IMED`) proposed by Honda and Takemura (2011). The general idea behind this algorithm is, following the intuition given by the lower bound, to narrow on the current best arm and its neighbourhood for pulling an arm at a given time step.

Notations. The empirical mean of the rewards from the arm a at time t is denoted by $\widehat{\mu}_a(t) = \sum_{s=1}^{t} \mathbb{I}_{\{a_s=a\}} \frac{X_s}{N_a(t)}$ 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 $\widehat{\mathcal{A}}^\star(t) = \operatorname*{argmax}_{a\in\mathcal{A}} \widehat{\mu}_a(t)$ respectively the current best mean and the current set of optimal arms.

2.3.1 `OSUB` algorithm

For completeness, we recall in this subsection `OSUB` (Optimal Sampling for Unimodal Bandits) algorithm from Combes and Proutiere (2014a).

Algorithm 6 `OSUB`

Pull each arm once
for $t = |\mathcal{A}| \dots T-1$ **do**
Choose $\widehat{a}_t^\star \in \operatorname*{argmin}_{\widehat{a}^\star\in\widehat{\mathcal{A}}^\star(t)} N_{\widehat{a}^\star}(t)$ (chosen arbitrarily) so that $\widehat{a}_t^\star \in \widehat{\mathcal{A}}^\star(t)$ and $N_{\widehat{a}_t^\star}(t) \leqslant N_{\widehat{a}^\star}(t), \forall \widehat{a}^\star \in \widehat{\mathcal{A}}^\star(t)$

$$\text{Pull } a_{t+1} = \begin{cases} \widehat{a}_t^\star & \text{if } \frac{L_t(\widehat{a}_t^\star)-1}{d+1} \in \mathbb{N} \\ \operatorname*{argmax}_{a\in\mathcal{V}_{\widehat{a}_t^\star}} u_a(t) & \text{else} \end{cases}$$

end for

In Algorithm 6, for some numerical constant $c>0$, the index computed by OSUB algorithm for arm $a\in\mathcal{A}$ and step $t\geqslant|\mathcal{A}|$ is

$$u_a(t)=\sup\left\{u\geqslant\widehat{\mu}_a(t)\colon N_a(t)\mathrm{KL}(\widehat{\mu}_a(t)|u)\leqslant f_c(L_t(\widehat{a}_t^\star))\right\},$$

where $L_t(a)=\sum_{t'=1}^{t}\mathbb{I}_{\{\widehat{a}_{t'}^\star=a\}}$ counts how many times arm a was a leader (best empirical arm), $d=\max_{a\in\mathcal{A}}|\mathcal{V}_a|$ is the maximum degree of nodes in G, and $f_c(\cdot)=\log(\cdot)+c\log\log(\cdot)$. We set $c=1$ for the numerical experiments.

2.3.2 IMED-UB algorithm

We first pull each arm once. For all arm $a\in\mathcal{A}$ and time step $t\geqslant 1$ we introduce the IMED index

$$I_a(t)\,=\,N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t))+\log(N_a(t))\,,$$

with the convention $0\times\infty=0$. This index can be seen as a transportation cost for moving a sub-optimal arm to an optimal one plus an exploration term: the logarithm of the number of pulls. When an optimal arm is considered, the transportation cost is null and there is only the exploration part. Note that, as stated in Honda and Takemura (2011), $I_a(t)$ is an index in the weaker sense since it cannot be determined only by samples from the arm a but also uses the empirical mean of the current optimal arm. We define IMED-UB (Indexed Minimum Empirical Divergence for Unimodal Bandits), described in Algorithm 7, to be the algorithm consisting of pulling an arm $a_t\in\{\widehat{a}_t^\star\}\cup\mathcal{V}_{\widehat{a}_t^\star}$ with minimum index at each time step t, where is $\widehat{a}_t^\star\in\mathrm{argmin}_{\widehat{a}^\star\in\widehat{\mathcal{A}}^\star(t)}N_{\widehat{a}^\star}(t)$ is a current best arm in $\widehat{\mathcal{A}}^\star(t)$ such that for all current best arm $\widehat{a}^\star\in\widehat{\mathcal{A}}^\star(t)$, $N_{\widehat{a}^\star}(t)\geqslant N_{\widehat{a}_t^\star}(t)$. This is a natural algorithm since the lower bound on the regret given in Proposition 1 involves only the arms in $\mathcal{V}_{a^\star}$, the neighbourhood of the arm $a^\star$ of maximal mean.

Algorithm 7 IMED-UB

Pull each arm once
for $t=|\mathcal{A}|\dots T-1$ **do**
  Choose $\widehat{a}_t^\star\in\underset{\widehat{a}^\star\in\widehat{\mathcal{A}}^\star(t)}{\mathrm{argmin}}\,N_{\widehat{a}^\star}(t)$ (chosen arbitrarily) so that $\widehat{a}_t^\star\in\widehat{\mathcal{A}}^\star(t)$ and $N_{\widehat{a}_t^\star}(t)\leqslant N_{\widehat{a}^\star}(t),\forall\widehat{a}^\star\in\widehat{\mathcal{A}}^\star(t)$
  Pull $a_{t+1}\in\underset{a\in\{\widehat{a}_t^\star\}\cup\mathcal{V}_{\widehat{a}_t^\star}}{\mathrm{argmin}}\,I_a(t)$ (chosen arbitrarily)
end for

2.3.3 Asymptotic optimality of IMED-UB

In this section, we state the main theoretical result of this chapter.

Theorem 2 (Upper bounds). *Let us consider a set of distributions $\nu\in\mathcal{D}_{uni}$ with means $\mu\in\mathrm{I}^{\mathcal{A}}$ and let $a^\star$ its optimal arm. Let $\mathcal{V}_{a^\star}$ be the sub-optimal arms in the neighbourhood of $a^\star$. Then under IMED-UB algorithm for all $0<\varepsilon<\varepsilon_\nu$, for all horizon time $T\geqslant 1$, for all $a\in\mathcal{V}_{a^\star}$,*

$$\begin{aligned}\mathbb{E}_\nu[N_a(T)]\;\leqslant\;&\frac{1+\alpha_\nu(\varepsilon)}{\mathrm{KL}(\mu_a|\mu_{a^\star})}\log(T)+2\min\{|\mathcal{A}|\,,d(d+1)\}\;C_\varepsilon\sqrt{\log(c_\varepsilon T)}\\&+\;\min\{|\mathcal{A}|\,,d(d+1)\}\big(1+c_{\varepsilon_\nu}^{-1}\big)+\min\{3\,|\mathcal{A}|\,,(d+1)(d+3)\}\,\frac{2\sigma_{\varepsilon_\nu}^2\,e^{\varepsilon_\nu^2/2\sigma_\varepsilon^2}}{\varepsilon^2}+1\end{aligned}$$

and, for all $a\notin\{a^\star\}\cup\mathcal{V}_{a^\star}$,

$$\begin{aligned}\mathbb{E}_\nu[N_a(T)]\;\leqslant\;&2\min\{|\mathcal{A}|\,,d(d+1)\}\;C_\varepsilon\sqrt{\log(c_\varepsilon T)}\\&+\;\min\{|\mathcal{A}|\,,d(d+1)\}\big(1+c_{\varepsilon_\nu}^{-1}\big)+\min\{3\,|\mathcal{A}|\,,(d+1)(d+3)\}\,\frac{2\sigma_{\varepsilon_\nu}^2\,e^{\varepsilon_\nu^2/2\sigma_\varepsilon^2}}{\varepsilon^2}+1\,,\end{aligned}$$

where d is the maximum degree of nodes in G, $\varepsilon_\nu = \frac{1}{2}\min_{a\neq a^\star}\max_{a'\in\mathcal{V}_a}\mu_{a'}-\mu_a$,
$\sigma_\varepsilon^2 = \max_{a\in\mathcal{A}}\left\{\mathbb{V}_{X\sim p(\mu')}(X)\colon \mu'\in[\mu_a-\varepsilon\,,\mu_a]\right\}$ and $c_\varepsilon, C_\varepsilon > 0$ are the constants involved in Theorem 6. $\alpha_\nu(\cdot)$ is a non-negative function depending only on ν such that $\lim_{\varepsilon\to 0}\alpha_\nu(\varepsilon)=0$ (see Section 2.4.1 for more details).

In particular one can note that the arms in the neighbourhood of the optimal one are pulled $\mathcal{O}(\log(T))$ times while the other sub-optimal arms are pulled $\mathcal{O}\left(\sqrt{\log(T)}\right)$ of times under `IMED-UB`. This is coherent with the lower bound that only involves the neighbourhood of the best arm. More precisely, combining Theorem 2 and the tower rule (2.1) gives the asymptotic optimality of `IMED-UB` with respect to the lower bound of Proposition 1.

Corollary 1 (Asymptotic optimality)**.** *With the same notations as in Theorem 2, then under `IMED-UB` algorithm*

$$\limsup_{T\to\infty}\frac{R(\nu,T)}{\log(T)} \leqslant \mathfrak{C}_{uni}(\mu) = \sum_{a\in\mathcal{V}_{a^\star}}\frac{\Delta_a}{\mathrm{KL}(\mu_a|\mu_{a^\star})}\,.$$

A finite time analysis of `IMED-UB` is provided in following Section 2.4.

2.3.4 Asymptotic optimality of `IMED`

The non-structured multi-armed bandit problem can be seen as a particular case of unimodal bandits. Indeed, if we consider $E=\{(a,a')\colon a\neq a'\}$ then G is the fully connected graph and $\mathcal{V}_a=\mathcal{A}\setminus\{a\}$ for all arm $a\in\mathcal{A}$. In particular, the path from any sub-optimal arm a to $a^\star$, the optimal one, is simply $(a,a^\star)\in E$. Furthermore, in such a case, `IMED` and `IMED-UB` algorithms, respectively summarized in Algorithms 2 and 7, coincide perfectly. Thus, a straightforward application of Theorem 2 gives us optimal finite time guarantees for `IMED` algorithm. Note that this natural shift from the structured unimodal case to the unstructured case from both an algorithmic and analytical point of view makes the `IMED` approach an innovative approach to deal with structured bandit problems.

Corollary 2 (Upper bounds under `IMED` algorithm)**.** *Let us consider a set of distributions $\nu\in\mathcal{D}$ with means $\mu\in\mathrm{I}^{\mathcal{A}}$ and let $a^\star$ its optimal arm. Then under `IMED` algorithm for all $0<\varepsilon<\varepsilon_\nu$, for all horizon time $T\geqslant 1$, for all sub-optimal arm $a\neq a^\star$,*

$$\mathbb{E}_\nu[N_a(T)] \leqslant \frac{1+\alpha_\nu(\varepsilon)}{\mathrm{KL}(\mu_a|\mu_{a^\star})}\log(T) + 2\,|\mathcal{A}|\,C_\varepsilon\sqrt{\log(c_\varepsilon T)} + |\mathcal{A}|\left(1+c_{\varepsilon_\nu}^{-1}\right) + 3\,|\mathcal{A}|\,\frac{2\sigma_{\varepsilon_\nu}^2\, e^{\varepsilon_\nu^2/2\sigma_\varepsilon^2}}{\varepsilon^2} + 1$$

where $\varepsilon_\nu=\frac{1}{2}\min_{a\neq a^\star}\mu^\star-\mu_a$, $\sigma_\varepsilon^2=\max_{a\in\mathcal{A}}\left\{\mathbb{V}_{X\sim p(\mu')}(X)\colon \mu'\in[\mu_a-\varepsilon\,,\mu_a]\right\}$ and $c_\varepsilon, C_\varepsilon>0$ are the constants involved in Theorem 6. $\alpha_\nu(\cdot)$ is a non-negative function depending only on ν such that $\lim_{\varepsilon\to 0}\alpha_\nu(\varepsilon)=0$ (see Section 2.4.1 for more details). Thus under `IMED` algorithm, we have

$$\limsup_{T\to\infty}\frac{R(\nu,T)}{\log(T)} \leqslant \mathfrak{C}_1(\mu) = \sum_{a\neq a^\star}\frac{\Delta_a}{\mathrm{KL}(\mu_a|\mu_{a^\star})}\,.$$

2.4 `IMED-UB` 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 (see Lemma 1, Lemma 2). Then, based on

concentration lemmas (see Section B.2), the algorithm-based empirical lower bounds ensure the reliability of the estimators of interest (Lemma 4). Interestingly, this makes use of arguments based on recent concentration of measure that enable to control the concentration without adding some $\log\log$ bonus (such a bonus was required for example in the initial analysis of KLUCB algorithm from Cappé et al. (2013)). Then, combining the reliability of these estimators with the obtained algorithm-based empirical upper bounds, we obtain upper bounds on the average numbers of pulls (Theorem 2). The proof is concise to fit mostly in the next few pages.

2.4.1 Notations

Let us consider $\nu \in \mathcal{D}$ and let us denote by $a^\star$ its best arm. We recall that for all $a \in \mathcal{A}$, $\mathcal{V}_a = \{a' \in \mathcal{A} : \ (a,a') \in E\}$ is the neighbourhood of arm a in graph $G=(\mathcal{A},E)$, and that

$$d = \max_{a\in\mathcal{A}} |\mathcal{V}_a| \, , \ \varepsilon_\nu = \frac{1}{2} \min_{a\neq a^\star} \max_{a'\in\mathcal{V}_a} \mu_{a'} - \mu_a \, . \tag{2.2}$$

Then, there exists a function $\alpha_\nu(\cdot)$ such that for all $0<\varepsilon<\varepsilon_\nu$, for all $a \neq a^\star$,

$$\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon) \geqslant (1+\alpha_\nu(\varepsilon))^{-1} \, \mathrm{KL}(\mu_a|\mu^\star) \tag{2.3}$$

and $\lim\limits_{\varepsilon\downarrow 0} \downarrow \ \alpha_\nu(\varepsilon) = 0$. Indeed, since $\mathrm{KL}(\cdot|\cdot)$ is a convex function, it is a continuous function within the interior of its domain of definition. At each time step $t \geqslant 1$, $\widehat{a}_t^\star$ is arbitrarily chosen in $\operatorname{argmin}\limits_{a\in\widehat{\mathcal{A}}^\star(t)} N_a(t)$ where $\widehat{\mathcal{A}}^\star(t) = \operatorname{argmax}\limits_{a\in\mathcal{A}} \widehat{\mu}_a(t)$, that is

$$\begin{cases} \widehat{a}_t^\star \in \widehat{\mathcal{A}}^\star(t) \\ \forall \widehat{a}^\star \in \widehat{\mathcal{A}}^\star(t), \quad N_{\widehat{a}^\star}(t) \geqslant N_{\widehat{a}_t^\star}(t) \, . \end{cases}$$

2.4.2 Algorithm-based empirical bounds

IMED-UB 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 $\log(t)$.
We believe that establishing these empirical lower and upper bounds is a key element of our proof technique, that is of independent interest and not *a priori* restricted to the unimodal structure.

Lemma 1 (Empirical lower bounds)**.** *Under* IMED-UB*, at each step time $t \geqslant |\mathcal{A}|$, for all $a \in \mathcal{V}_{\widehat{a}_t^\star}$,*

$$\log\big(N_{a_{t+1}}(t)\big) \leqslant N_a(t) \, \mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) + \log(N_a(t)) \tag{2.4}$$

and

$$N_{a_{t+1}}(t) \leqslant N_{\widehat{a}_t^\star}(t) \, . \tag{2.5}$$

Proof. For $a \in \mathcal{A}$, 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

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

This implies, since the arm with minimum index is pulled, $\log\big(N_{a_{t+1}}(t)\big) \leqslant I_{a_{t+1}}(t) = \min\limits_{a'\in\{\widehat{a}_t^\star\}\cup\mathcal{V}_{\widehat{a}_t^\star}} I_{a'}(t) \leqslant I_{\widehat{a}_t^\star}(t) = \log\big(N_{\widehat{a}_t^\star}(t)\big)$. By taking the $\log^{-1}(\cdot)$, the last inequality allows us to conclude. □

Lemma 2 (Empirical upper bounds). *Under* `IMED-UB` *at each step time* $t \geqslant |\mathcal{A}|$,

$$N_{a_{t+1}}(t)\,\mathrm{KL}\big(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^{\star}(t)\big) \leqslant \log(t)\,. \tag{2.6}$$

Proof. As above, by construction we have

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

It remains, to conclude, to note that

$$N_{a_{t+1}}(t)\mathrm{KL}\big(\widehat{\mu}_{a_{t+1}}(t)|\widehat{\mu}^{\star}(t)\big) \leqslant I_{a_{t+1}}(t)\,,$$

and

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

□

2.4.3 Non-reliable current best arm

Before going further in the analysis, we inform the reader that sets $\mathcal{E}^{+}_{a,a'}(\varepsilon)$, $\mathcal{E}^{-}_{a,a'}(\varepsilon)$, $\mathcal{K}^{-}_{a,a'}(\varepsilon)$ for $a, a' \in \mathcal{A}$, $\varepsilon > 0$, used in this section are introduced and study in Section B.1.

For accuracy $\varepsilon > 0$, let $\mathcal{M}^{\star}(\varepsilon)$ be the set of times $t \geqslant 1$ that do not belong to $\mathcal{E}^{+}_{\widehat{a}^{\star}_t, a_{t+1}}(\varepsilon)$ and where some of the current best arm $\widehat{a}^{\star}_t$ differs from $a^{\star}$,

$$\mathcal{M}^{\star}(\varepsilon) \coloneqq \left\{ t \geqslant |\mathcal{A}| : \begin{array}{ll} (1) & t \notin \mathcal{E}^{+}_{\widehat{a}^{\star}_t, a_{t+1}}(\varepsilon) \\ (2) & \widehat{a}^{\star}_t \neq a^{\star} \end{array} \right\}. \tag{2.7}$$

Lemma 3 (Relation between subsets of times). *Under* `IMED-UB`*, for all accuracy* $0 < \varepsilon < \varepsilon_{\nu}$,

$$\mathcal{M}^{\star}(\varepsilon) \subset \bigcup_{\substack{t \geqslant 1 \\ a \in \mathcal{V}_{\widehat{a}^{\star}_t}}} \mathcal{K}^{-}_{a, a_{t+1}}(\varepsilon_{\nu})\,, \tag{2.8}$$

where $\varepsilon_{\nu} = \dfrac{1}{2} \min_{a \neq a^{\star}} \max_{a' \in \mathcal{V}_a} \mu_{a'} - \mu_a$.

Proof. Let us consider $t \in \mathcal{M}^{\star}(\varepsilon)$. Since $\widehat{a}^{\star}_t \neq a^{\star}$, there exists $a \in \operatorname{argmax}_{a' \in \mathcal{V}_{\widehat{a}^{\star}_t}} \mu_{a'}$ such that

$$\mu_a > \mu_{\widehat{a}^{\star}_t}\,. \tag{2.9}$$

Then, since $\widehat{a}^{\star}_t \in \operatorname{argmax}_{a \in \mathcal{A}} \widehat{\mu}_a(t)$, we have

$$\widehat{\mu}_{\widehat{a}^{\star}_t}(t) = \widehat{\mu}^{\star}(t) \geqslant \widehat{\mu}_a(t)\,. \tag{2.10}$$

Since $t \in \mathcal{M}^{\star}(\varepsilon)$, $t \notin \mathcal{E}^{+}_{\widehat{a}^{\star}_t, a_{t+1}}(\varepsilon)$, where $\mathcal{E}^{+}_{\widehat{a}^{\star}_t, a_{t+1}}(\varepsilon) = \big\{ t \in [\![1, T-1]\!] : N_{a_{t+1}}(t) \leqslant N_{\widehat{a}^{\star}_t}(t),\ \widehat{\mu}_{\widehat{a}^{\star}_t}(t) \geqslant \mu_{\widehat{a}^{\star}_t} + \varepsilon \big\}$ according to Equation (B.1). From empirical lower bounds (2.5), we have $N_{a_{t+1}}(t) \leqslant N_{\widehat{a}^{\star}_t}(t)$. This implies

$$\mu_{\widehat{a}^{\star}_t} + \varepsilon \geqslant \widehat{\mu}_{\widehat{a}^{\star}_t}(t)\,. \tag{2.11}$$

By combining Equations (2.10) and (2.11), it comes

$$\mu_{\widehat{a}^{\star}_t} + \varepsilon \geqslant \widehat{\mu}^{\star}(t) \geqslant \widehat{\mu}_a(t)\,. \tag{2.12}$$

Since $\varepsilon < \varepsilon_\nu \leqslant \left|\mu_a - \mu_{\widehat{a}_t^\star}\right|/2$, Equation (2.9) and previous Equation (2.12) imply

$$\mu_a - \varepsilon_\nu > \widehat{\mu}_{\widehat{a}_t^\star}(t) \geqslant \widehat{\mu}_a(t)\,. \tag{2.13}$$

Since $a \in \mathcal{V}_{\widehat{a}_t^\star}$, empirical lower bounds (2.4) imply

$$\log\big(N_{a_{t+1}}(t)\big) \leqslant N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) + \log(N_a(t))\,. \tag{2.14}$$

The classical monotonic properties of $\mathrm{KL}(\cdot|\cdot)$ and Equation (2.13) imply

$$\left\{\begin{array}{l} \widehat{\mu}_a(t) < \mu_a - \varepsilon_\nu \\ \mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) \leqslant \mathrm{KL}(\widehat{\mu}_a(t)|\mu_a - \varepsilon_\nu)\,. \end{array}\right. \tag{2.15}$$

Combining Equations (2.14) and (2.15), we get

$$\left\{\begin{array}{l} \widehat{\mu}_a(t) < \mu_a - \varepsilon_\nu \\ \log\big(N_{a_{t+1}}(t)\big) \leqslant N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\mu_a - \varepsilon_\nu) + \log(N_a(t))\,, \end{array}\right. \tag{2.16}$$

which means $t \in \mathcal{K}^-_{a,a_{t+1}}(\varepsilon_\nu)$. □

2.4.4 Reliable current means and current best arm

In this subsection, we characterize subsets of times where both the mean of current pulled arm and the optimal mean are well-estimated.

Let us consider for $0 < \varepsilon < \varepsilon_\nu$, for $a \neq a^\star$,

$$\mathcal{U}_a(\varepsilon) = \{t \geqslant |\mathcal{A}| : a_{t+1} = a\} \bigcap \left(\bigcup_{t \geqslant 1} \mathcal{E}^+_{a_{t+1},a_{t+1}}(\varepsilon) \cup \mathcal{E}^-_{\widehat{a}_t^\star,a_{t+1}}(\varepsilon) \cup \mathcal{E}^+_{\widehat{a}_t^\star,a_{t+1}}(\varepsilon) \cup \mathcal{M}^\star(\varepsilon) \right). \tag{2.17}$$

According to `IMED-UB` algorithm (summarized in Algorithm 7), for all $t \geqslant |\mathcal{A}|$, current pulled arm a_{t+1} is chosen in $\{\widehat{a}_t^\star\} \cup \mathcal{V}_{\widehat{a}_t^\star}$ so that $\widehat{a}_t^\star \in \{a_{t+1}\} \cup \mathcal{V}_{a_{t+1}}$. Furthermore, from Lemma 3 we have $\mathcal{M}^\star(\varepsilon) \subset \bigcup\limits_{\substack{t \geqslant 1 \\ a \in \mathcal{V}_{\widehat{a}_t^\star}}} \mathcal{K}^-_{a,a_{t+1}}(\varepsilon_\nu)$.

This implies

$$\mathcal{U}_a(\varepsilon) \subset \bigcup_{\substack{a' \in \{a\} \cup \mathcal{V}_a \\ a'' \in \mathcal{V}_{a'}}} \mathcal{E}^-_{a',a}(\varepsilon) \cup \mathcal{E}^+_{a',a}(\varepsilon) \cup \mathcal{K}^-_{a'',a}(\varepsilon_\nu) \subset \bigcup_{a''' \in \mathcal{A}} \mathcal{E}^-_{a''',a}(\varepsilon) \cup \mathcal{E}^+_{a''',a}(\varepsilon) \cup \mathcal{K}^-_{a''',a}(\varepsilon_\nu)\,, \tag{2.18}$$

where dummy variables a_{t+1} and $\widehat{a}_t^\star$ have been respectively replaced by a and a'.

In particular, from Lemma 30 and previous Equation (2.18) we have

$$\begin{aligned} \mathbb{E}_\nu[\mathcal{U}_a(\varepsilon)] &\leqslant 2\min\{|\mathcal{A}|\,,(d+1)\}\,\frac{2\sigma_\varepsilon^2 e^{\varepsilon^2/2\sigma_\varepsilon^2}}{\varepsilon^2} \\ &\quad + \min\{|\mathcal{A}|\,,d(d+1)\}\left(\frac{2\sigma_{\varepsilon_\nu}^2 e^{\varepsilon_\nu^2/2\sigma_{\varepsilon_\nu}^2}}{\varepsilon_\nu^2} + 1 + c_{\varepsilon_\nu}^{-1} + 2C_\varepsilon\sqrt{\log(c_\varepsilon T)}\right) \\ &\leqslant \min\{3\,|\mathcal{A}|\,,(d+1)(d+3)\}\,\frac{2\sigma_{\varepsilon_\nu}^2 e^{\varepsilon_\nu^2/2\sigma_\varepsilon^2}}{\varepsilon^2} \\ &\quad + \min\{|\mathcal{A}|\,,d(d+1)\}\Big(1 + c_{\varepsilon_\nu}^{-1} + 2C_\varepsilon\sqrt{\log(c_\varepsilon T)}\Big)\,, \end{aligned} \tag{2.19}$$

where $d = \max_{a \in \mathcal{A}} |\mathcal{V}_a|$ is the maximum degree of nodes in graph $\mathcal{G}$.

Lemma 4 (Reliable current means). *Under* `IMED-UB`*, for all accuracy* $0<\varepsilon<\varepsilon_\nu$*, for all sub-optimal arm* $a\neq a^\star$*, for all time step* $t\notin\mathcal{U}_a(\varepsilon)$*,* $t\geqslant|\mathcal{A}|$*, such that* $a_{t+1}=a$*,*

$$\begin{cases}\widehat{a}_t^\star = a^\star\\ \widehat{\mu}^\star(t)\geqslant\mu^\star-\varepsilon\\ \widehat{\mu}_a(t)\leqslant\mu_a+\varepsilon\,.\end{cases}$$

Proof. For $0<\varepsilon<\varepsilon_\nu=\min\limits_{a\neq a'}|\mu_a-\mu_{a'}|/2$, for $a\neq a^\star$, let us consider a time step $t\notin\mathcal{U}_a(\varepsilon)$, $t\geqslant|\mathcal{A}|$ such that $a_{t+1}=a$.

Since $a_{t+1}=a$ and $t\notin\mathcal{U}_{a_{t+1}}(\varepsilon)$ then $t\notin\mathcal{E}^+_{a_{t+1},a_{t+1}}(\varepsilon)$, that is $\widehat{\mu}_{a_{t+1}}(t)<\mu_{a_{t+1}}+\varepsilon$ or $\widehat{\mu}_a(t)<\mu_a+\varepsilon$ (since $a_{t+1}=a$).

Since $a_{t+1}=a$ and $t\notin\mathcal{U}_{a_{t+1}}(\varepsilon)$ then $t\notin\mathcal{E}^-_{\widehat{a}_t^\star,a_{t+1}}(\varepsilon)$, where

$$\mathcal{E}^-_{\widehat{a}_t^\star,a_{t+1}}(\varepsilon)=\left\{t\in[\![1,T-1]\!]\colon N_{a_{t+1}}(t)\leqslant N_{\widehat{a}_t^\star}(t),\ \widehat{\mu}_{\widehat{a}_t^\star}(t)\leqslant\mu_{\widehat{a}_t^\star}-\varepsilon\right\}$$

according to Equation (B.2). From empirical lower bounds (2.5), we have $N_{a_{t+1}}(t)\leqslant N_{\widehat{a}_t^\star}(t)$. This implies

$$\widehat{\mu}^\star(t)=\widehat{\mu}_{\widehat{a}_t^\star}(t)>\mu_{\widehat{a}_t^\star}-\varepsilon\,. \tag{2.20}$$

Since $a_{t+1}=a$ and $t\notin\mathcal{U}_{a_{t+1}}(\varepsilon)$ then $t\notin\mathcal{E}^+_{\widehat{a}_t^\star,a_{t+1}}(\varepsilon)\cup\mathcal{M}^\star(\varepsilon)$. From Equation (2.7), this implies

$$\widehat{a}_t^\star=a^\star\,. \tag{2.21}$$

By combining Equations (2.20) and (2.21), we get

$$\widehat{\mu}^\star(t)>\mu_{a^\star}-\varepsilon=\mu^\star-\varepsilon\,. \tag{2.22}$$

□

2.4.5 Upper bounds on the numbers of pulls of sub-optimal arms

In this subsection, we now combine the different results of the previous subsections to prove Theorem 2.

Proof of Theorem 2. For $0<\varepsilon<\varepsilon_\nu$, for $a\neq a^\star$, let us consider $t\notin\mathcal{U}_a(\varepsilon)$, $t\geqslant|\mathcal{A}|$, such that $a_{t+1}=a$. From empirical upper bounds (2.6), we have

$$N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t))\leqslant\log(t)\,. \tag{2.23}$$

From Lemma 4 and Algorithm 7, we have $a\in\mathcal{V}_{a^\star}$ and $\widehat{\mu}_a(t)\leqslant\mu_a+\varepsilon<\mu^\star-\varepsilon\leqslant\widehat{\mu}^\star(t)$. From classical monotonic properties of $\mathrm{KL}(\cdot|\cdot)$ and Equation (2.3), we have $\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t))\geqslant\mathrm{KL}(\mu_a+\varepsilon|\mu^\star-\varepsilon)\geqslant(1+\alpha_\nu(\varepsilon))^{-1}\,\mathrm{KL}(\mu_a|\mu^\star)$. In view of Equation (2.23), this implies

$$\forall t\notin\mathcal{U}_a(\varepsilon),t\geqslant|\mathcal{A}|\,,\ \text{such that } a_{t+1}=a,\quad\begin{cases}a\in\mathcal{V}_{a^\star}\\ N_a(t)\leqslant\dfrac{(1+\alpha_\nu(\varepsilon))\log(t)}{\mathrm{KL}(\mu_a|\mu^\star)}\,.\end{cases} \tag{2.24}$$

For all arm $a\in\mathcal{A}$, for all time step $t\geqslant|\mathcal{A}|$, we denote by

$$\tau_a(t)=\max\left\{t'\in[\![|\mathcal{A}|\,;t]\!]:\ a_{t'+1}=a\quad\text{and}\quad t'\notin\mathcal{U}_a(\varepsilon)\right\} \tag{2.25}$$

the last time step before time step t that does not belong to $\mathcal{U}_a(\varepsilon)$ such that we pull arm a.

Then, from Equations (2.24) and (2.25) we have

$$\begin{aligned}
\forall a \neq a^\star,\ \forall t \geqslant 1, \quad N_a(t) &= N_a(|\mathcal{A}|) + \sum_{t'\geqslant|\mathcal{A}|}^{t-1} \mathbb{I}_{\left\{a_{t'+1}=a\right\}} \\
&\leqslant 1 + \sum_{t'\geqslant 1}^{t-1} \mathbb{I}_{\left\{a_{t'+1}=a,\ t'\in\mathcal{U}_a(\varepsilon)\right\}} + \sum_{t'\geqslant|\mathcal{A}|}^{t-1} \mathbb{I}_{\left\{a_{t'+1}=a,\ t'\notin\mathcal{U}_a(\varepsilon)\right\}} \\
&\leqslant 1 + |\mathcal{U}_a(\varepsilon)| + \sum_{t'\geqslant|\mathcal{A}|}^{t-1} \mathbb{I}_{\left\{a_{t'+1}=a,\ t'\notin\mathcal{U}_a(\varepsilon)\right\}} \\
&\leqslant 1 + |\mathcal{U}_a(\varepsilon)| + \mathbb{I}_{\{a\notin\mathcal{V}_{a^\star}\}} \times 0 + \mathbb{I}_{\{a\in\mathcal{V}_{a^\star}\}} \times N_a(\tau_a(t)) \\
&\leqslant 1 + |\mathcal{U}_a(\varepsilon)| + \mathbb{I}_{\{a\in\mathcal{V}_{a^\star}\}} \frac{(1+\alpha_\nu(\varepsilon))\log(\tau_a(t))}{\mathrm{KL}(\mu_a|\mu^\star)} \\
&\leqslant 1 + |\mathcal{U}_a(\varepsilon)| + \mathbb{I}_{\{a\in\mathcal{V}_{a^\star}\}} \frac{(1+\alpha_\nu(\varepsilon))\log(t)}{\mathrm{KL}(\mu_a|\mu^\star)}\,.
\end{aligned}$$

This implies

$$\forall a \neq a^\star, \forall t \geqslant 1, \quad N_a(t) \leqslant \begin{cases} \dfrac{(1+\alpha_\nu(\varepsilon))\log(t)}{\mathrm{KL}(\mu_a|\mu^\star)} + |\mathcal{U}_a(\varepsilon)| + 1 & \text{if } a \in \mathcal{V}_{a^\star} \\ |\mathcal{U}_a(\varepsilon)| + 1 & \text{if } a \notin \mathcal{V}_{a^\star}\,. \end{cases} \tag{2.26}$$

From Equation (2.19), averaging these inequalities allows us to conclude. □

2.5 Numerical experiments

In this section, we compare empirically the following algorithms : `OSUB`, `UTS` (Combes and Proutiere, 2014a; Trinh et al., 2020) and `IMED-UB` described in Algorithm 7. We illustrate how performs `IMED-UB` algorithm under Bernoulli, Gaussian (variance $\sigma^2 = 0.25$) or Exponential distribution assumption. For the experiments we consider a graph $\mathcal{G}$ with maximal degree $d = 2$ and the unimodal unimodal vectors of means $\mu = (0.05, 0.10, 0.15, 0.20, 0.25, 0.20, 0.15, 0.10, 0.05)$, and average regrets over 500 runs for each distribution family. Based on these experiments (Figure 2.1), it seems that `IMED-UB` competes with `OSUB` and `UTS`.

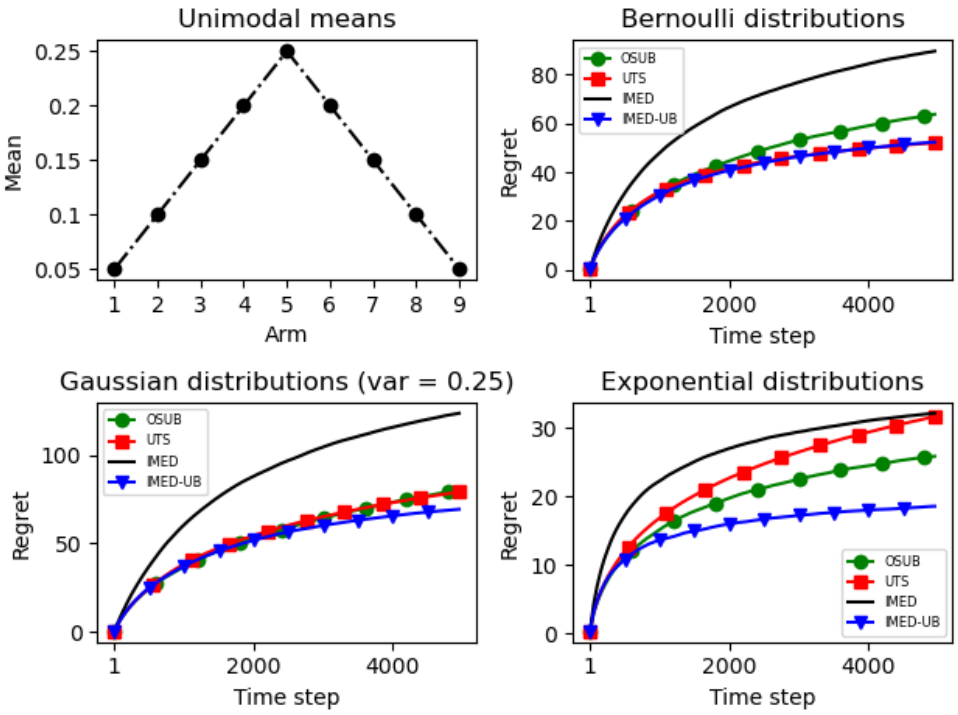


Figure 2.1: Cumulative regrets averaged over 500 runs.

Chapter 3

Multimodal Bandits

We consider a multi-armed bandit problem specified by a set of one-dimensional family exponential distributions endowed with a multimodal structure. The multimodal structure naturally extends the unimodal structure studied in Chapter 2. We introduce `IMED-MB`, an algorithm that exploits the multimodal structure, by adapting to this setting the popular Indexed Minimum Empirical Divergence (`IMED`) algorithm. Numerical experiments show that `IMED-MB` performs well in practice and competes with the state-of-the-art algorithms when assuming unimodal structure.

3.1 Introduction

We assume there exists an undirected graph $G = (\mathcal{A}, E)$ whose vertices are arms $\mathcal{A}$, and whose edges E characterize a partial order among means $(\mu_a)_{a\in\mathcal{A}}$. This partial order is assumed unknown 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 $\mathcal{A}^+$. We assume that $|\mathcal{A}^+(\nu)|$,the size of subset $\mathcal{A}^+(\nu)$, is equal to M, the number of local maximums. We assume that M is known to the learner (Assumption 2). Lastly, we assume that $\nu \subset \mathcal{P} \coloneqq \{p(\mu), \mu \in \mathrm{I}\}$, 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. Thus, for all $a \in \mathcal{A}$ we have $\nu_a = p(\mu_a)$. We denote by $\mathcal{D}_{(\mathcal{P},G)}$ or $\mathcal{D}_{\text{M-modal}}$ (or simply $\mathcal{D}$ when there is no confusion) the structured set of such unimodal-bandit distributions characterized by $(\mathcal{P}, G)$. In the following, we assume that $\mathcal{P}$ is a set of one-dimensional exponential family distributions. For $\nu \subset \mathcal{P}$, we denote by $\mathcal{A}^\star(\nu) = \operatorname{argmax}_{a\in\mathcal{A}} \mu_a$ the set of optimal arms of ν. When there is no possible confusion $\mathcal{A}^\star(\nu)$ is simply denoted $\mathcal{A}^\star$. In particular, we have

$$\mathcal{A}^\star \subset \mathcal{A}^+ \,. \tag{3.1}$$

Assumption 2 (Local maximums)**.** *The number $M = |\mathcal{A}^+|$ of arms with locally maximal means is known to the learner.*

Assumption 3 (Unique optimal arm)**.** *We assume there exists $a^\star \in \mathcal{A}$ such that $\mathcal{A}^\star = \{a^\star\}$.*

Unimodal Structure. When $\mathcal{A}^+ = \mathcal{A}^\star = \{a^\star\}$, the introduced *multimodal* coincides with the *unimodal* structure (Chapter 2) that has been first considered in Yu and Mannor (2011) from a bandit perspective. The study of unimodal structure naturally appears in many contexts, e.g. single-peak preference economics, voting theory or wireless communications, and Combes and Proutiere (2014a); Trinh et al. (2020) provide an explicit lower bound and optimal corresponding algorithms, respectively `OSUB`, `UTS`.

Notations. Let $\nu \in \mathcal{D}$. Let $\mu^\star = \max_{a\in\mathcal{A}} \mu_a$ be the optimal mean. We define for an arm $a \in \mathcal{A}$ its sub-optimality gap $\Delta_a = \mu^\star - \mu_a$. Considering an horizon $T \geqslant 1$, thanks to the tower rule we can rewrite the regret as follows:

$$R(\nu, T) = \sum_{a\in\mathcal{A}} \Delta_a \, \mathbb{E}_\nu\big[N_a(T)\big] \,, \tag{3.2}$$

where $N_a(t) = \sum_{s=1}^t \mathbb{I}_{\{a_s=a\}}$ is the number of pulls of arm a at time t.

3.2 Regret lower bound

In order to obtain non trivial lower bound on the regret we consider algorithms that are *consistent* (Definition 1). We can derive from the notion of consistency an asymptotic lower bound on the regret.

Proposition 2 (Lower bounds on the regret)**.** *Let us consider a consistent algorithm. Let us consider a configuration $\nu \in \mathcal{D}_{\text{M-modal}}$ with means $\mu \in \mathrm{I}^{\mathcal{A}}$ such that for all $a_1^+, a_2^+ \in \mathcal{A}^+(\nu)$,*

$$a_1^+ \neq a_2^+ \quad \Rightarrow \quad \mathcal{V}_{a_1^+} \cap \mathcal{V}_{a_2^+} = \emptyset \,. \tag{3.3}$$

Then it must be that

$$\liminf_{T\to\infty} \frac{R(\nu,T)}{\log(T)} \geqslant \mathfrak{C}_{\text{multi}}(\mu) := \sum_{a^+\in\mathcal{A}^+} \sum_{\substack{a\in\{a^+\}\cup\mathcal{V}_{a^+} \\ \mu_a\neq\mu^\star}} \frac{\Delta_a}{\mathrm{KL}(\mu_a|\mu^\star)} \,,$$

where $\mathrm{KL}(\mu|\mu') = \int_{\mathbb{R}} \log(f(x,\mu)/f(x,\mu'))f(x,\mu)\lambda(\mathrm{d}x)$ denotes the Kullback-Leibler divergence between $\nu = p(\mu)$ and $\nu' = p(\mu')$, for $\mu, \mu' \in \mathrm{I}$, and where $\mathcal{V}_a$ is the neighbourhood of arm $a \in \mathcal{A}$.

A proof of Proposition 2 is provided after the discussion below.

Remark 5. *$\mathfrak{C}_{\mathcal{D}_{\text{M-modal}}}(\mu)$ is simply denoted $\mathfrak{C}_{\text{multi}}(\mu)$.*

Remark 6. *The quantity $\mathfrak{C}_{\text{multi}}(\mu)$ is a fully explicit function of μ (it does not require solving any optimization problem) for some set of distributions ν (see Remark 1). This useful property no longer holds in general for arbitrary structures. Also, it is noticeable that $\mathfrak{C}_{\text{multi}}(\mu)$ 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 $\cup_{a^+\in\mathcal{A}^+}\{a^+\} \cup \mathcal{V}_{a^+}$ are sampled $o(\log(T))$, which contrasts with the unstructured stochastic multi-armed bandits.*

Discussion. In Proposition 2, the conditions on $\mathcal{A}^+(\nu)$ detailed in Equation 3.3 ensure that for all local maximum $a^+ \in \mathcal{A}^+$, for all sub-optimal $a \in \mathcal{V}_{a^+}$ in its neighbourhood, all "most confusing"[1] configuration $\nu^{(a)}$ (defined in Equation (3.4)) still remains in $\mathcal{D}_{\text{M-modal}}$. This allows us to use the consistency on $\mathcal{D}_{\text{M-modal}}$ of the considered algorithm. Without the conditions detailed in Equation (3.3), a most confusing configuration $\nu^{(a)}$ could have only $|\mathcal{A}^+(\nu)| - 1 = M - 1$ local maximums. However, these conditions are not very restrictive and only imply that two local maximums cannot follow each other in graph G. For clarity, let us illustrate in Figure 3.1 the situation in dimension one when the initial configuration has 2-modal means but may have a most confusing configuration with unimodal means.

[1] This notion of "most confusing" refers to the generic proof technique used to derive regret lower bounds. It involves a change-of-measure argument, from the initial configuration in which the arm is sub-optimal to another one chosen to make it optimal.

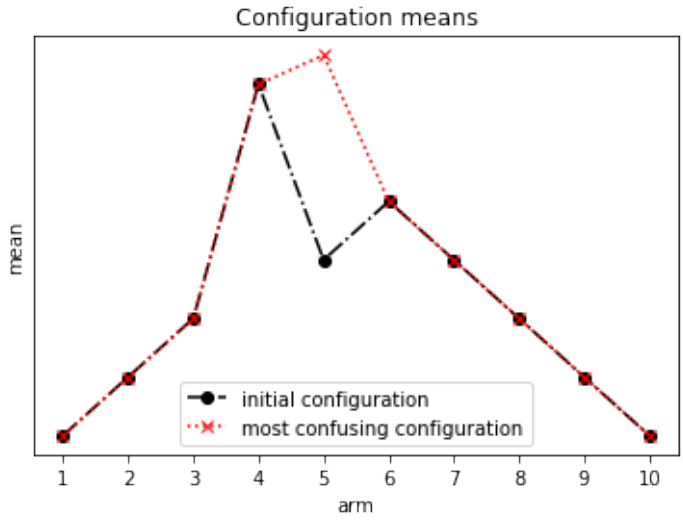


Figure 3.1: Configuration with 1-dimensional 2-modal means admitting unimodal most confusing configuration. Here, $\mathcal{A}=[\![1,10]\!]$ and $\mathcal{V}_a=\{a-1,a+1\}\cap\mathcal{A}$ for all $a\in\mathcal{A}$ so we have $\mathcal{A}^+=\{4,6\}$ and $\mathcal{V}_4\cap\mathcal{V}_6=\{5\}\neq\emptyset$.

Proof. We consider a configuration $\nu\in\mathcal{D}$ satisfying Equation (3.3), that is such that for all $a_1^+,a_2^+\in\mathcal{A}^+(\nu)$,

$$a_1^+\neq a_2^+ \quad\Rightarrow\quad \mathcal{V}_{a_1^+}\cap\mathcal{V}_{a_2^+}=\emptyset\,,$$

where $\mathcal{V}_a$ is the neighbourhood of arm $a\in\mathcal{A}$.

Let us consider a sub-optimal $a\in\cup_{a^+\in\mathcal{A}^+}\{a^+\}\cup\mathcal{V}_{a^+}\setminus\mathcal{A}^\star(\nu)$. The proof consists in used Lemma 5 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

$$\forall a'\in\mathcal{A},\quad \mu_{a'}^{(a)}(\varepsilon)=\begin{cases}\mu_{a'} & \text{if } a'\neq a\\ \mu^\star+\varepsilon & \text{if } a'=a\,.\end{cases} \tag{3.4}$$

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, due to Equation (3.3), the most confusing configuration $\nu^{(a)}(\varepsilon)$ still belongs to $\mathcal{D}_{\text{M-modal}}$, that is $\mu^{(a)}(\varepsilon)$ has exactly $|\mathcal{A}^+(\nu)|=M$ local maximums.

Let us consider the random variable $Z_T=N_a(T)/T\in[0,1]$. Then previous Lemma 5 implies

$$\sum_{a'\in\mathcal{A}}\mathbb{E}_\nu[N_{a'}(T)]\,\mathrm{KL}\big(\mu_{a'}\big|\mu_{a'}^{(a)}(\varepsilon)\big)\geqslant \mathrm{kl}\big(\mathbb{E}_\nu[Z_T]\big|\mathbb{E}_{\nu^{(a)}(\varepsilon)}[Z_T]\big)\,. \tag{3.5}$$

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 (3.5) rewrites

$$\mathbb{E}_\nu[N_a(T)]\,\mathrm{KL}\big(\mu_a\big|\mu^\star+\varepsilon\big)\geqslant \mathrm{kl}\big(\mathbb{E}_\nu[Z_T]\big|\mathbb{E}_{\nu^{(a)}(\varepsilon)}[Z_T]\big)\,. \tag{3.6}$$

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 $\mathcal{D}_{\text{M-modal}}$ and $\begin{cases} \nu \in \mathcal{D}_{\text{M-modal}} \\ a \notin \mathcal{A}^\star(\nu) \end{cases}$, the averaged number of pulls of arm a for configuration ν is sub-linear and

$$\lim_{T\to\infty} \mathbb{E}_\nu[Z_T] = \lim_{T\to 0} \mathbb{E}_\nu[N_a(T)]/T = 0\,. \tag{3.7}$$

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

$$\lim_{T\to\infty} \mathbb{E}_{\nu^{(a)}(\varepsilon)}[Z_T] = \lim_{T\to 0} \mathbb{E}_{\nu^{(a)}(\varepsilon)}[N_a(T)]/T = 1\,. \tag{3.8}$$

By combining Equation (3.7) and (3.8), we have in particular when T tends to ∞ that

$$\mathrm{kl}\big(\mathbb{E}_\nu[Z_T]\big|\mathbb{E}_{\nu^{(a)}(\varepsilon)}[Z_T]\big) \underset{T\to\infty}{\sim} \log\left(\frac{1}{1-\mathbb{E}_{\nu^{(a)}(\varepsilon)}[Z_T]}\right)\,. \tag{3.9}$$

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

$$\log\left(\frac{1}{1-\mathbb{E}_{\nu^{(a)}(\varepsilon)}[Z_T]}\right) = \log\left(\frac{T}{\sum\limits_{a'\notin\mathcal{A}^\star(\nu^{(a)}(\varepsilon))} \mathbb{E}_{\nu^{(a)}(\varepsilon))}[N_{a'}(T)]}\right) = \log\left(\frac{T}{O(T^\alpha)}\right)\,, \quad \forall \alpha > 0\,. \tag{3.10}$$

In particular, by combining previous Equation (3.10) and Equation (3.9) we get the following asymptotic result,

$$\lim_{T\to\infty} \frac{\mathrm{kl}\big(\mathbb{E}_\nu[Z_T]\big|\mathbb{E}_{\nu^{(a)}(\varepsilon)}[Z_T]\big)}{\log(T)} = 1\,. \tag{3.11}$$

We prove Proposition 2 by combining this last Equation (3.11) with Equation (3.6). □

Lemma 5 (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 $[0,1]$,*

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

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

3.3 Numerically efficient algorithm for multimodal bandits

We start this section by considering some useful notations before introducing and defining `IMED-MB` algorithm.

3.3.1 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 $\widehat{\mu}^\star(t) = \max\limits_{a\in\mathcal{A}} \widehat{\mu}_a(t)$ and $\widehat{\mathcal{A}}^\star(t) = \operatorname{argmax}\limits_{a\in\mathcal{A}} \widehat{\mu}_a(t)$ respectively the current best mean and the current set of optimal arms. We denote by $\widehat{a}_t^\star$ an arm arbitrarily chosen in $\widehat{\mathcal{A}}^\star(t)$. We denote by $\overline{\mathcal{A}}(t) := \{a \in \mathcal{A} : \forall a' \in \mathcal{V}_a, \widehat{\mu}_{a'}(t) \leqslant \widehat{\mu}_a(t)\}$ the set of arms with locally maximal empirical means and by

$\widehat{\mathcal{A}}^+(t) \subset \overline{\mathcal{A}}(t)$ the set of no more than $|\mathcal{A}^+|$ arms with locally maximal empirical means (ties are broken arbitrarily),

$$\begin{cases} \text{if } \left|\overline{\mathcal{A}}(t)\right| < |\mathcal{A}^+| : & \widehat{\mathcal{A}}^+(t) = \overline{\mathcal{A}}(t) \\ \text{if } \left|\overline{\mathcal{A}}(t)\right| \geqslant |\mathcal{A}^+| : & \widehat{\mathcal{A}}^+(t) \in \underset{a_1,\dots,a_{|\mathcal{A}^+|} \in \overline{\mathcal{A}}(t)}{\operatorname{argmax}} \sum_{a \in \{a_1,\dots,a_{|\mathcal{A}^+|}\}} \widehat{\mu}_a(t)\,. \end{cases}$$

Note that this choice of $\widehat{\mathcal{A}}^+(t)$ gives priority to local maximums with greater current empirical means.

3.3.2 `IMED-MB` algorithm.

For all arm $a \in \mathcal{A}$ and time step $t \geqslant 1$ we introduce the `IMED` index from Honda and Takemura (2015),

$$I_a(t) = N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) + \log(N_a(t))\,, \tag{3.12}$$

with the convention $0 \times \infty = 0$ and $\log(0) = -\infty$, and denote by $\overline{a}_t$ the arm with minimal `IMED` index

$$\overline{a}_t \in \underset{a \in \mathcal{A}}{\operatorname{argmin}}\, I_a(t) \quad \text{(arbitrarily chosen)}. \tag{3.13}$$

This index can be seen as a transportation cost for moving a sub-optimal arm to an optimal one plus an exploration term: the logarithm of the numbers of pulls. When an optimal arm is considered, the transportation cost is null and there is only the exploration part. Note that, as stated in Honda and Takemura (2015), $I_a(t)$ is an index in the weaker sense since it cannot be determined only by samples from the arm a but also uses empirical means of current optimal arms. If multimodal structure is not considered, arm $\overline{a}_t$ may be seen as the current most informative arm. However, regarding the lower bound for multimodal structure (Proposition 2), the current most informative arm may be

$$\overline{\overline{a}}_t \in \underset{a \in \bigcup_{\widehat{a}^+ \in \widehat{\mathcal{A}}^+(t)} \{\widehat{a}^+\} \cup \mathcal{V}_{\widehat{a}^+}}{\operatorname{argmin}} I_a(t) \quad \text{(arbitrarily chosen)}. \tag{3.14}$$

When $\left|\widehat{\mathcal{A}}^+(t)\right| < |\mathcal{A}^+|$, obviously the underlying multimodal structure is poorly estimated and arm $\overline{a}_t$ is used to deal with the trade-off exploitation versus exploration: if $\overline{a}_t$ coincides with the current best arm, that is $\overline{a}_t = \widehat{a}_t^\star$, we exploit and then pull best arm $\widehat{a}_t^\star$, otherwise we explore in order to better estimate the underlying multimodal structure and then pull the arm with the minimal number of pull in exploration phases

$$\underline{\underline{a}}_t \in \underset{a \in \mathcal{A}}{\operatorname{argmin}}\, N_a(t) \quad \text{(arbitrarily chosen)}. \tag{3.15}$$

When $\left|\widehat{\mathcal{A}}^+(t)\right| = |\mathcal{A}^+|$, arm $\overline{\overline{a}}_t$ is now used to deal with the trade-off exploitation versus exploration. We do not directly pull $\overline{\overline{a}}_t$ during exploration phases but consider second order `IMED` type indexes (Eq. 3.17) to ensure better estimations of the arms with locally maximal empirical means, that is the arms from set $\widehat{\mathcal{A}}^+(t)$. We explore in order to improve confidence in arms with locally maximal empirical means and then pull an arm with minimal second order index (Eq. 3.18). Let us consider

$$\widehat{a}_t^+ \in \underset{a \in \{\overline{\overline{a}}_t\} \cup \mathcal{V}_{\overline{\overline{a}}_t}}{\operatorname{argmax}}\, \widehat{\mu}_a(t) \quad \text{(arbitrarily chosen)}. \tag{3.16}$$

In particular, we have $\widehat{a}_t^+ \in \widehat{\mathcal{A}}^+(t)$. We then consider the second order index for arm $a \in \mathcal{A}$,

$$\underline{I}_a(t) = \begin{cases} \text{if } \widehat{\mu}_a(t) < \widehat{\mu}_{\widehat{a}_t^+}(t) : \\ \quad N_a(t)\, \mathrm{KL}\Big(\widehat{\mu}_a(t) \Big| \widehat{\mu}_{\widehat{a}_t^+}(t)\Big) + \log(N_a(t)) \\ \text{if } \widehat{\mu}_a(t) \geqslant \widehat{\mu}_{\widehat{a}_t^+}(t) : \\ \quad \log(N_a(t)) \end{cases} \tag{3.17}$$

and

$$\underline{a}_t \in \underset{a \in \{\widehat{a}_t^+\} \cup \mathcal{V}_{\widehat{a}_t^+}}{\operatorname{argmin}} \underline{I}_a(t) \quad \text{(arbitrarily chosen)}. \tag{3.18}$$

We speak of second order exploration because of the following inequalities on the indexes. At each time step $t \geqslant 1$, we have

$$\underline{I}_{\underline{a}_t}(t) \leqslant \log\Big(N_{\widehat{a}_t^+}(t)\Big) \leqslant I_{\widehat{a}_t^+}(t)\,. \tag{3.19}$$

Note that when $\widehat{a}_t^+ = \widehat{a}_t^\star$, then both `IMED` type indexes coincide, that is $\underline{I}_{\underline{a}_t}(t) = I_{\underline{a}_t}(t)$. However, when $\widehat{a}_t^+ \neq \widehat{a}_t^\star$, we expect arm $\widehat{a}_t^+$ to be a sub-optimal one and so to be pulled a logarithmic number of times. In that particular case and according to previous Equation (3.19), we then expect $\underline{I}_{\underline{a}_t}(t)$ to be a $O(\log\log t)$.

`IMED-MB` algorithm is summarized in Algorithm 8.

Algorithm 8 `IMED-MB`

for $t = 1 \dots T-1$ **do**
  if $\Big|\widehat{\mathcal{A}}^+(t)\Big| < |\mathcal{A}^+|$ **then** ▷▷▷ `MULTIMODAL STRUCTURE BADLY ESTIMATED`
    if $\overline{a}_t = \widehat{a}_t^\star$ **then** ▷ `Exploitation`
      Pull arm $a_{t+1} = \overline{a}_t$ (Eq. 3.13)
    else ▷ `Exploration`
      Pull arm $a_{t+1} = \underline{a}_t$ (Eq. 3.15)
    end if
  else ▷▷▷ `MULTIMODAL STRUCTURE WELL ESTIMATED`
    if $\underline{a}_t = \widehat{a}_t^+$ **then** ▷ `Reliable local maximum` $\widehat{a}_t^+$ `: exploration-exploitation`
      Pull arm $a_{t+1} = \overline{\overline{a}}_t$
    else ▷▷ `Non-reliable local maximum` $\widehat{a}_t^+$ `: second order exploration`
      Pull arm $a_{t+1} = \underline{a}_t$ (Eq. 3.18)
    end if
  end if
end for

3.4 Numerical experiments

In this section, we compare empirically the following algorithms : `OSUB`, `UTS` (Combes and Proutiere, 2014a; Trinh et al., 2020) and `IMED-MB` described in Algorithm 8. We illustrate how perform `IMED-MB` algorithm under Bernoulli, Gaussian (variance $\sigma^2 = 0.25$) or Exponential distribution assumption. For the experiments we consider a graph $\mathcal{G}$ with maximal degree $d = 8$ and the multimodal vectors of means is of dimension 2, and average regrets over 100 runs for each distribution family. Additional details are provided in the paragraph below. Based on these experiments (Figures 3.2, 3.3), it seems that `IMED-MB` competes with `OSUB` and `UTS` for the unimodal structure.

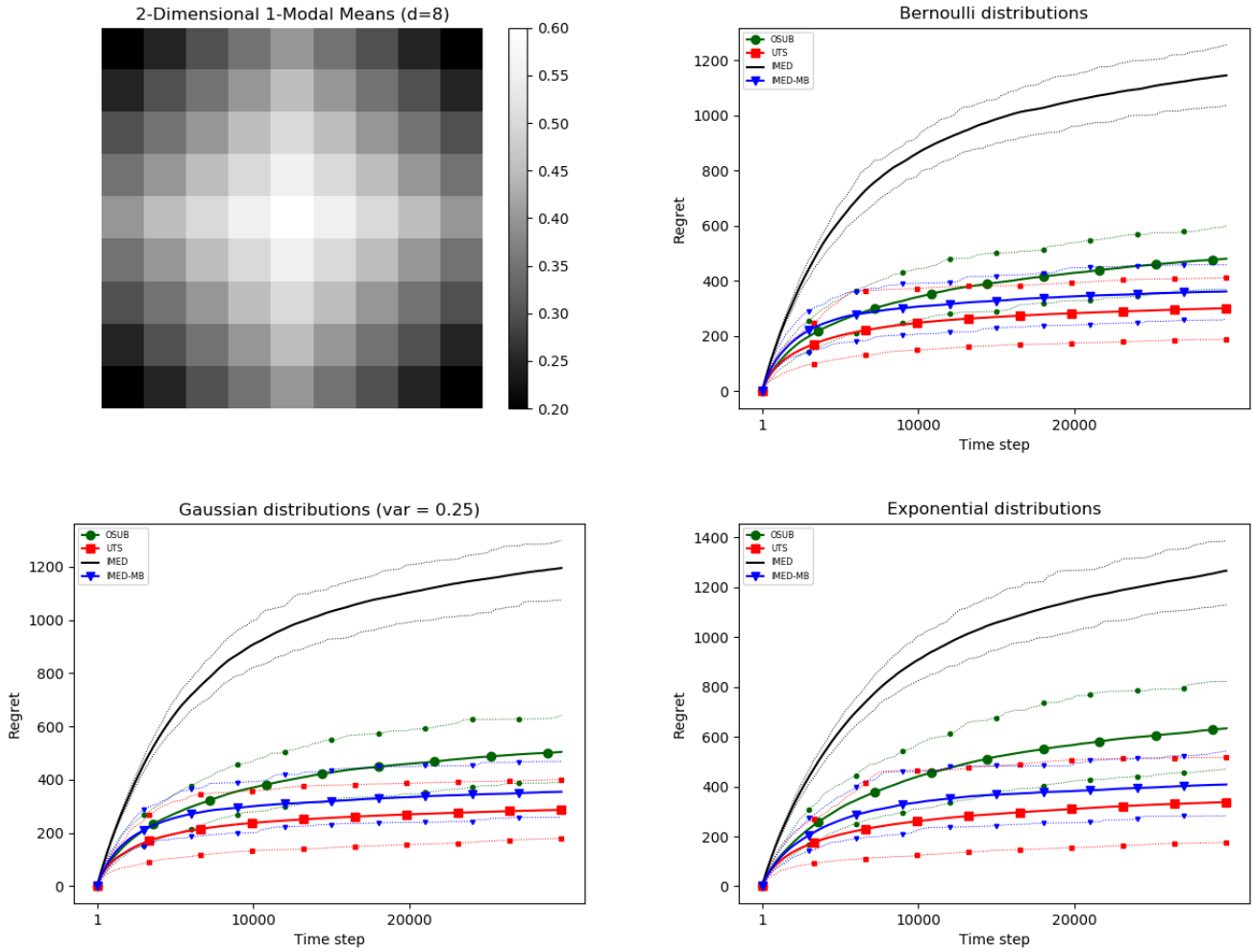


Figure 3.2: Cumulative regrets averaged over 100 runs for ν with 2-dimensional 1-modal means.

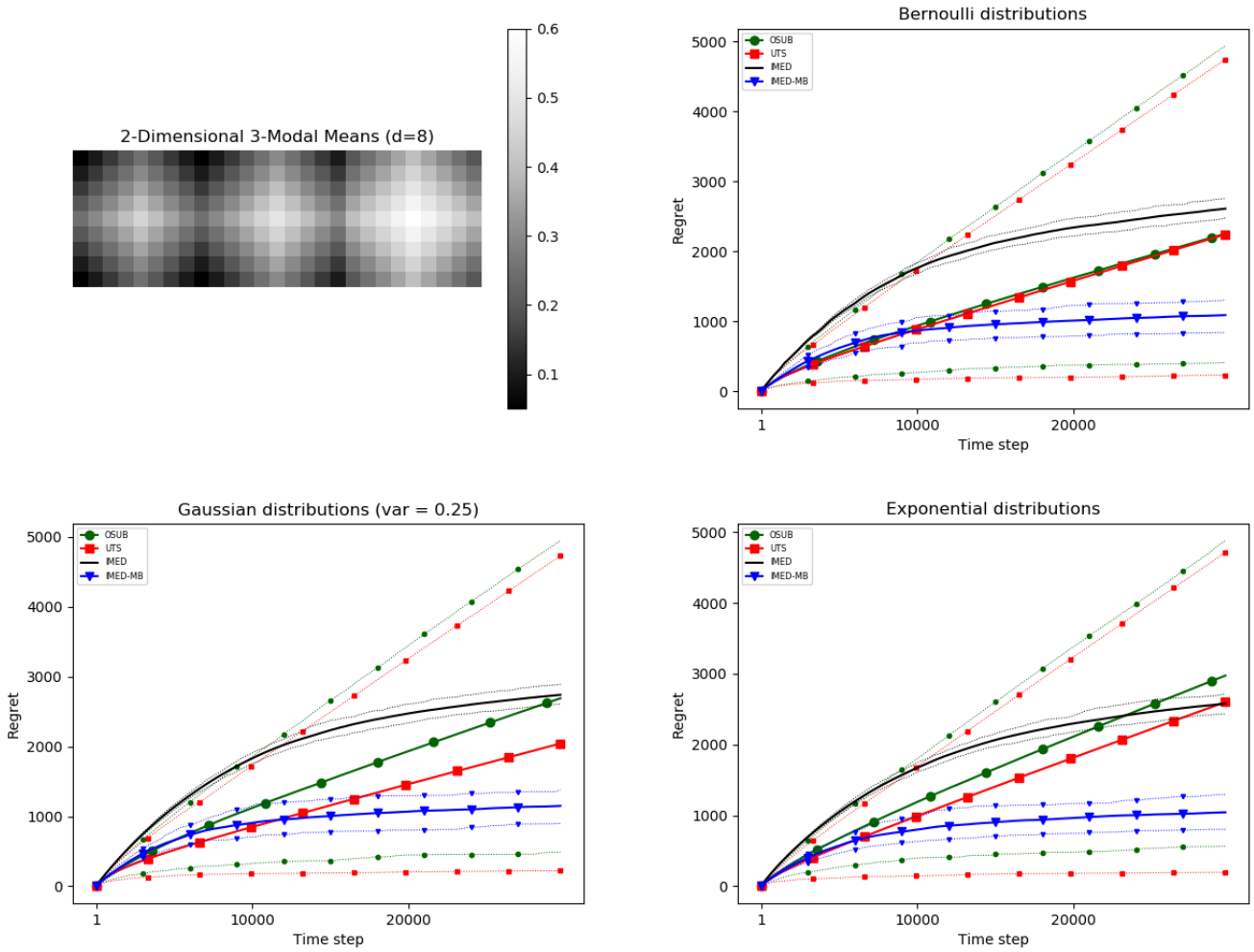


Figure 3.3: Cumulative regrets averaged over 100 runs for ν with 2-dimensional 3-modal means.

Additional details. In the following we provide additional details about experiments summarized in Figures 3.2 and 3.3.

The parameter of OSUB algorithm is set equal to $\gamma_{\text{OSUB}} = d$, where $d = \max_{a \in \mathcal{A}} |\mathcal{V}_a|$ is the maximal degree of nodes. The parameter of UTS algorithm is set equal to $\gamma_{\text{UTS}} = 2$. These parameters are those recommended in Combes and Proutiere (2014a) and Trinh et al. (2020).

For all the algorithms, we do not consider initial phase consisting in pulling each arm once. Indeed, we are interested in algorithm that do not necessarily explore all the arms. This is motivated by practical considerations: since the set of arms $\mathcal{A}$ can be large, we focus on algorithms that a priori do not explore all the arms. We show in Figures 3.4 and 3.5 that IMED-MB indeed concentrate the pulls around the arms with local maximal means. In both figures we represent the averaged numbers of pulls of sub-optimal arms under IMED-MB algorithm used to plot the averaged regrets in Figures 3.2 and Figures 3.3. Note that for readability purpose, we removed the pulls of optimal arms in both Figures.

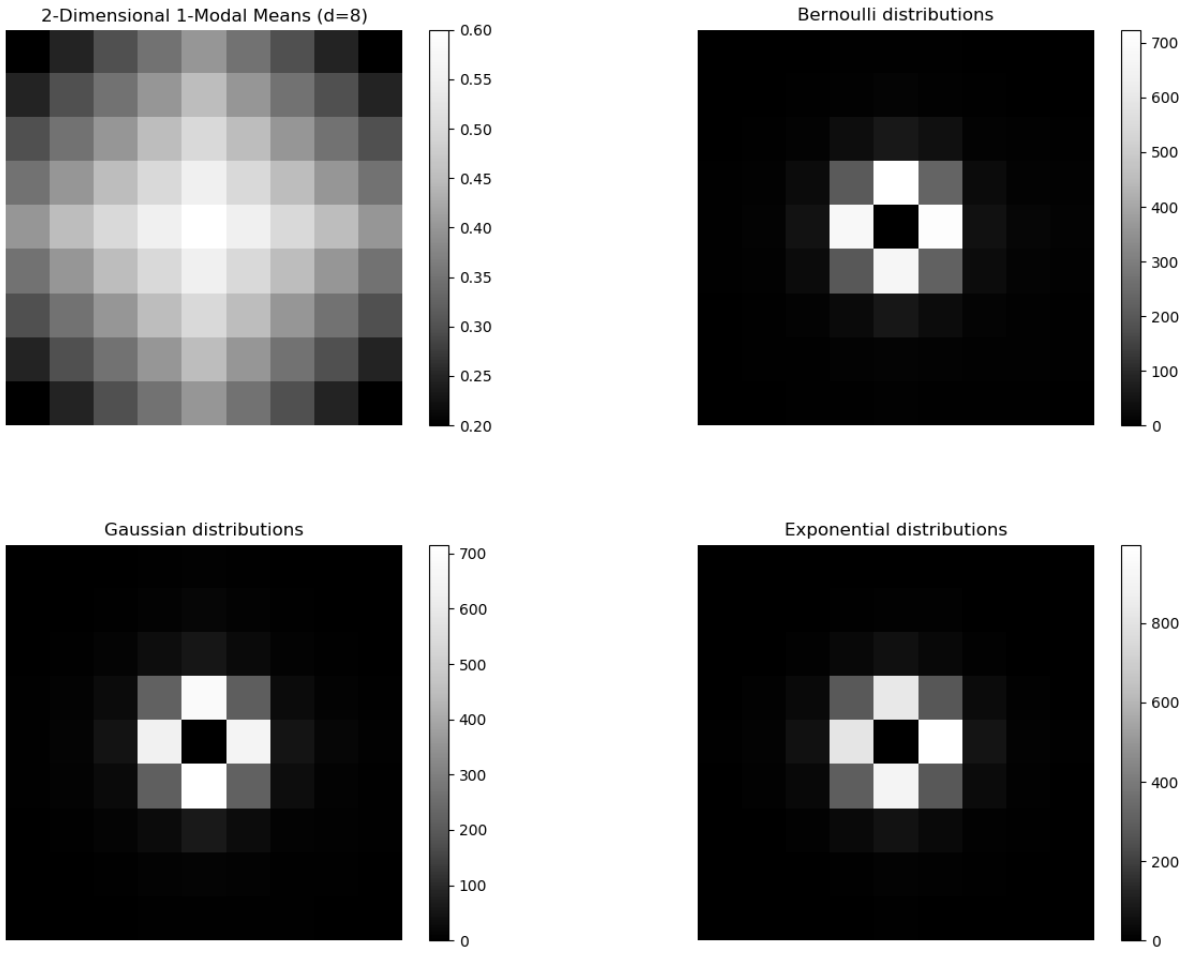


Figure 3.4: Number of pulls of sub-optimal arms under `IMED-MB` used in the unimodal case of Figure 3.2.

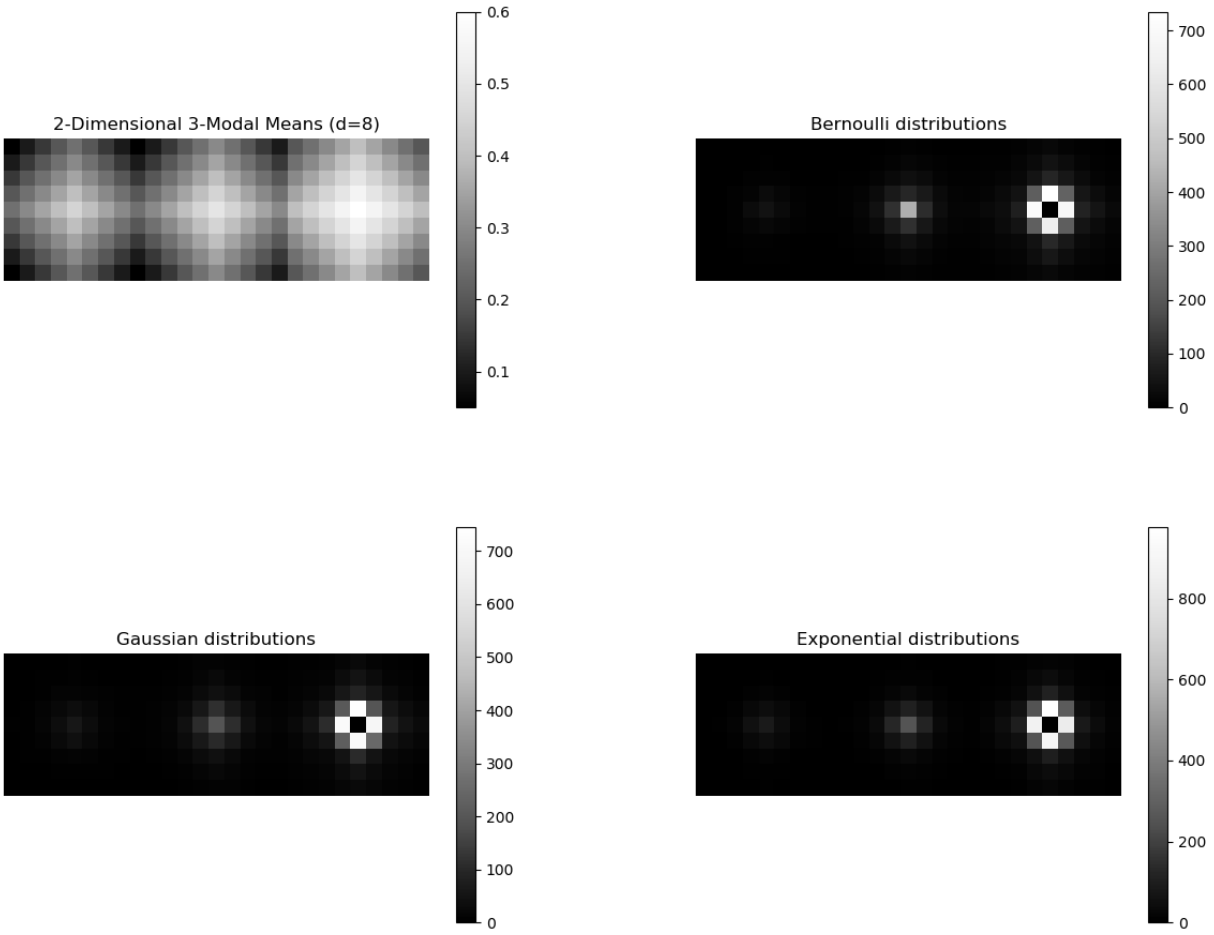


Figure 3.5: Numbers of pulls of sub-optimal arms under `IMED-MB` used in the 3-modal case of Figure 3.3.

Chapter 4

Graph-Structured Bandits

We consider a structured variant of the multi-armed bandit problem when the difference of means between any pair of arms is constrained not to exceed some value. This graph structure is introduced to encompass as special cases the classical structures Unimodal and Lipschitz. We derive the asymptotic lower bound on the cumulative regret for this structure, and introduce IMED-GS an extension of the popular Indexed Minimum Empirical Divergence (IMED) algorithm to such structured configurations. In order to show asymptotic optimality in a graph-structured scenario, we further add a tracking step to the IMED approach whose aim is to ensure the sub-optimal arms are played with correct asymptotic frequencies. Interestingly, this tracking step that requires solving an optimization problem is only triggered when the IMED index suggests exploration, which provably happens no more than $O(\log(T))$ times within T rounds. We further carefully handle the rounds when the structure cannot be exploited due to large uncertainty (e.g. initial rounds), by combining structured and unstructured versions of IMED. Our analysis enables an explicit finite-time regret bound emphasizing the role of the second-order terms. Last, we illustrate the benefit of IMED-GS over alternative structured bandit algorithms on numerical experiments.

4.1 Introduction

For a given *closed set of relationship matrices* $\Theta \subset [-1,1]^{\mathcal{A}^2}$ we introduce

$$\mathcal{D}_\Theta := \left\{ \nu \in \mathcal{B} \colon \exists \theta \in \Theta,\ \forall a, a' \in \mathcal{A},\ \mu_a - \mu_{a'} \leqslant \theta_{a,a'} \right\}, \tag{4.1}$$

where $\mathcal{B}$ is the set of Bernoulli distributions with means in $(0,1)$. We call this a *graph structure*, as it constraints pairs of means. We further impose a pseudo-metric property on each matrix $\theta \in \Theta$.

Remark 7. *$\mathcal{D}_\Theta$ may be denoted $\mathcal{D}$ when there is no possible confusion.*

Assumption 4 (Pseudo-metric)**.** *For all $\theta \in \Theta$, for all arms $a, a', a'' \in \mathcal{A}$, one has $\theta_{a,a} = 0$ (definiteness) and $\theta_{a,a''} \leqslant \theta_{a,a'} + \theta_{a',a''}$ (triangular inequality).*

This formulation is flexible enough to capture as special cases popular structures such as Unimodal, Lipschitz or Aggregate of bandits (see details in Section A.1), which makes it appealing to study. Note that it can also naturally interpolates between a fully structured and fully unstructured case.

Remark 8. *Assumption 4 does not require $\theta_{a,a'}$ to be non-negative for $a, a' \in \mathcal{A}$.*

Remark 9. *The notion of Lipschitz bandit we refer to is the one used in Magureanu et al. (2014). When $\Theta = \{\theta\}$ is a singleton and θ is a symmetric relationship matrix, that is $\theta_{a,a'} = \theta_{a',a}$ for $a, a' \in \mathcal{A}$, we then recover a generic notion of Lipschitz structure as a particular case of graph structure. However, the Unimodal structure, for example, cannot be described with Lipschitz properties but still remains a particular case of graph structure (see details in Section A.1).*

Goal. First, we target the design of a single algorithm able to achieve, for any given graph structure, instance-dependent asymptotic optimality against any instance. We also want this algorithm to be computationally efficient, in the sense it avoids computing (a version of) $\mathfrak{C}_{\mathcal{D}}(\mu)$ when estimations are "too noisy", and only solves this costly problem provably scarcely. Second, we want to obtain finite-time analysis with explicit terms using the same generic proof technique for each structure. Last, although our algorithm applies to other structures, we want it to be competitive against both structure-dependent (like e.g. `OSUB`, `CKL-UCB` respectively for Unimodal and Lipschitz) and generic (like `OSSB`) state-of-the-art algorithms in practice on classical structures. We now address these challenges.

4.2 Regret lower bound

In this section, we specify the regret lower bound when assuming a graph structure. To this end, we follow the classical approach from Lai and Robbins (1985); Graves and Lai (1997). In order to obtain non trivial lower bound we consider algorithms that are consistent on $\mathcal{D}_\Theta$ (Definition 1). Using this notion, asymptotic regret lower bounds are proved using change-of-measure argument by considering *most confusing* bandit configurations. We identify them first introducing, for each sub-optimal arm $a \notin \mathcal{A}^\star(\nu) = \operatorname*{argmax}_{a\in\mathcal{A}} \mu_a$, the following distribution-dependent subset of relationship matrices

$$\Theta_a(\nu) \coloneqq \left\{ \theta \in \Theta^\star(\nu) \colon \begin{array}{l} (1)\ \forall a^\star \in \mathcal{A}^\star(\nu),\ \theta_{a,a^\star} > 0 \\ (2)\ \forall a' \notin \mathcal{A}^\star(\nu),\ \theta_{a,a'} \geqslant 0 \end{array} \right\}, \tag{4.2}$$

where $\Theta^\star(\nu) = \left\{ \theta \in \Theta \colon \nu \in \mathcal{D}_{\{\theta\}} \right\}$, and then the corresponding set of *informative* sub-optimal arms

$$\mathcal{A}_a(\nu,\theta) \coloneqq \left\{ a' \in \mathcal{A} \colon \mu_{a'} \leqslant \mu^\star - \theta_{a,a'} \right\}, \quad \theta \in \Theta_a(\nu)\,. \tag{4.3}$$

When "moving" sub-optimal arm a to make it optimal in a most confusing configuration, $\mathcal{A}_a(\nu,\theta)$ represents the set of sub-optimal arms which must also be "moved" to ensure the "most confusing"[1] bandit for sub-optimal a belongs to the structure $\mathcal{D}_\theta$. Establishing lower bounds is also a key component towards building an efficient algorithm. To avoid technical issues, we make the mild assumption that there exists most confusing configurations in $\mathcal{D}_\Theta$ for each sub-optimal arm.

Assumption 5. *Each sub-optimal arm $a \notin \mathcal{A}^\star(\nu)$ admits a most confusing instance, that is $\Theta_a(\nu) \neq \emptyset$.*

Proposition 3 (Lower bound on the regret)**.** *Let us consider a consistent algorithm. Then, for all configuration $\nu \in \mathcal{D}_\Theta$ with means $\mu = (\mu_a)_{a\in\mathcal{A}}$, under Assumptions 4 and 5 it must be that*

$$\begin{aligned} \liminf_{T\to\infty} \frac{R(\nu,T)}{\log(T)} &\geqslant \mathfrak{C}_\Theta(\mu) \coloneqq \min_{\eta\in\mathbb{R}_+^{\mathcal{A}}} \sum_{a\in\mathcal{A}} \eta_a \left(\max(\mu) - \mu_a\right) \\ \text{s.t. } \forall a \notin \operatorname{argmax}(\mu),& \\ \min_{\theta\in\overline{\Theta}_a(\nu)} \sum_{\substack{a'\in\mathcal{A} \\ \mu_{a'} \leqslant \max(\mu) - \theta_{a,a'}}} & \operatorname{kl}\big(\mu_{a'} | \max(\mu) - \theta_{a,a'}\big) \eta_{a'} \geqslant 1, \end{aligned} \tag{4.4}$$

where $\overline{\Theta}_a(\nu) = \{\theta \in \Theta^\star(\nu) \colon \forall a' \in \mathcal{A}, \theta_{a,a'} \geqslant 0\}$ denotes the closure of $\Theta_a(\nu)$ in Θ, for all $a \notin \mathcal{A}^\star(\nu)$.

[1] These notions of "moving" and "most confusing" refer to the generic proof technique used to derive regret lower bounds. It involves a change-of-measure argument, from the initial configuration in which the arm is sub-optimal to another one chosen to make it optimal.

The proof of this result uses a change-of-measure argument and follows classical proof techniques from the literature, see Lai and Robbins (1985); Agrawal et al. (1989); Graves and Lai (1997); Cappé et al. (2013). We detail it in Section A.2 for completeness. Intuitively, the vector η represents an asymptotic number of pulls of arms rescaled by $\log(T)$.

Remark 10. *$\mathfrak{C}_{\mathcal{D}_\Theta}(\mu)$ is simply denoted $\mathfrak{C}_\Theta(\mu)$.*

Remark 11. *Proposition 3 is a specification to the graph structure of the known lower bounds on the regret for generic structures. Assumption 5 is satisfied by both Unimodal and Lipschitz structures (Section A.1). By straightforward calculations, we recover for these structures the regret lower bounds respectively established in Combes and Proutiere (2014a) and Magureanu et al. (2014). The regret lower bound formulation from Proposition 3 motivates the shape of the indexes for `IMED-GS` algorithm (Section 4.3.1).*

4.3 Optimal algorithm for graph-structured bandits

In this section, we introduce our main algorithm `IMED-GS` to handle graph-structured bandits. It is primarily based on an `IMED`-type approach (see Honda and Takemura (2015)). `IMED` has been introduced for the case of *unstructured* bandit configurations but is interesting for two reasons. First, it has been shown to be asymptotically optimal, in the sense of matching the asymptotic (unstructured) instance-dependent lower bounds like `KLUCB` or Thompson-sampling algorithms. Then, its index is directly derived from the analysis of the lower bound, that does not requires an optimization procedure. For this reason, it constitutes an interesting basis in order to build regret efficient algorithms for structured configurations. When extended to the structured case, the `IMED` approach naturally targets satisfying the constraints of the optimization problem (which is sometimes called Pareto-optimality), but not solving the optimization problem, which offers an interesting advantage for the practitioner. In order to go beyond Pareto-optimality, `IMED-GS` employs a weak form of (data dependent) forcing mechanisms called tracking, in which a solution to the constrained optimization problem is computed only when Pareto-optimality is already ensured. Also, in initial rounds when the structure is "hidden" by the large uncertainty about the distributions, `IMED-GS` effectively reduces to the unstructured `IMED`. These innovations yield provable optimality for any graph-structure at the price of a slight but controlled increase in exploration.
One idea behind `IMED-GS` algorithm is, following the intuition given by the lower bound, to define structured `IMED`-type indexes, which requires being able to estimate $\overline{\Theta}_a(\nu)$ for sub-optimal arms $a \notin \mathcal{A}^\star(\nu)$ and $\Theta^\star(\nu)$, where ν is unknown. The following assumptions provide a convenient setting to do so and capture Unimodal and Lipschitz structures. Assumption 6 implies in particular that either $\Theta^\star(\nu) = \Theta$, or $\Theta^\star(\nu)$ can be well-estimated when the best arm is well-identified.

Assumption 6 (Arm-supported structure)**.** *Θ is either a singleton or the structure is supported by the best arm, that is $\Theta = \cup_{a\in\mathcal{A}} \left\{\theta^{(a)}\right\}$ and for $a \in \mathcal{A}$, $\nu \in \mathcal{D}_{\{\theta^{(a)}\}} \Leftrightarrow a = a^\star$.*

In the following Section 4.3.1, we introduce and detail the `IMED-GS` algorithm. It is summarized in Algorithm 9 (all ties are broken arbitrarily). We discuss its optimality properties in Section 4.3.2.

Algorithm 9 `IMED-GS`

Input: Structure Θ, sequences $(\gamma_t)_{t\geqslant 1}$, positive constant Γ, ξ, integer d.
Pull each arm once
for $t = |\mathcal{A}| \dots T-1$ **do**
  Compute structured `IMED` choice $\overline{a}_t \in \operatorname{argmin}_{a\in\mathcal{A}} \overline{I}_a^{(d)}(t)$ (Eq. 4.9)
  if $\overline{a}_t \in \widehat{\mathcal{A}}^\star(t)$ **then** ▷ `Exploitation`
    Pull $a_{t+1} = \overline{a}_t$
  else ▷ `Exploration`
    Compute $n^{\text{opt}}(t)$ to form $N^{\text{opt}}(t)$, and compute $\overline{\mathcal{A}}_{\overline{a}_t}(t)$.
    Compute $a_t^{\text{opt}} \in \operatorname{argmax}_{a\in\overline{\mathcal{A}}_{\overline{a}_t}(t)} N_a^{\text{opt}}(t) - N_a(t)$ and $\underline{a}_t \in \operatorname{argmin}_{a\in\mathcal{A}} \underline{I}_a(t)$ (Eq. 4.13)
    if $\underline{a}_t == a_t^{\text{opt}}$ **then** ▷ `Reliable opt.`
      Pull $a_{t+1} = a_t^{\text{opt}}$ (Eq. 4.13)
    else ▷ `Unreliable opt.`
      if $\overline{I}_{\overline{a}_t}^{(d)}(t) \leqslant \Gamma \cdot I_{\dot{a}_t}(t)$ **then**
        Pull $a_{t+1} = \overline{a}_t$ (Eq. 4.9)
      else
        Pull $a_{t+1} = \dot{a}_t$ (Eq. 4.6)
end for

Unstructured `IMED`. For convenience, we introduce for $\xi \geqslant 0$, the function

$$\mathrm{f}_\xi : x \geqslant 1 \mapsto \log(x) + \xi \, \log(1 \vee \log(x)) \geqslant 0 \,. \tag{4.5}$$

We note that f_ξ is an increasing continuous function on $[1\,, +\infty[$. Using this notation, we redefine in this chapter the `IMED` indexes for unstructured bandits: for each $a \in \mathcal{A}$, for $t \geqslant 1$.

$$\begin{aligned} &(\textit{Unstructured } \texttt{IMED}) \\ &I_a(t) = N_a(t)\, \mathrm{kl}(\widehat{\mu}_a(t)\,|\widehat{\mu}^\star(t)) + \mathrm{f}_\xi(N_a(t)) \;. \end{aligned} \tag{4.6}$$

Remark 12. *The classical `IMED` indexes from Honda and Takemura (2015) are defined using $\xi = 0$. We introduce this minor extension to simplify the analysis of `IMED-GS` detailed in Section 4.5. In particular, later defining the structured index $\overline{I}_a(t)$ in (4.9), the following straightforward property occurs:*

$$\forall t \geqslant |\mathcal{A}|\,,\ \forall a \in \mathcal{A}, \quad \mathrm{f}_\xi(N_a(t)) \leq I_a(t) \leqslant \overline{I}_a(t) \,. \tag{4.7}$$

4.3.1 `IMED-GS` algorithm

At time t, we denote $\widehat{\mu}^\star(t) = \max_{a\in\mathcal{A}} \widehat{\mu}_a(t)$ the current (empirical) best mean, $\widehat{a}_t^\star \in \widehat{\mathcal{A}}^\star(t) = \operatorname{argmax}_{a\in\mathcal{A}} \widehat{\mu}_a(t)$ the current set of optimal arms and finally $\widehat{\theta}^\star(t) = \theta$ if $\Theta = \{\theta\}$, $\theta^{(\widehat{a}_t^\star)}$ otherwise.

Remark 13. *The definition of $\widehat{\theta}^\star(t)$, for $t \geqslant 1$, is inspired from Assumption 6, which is satisfied for the Unimodal, Lipschitz structures and the Aggregate of bandits (defined in Section A.1).*

We then define the current set of arms informative about arm $a \notin \widehat{\mathcal{A}}^\star(t)$ as

$$\widehat{\mathcal{A}}_a(t) := \left\{ a' \notin \widehat{\mathcal{A}}^\star(t) : \widehat{\mu}_{a'}(t) \leqslant \widehat{\mu}^\star(t) - \widehat{\theta}^\star_{a,a'}(t) \right\}. \tag{4.8}$$

In the initial rounds, not all informative arms are considered to compute the index. More precisely, (only) when it holds that $\left|\widehat{\mathcal{A}}_a(t)\right| > d+1$ and $\sum_{a' \in \widehat{\mathcal{A}}_a(t)} 2(\widehat{\mu}^\star(t) - \widehat{\mu}_{a'}(t))^2 N_{a'}(t) < \Phi(2|\widehat{\mathcal{A}}_a(t)|+1)$, for $a \notin \widehat{\mathcal{A}}^\star(t)$, where $d \in \{1, \dots, |\mathcal{A}|\}$ is a parameter and $\Phi(x) = x\log(x)$, we replace $\widehat{\mathcal{A}}_a(t)$ with the set $\widehat{\mathcal{A}}_a^{(d)}(t) \subset \widehat{\mathcal{A}}_a(t)$ consisting of $\{a\}$ plus the d-th most pulled arms from $\widehat{\mathcal{A}}_a(t) \setminus \{a\}$. It is justified by the fact that in the beginning, no structure can reasonably be exploited due to the poor estimates. We note that $a \in \widehat{\mathcal{A}}_a(t)$ since $\widehat{\theta}^\star_{a,a}(t) = 0$, for all current sub-optimal arm $a \notin \widehat{\mathcal{A}}^\star(t)$.

Graph-structured index. We are now ready to detail `IMED-GS` algorithm. Guided by the lower bound established in Proposition 3 we generalize in (4.9) the `IMED` index from Honda and Takemura (2011) to take into account the graph structure as follows. For convenience, we consider $\widehat{\mathcal{A}}_{\widehat{a}^\star}(t) = \widehat{\mathcal{A}}^{(d)}_{\widehat{a}^\star}(t) = \{\widehat{a}^\star\}$ for all current optimal arm $\widehat{a}^\star \in \widehat{\mathcal{A}}^\star(t)$. Then, for all arm $a \in \mathcal{A}$ and for all time step $t \geqslant |\mathcal{A}|$, we define

$$\begin{aligned} \overline{I}_a(t) \coloneqq \sum_{a' \in \widehat{\mathcal{A}}_a(t)} N_{a'}(t)\, \mathrm{kl}\big(\widehat{\mu}_{a'}(t) \big| \widehat{\mu}^\star(t) - \widehat{\theta}^\star_{a,a'}(t)\big) \\ + \log\left(\sum_{a' \in \widehat{\mathcal{A}}_a(t)} N_{a'}(t)\right), \end{aligned} \tag{4.9}$$

and its reduced version $\overline{I}^{(d)}_a(t)$ defined using $\widehat{\mathcal{A}}^{(d)}_a(t)$ in lieu of $\widehat{\mathcal{A}}_a(t)$ (this coincides with $\overline{I}_a(t)$ when the condition for using $\widehat{\mathcal{A}}^{(d)}_a(t)$ does not hold). Note that

$$\forall \widehat{a}^\star \in \widehat{\mathcal{A}}^\star(t), \quad \overline{I}_{\widehat{a}^\star}(t) = \overline{I}^{(d)}_{\widehat{a}^\star}(t) = \mathrm{f}_\xi(N_{\widehat{a}^\star}(t))\,.$$

This generalized index can be seen as a transportation cost for "moving" a sub-optimal arm to an optimal one, plus an exploration term (the term $\mathrm{f}_\xi(\cdot)$). When a current optimal arm is considered, the transportation cost is null and only the exploration part remains. Then, the most informative arm at time step $t \geqslant |\mathcal{A}|$ is defined as

$$\overline{a}_t \in \operatorname*{argmin}_{a \in \mathcal{A}} \overline{I}^{(d)}_a(t)\,. \tag{4.10}$$

Pulling this arm intuitively ensures the constrains in the optimization problem of the regret lower bound are asymptotically satisfied (Pareto optimality). In order to reach optimality, we need to ensure we pull arms in a way that asymptotically matches $\mathfrak{C}_\Theta(\mu)$. In order to avoid repeatedly solving such an optimization problem too often, we now detail an innovative approach that may look rather intricate at first sight, but is carefully designed to avoid unnecessary computations. It is based on an interplay between exploration/exploitation phases and reliability tests that we explain precisely below.

Exploitation. In case the most informative arm is currently optimal, that is $\overline{a}_t \in \widehat{\mathcal{A}}^\star(t)$, we exploit, that is the algorithm simply pulls this arm: $a_{t+1} = \overline{a}_t$. We show later that this happens asymptotically often, as the other case detailed below is only triggered about $O(\log(T))$ times out of T steps.

Exploration. In the other cases, we explore by trading off low regret and information gathering. To this end, we introduce the pseudo-counts $\left(N^{\mathrm{opt}}_a(t)\right)_{a \in \mathcal{A}} = \left(n^{\mathrm{opt}}_a(t) \overline{I}_{\overline{a}_t}(t)\right)_{a \in \mathcal{A}}$, where $n^{\mathrm{opt}}(t)$ is a solution of $\mathfrak{C}_\Theta(\widehat{\mu}(t))$,

the empirical version of (4.4), that is

$$
\begin{aligned}
n^{\mathrm{opt}}(t) \in \operatorname*{argmin}_{n\in\mathbb{R}_+^{\mathcal{A}}} & \sum_{a\notin\widehat{\mathcal{A}}^\star(t)} \widehat{\Delta}_a(t)\, n_a \\
s.t. \quad & \forall a \notin \widehat{\mathcal{A}}^\star(t), \\
& \sum_{a'\in\widehat{\mathcal{A}}_a(t)} \mathrm{kl}\Big(\widehat{\mu}_{a'}(t)\Big|\widehat{\mu}^\star(t)-\widehat{\theta}^\star_{a,a'}(t)\Big) n_{a'} \geqslant 1\,.
\end{aligned}
\tag{4.11}
$$

Note that the pseudo-counts rescale these proportions using $\overline{I}_{\overline{a}_t}(t)$ and not $\log(t)$. $\overline{I}_{\overline{a}_t}(t)$ behaves asymptotically as $\log(t)$ but is less conservative and more natural in finite-time. We then *track* the current optimal numbers of pulls by computing

$$
a_t^{\mathrm{opt}} \in \operatorname*{argmax}_{a'\in\widehat{\mathcal{A}}_{\overline{a}_t}(t)} N_{a'}^{\mathrm{opt}}(t) - N_{a'}(t)\,. \tag{4.12}
$$

Solving (4.11) is intuitively only useful asymptotically, provided that the set of informative arms are well-estimated. In order to have meaningful current subsets of informative arms when computing current optimization problem (4.11), for all time step $t \geqslant |\mathcal{A}|$, for all current sub-optimal arm $a' \notin \widehat{\mathcal{A}}^\star(t)$ we introduce the index

$$
\underline{I}_{a'}(t) := \begin{cases} \mathrm{f}_\xi\Big(N_{a_t^{\mathrm{opt}}}(t)\Big) & \text{if } a' = a_t^{\mathrm{opt}} \\ N_{a'}(t)\cdot 2\gamma_t^2(\widehat{\Delta}_{a'}(t))^2 + \mathrm{f}_\xi(N_{a'}(t)) & \text{if } a' \neq a_t^{\mathrm{opt}}, \end{cases} \tag{4.13}
$$

where $\widehat{\Delta}_{a'}(t) = \widehat{\mu}^\star(t) - \widehat{\mu}_{a'}(t)$ and $(\gamma_t)_{t\geqslant 1}$ is a decreasing sequence such that $0 < \gamma_t < \gamma_1$ for all $t \geqslant 1$, with $\gamma_1 := 1/(7 \vee \sqrt{\Phi(2|\mathcal{A}|+1)}$, where $\Phi(x) = x\log(x)$. We further compute

$$
\underline{a}_t \in \operatorname*{argmin}_{a'\notin\widehat{\mathcal{A}}^\star(t)} \underline{I}_{a'}(t)\,. \tag{4.14}
$$

This trades-off between estimating the graph structure and achieving low regret. However, rather than directly pulling this arm, we use it as a test. Indeed, pulling this arm only makes sense when the structure is already well estimated.

Reliability test of current optimization problem. If $\underline{a}_t = a_t^{\mathrm{opt}}$, we consider the set of informative arms is well estimated hence we pull the current *tracked* arm $a_{t+1} = a_t^{\mathrm{opt}}$. Otherwise, the current optimization problem is considered to be unreliable and thus either we explore the current most informative arm $\overline{a}_t$ or we explore disregarding the considered structure. To trade-off between this two options, we compare $\overline{I}^{(d)}_{\overline{a}_t}(t)$ to the minimum index of an unstructured bandit. Let $\Gamma > 1$. If $\overline{I}^{(d)}_{\overline{a}_t}(t) \leqslant \Gamma \cdot I_{\dot{a}_t}(t)$, we explore the current most informative arm, that is we choose $a_{t+1} = \overline{a}_t$. Otherwise, we explore according to `IMED` indexes for unstructured bandit, that is we choose

$$
a_{t+1} = \dot{a}_t \in \operatorname*{argmin}_{a\in\mathcal{A}} I_a(t)\,. \tag{4.15}
$$

Comment. When γ_t is set equal to zero, pulling arm a_t^{opt} then corresponds to pulling the least pulled current sub-optimal arm in exploration phases, which would a priori lead to a non-optimal asymptotic behavior. Now when γ_t is set close to 1, pulling arm a_t^{opt} would asymptotically yields a numbers of pulls of current sub-optimal arms larger than $O(\mathrm{f}_\xi(N_{a_t^{\mathrm{opt}}}(t))) = O(\mathrm{f}_\xi(\mathrm{f}_\xi(t)))$, which seems insufficient to properly estimate the current optimization problem. The sequence $(\gamma_t)_{t\geqslant 1}$ is introduced in order to manage all the situations between these two borderline cases. Now Γ trades-off between gathering information about the structure and

reliability perspectives. When the structure is not informative, that is when $\overline{I}_a(t) = I_a(t)$ for all $a \in \mathcal{A}$, the ratio $\overline{I}_{\bar{a}_t}(t)/I_{\dot{a}_t}(t)$ is all equal to 1. A ratio greater than 1 for some current sub-optimal arm means that we gather information about its interactions with other current sub-optimal arms (from a structure perspective) at the price of a poorer estimation of its means. Reliable current means ensures reliable current optimization problems. Lastly, the re-scaling of $n^{\mathrm{opt}}(t)$, solution of $\mathfrak{C}_\Theta(\widehat{\mu}(t))$, by the index $\overline{I}_{\bar{a}_t}(t)$ and not $\log(t)$ results in better performance in practice since $\overline{I}_{\bar{a}_t}(t) \leqslant \mathrm{f}_\xi\left(\max_{a\in\mathcal{A}} N_a(t)\right)$ is upper bounded by the counts $(N_a(t))_{a\in\mathcal{A}}$ rather than current time step $t \geqslant 1$.

4.3.2 Asymptotic optimality of `IMED-GS`

In this section, we state the main theoretical result of this chapter about the instance-dependent regret bound of our algorithm for graph-structured bandits. To state this result, we first introduce a few technical notations and consider some technical assumptions detailed in Section 4.5.1. We define for the bandit configuration ν, its minimal optimality gap as $\Delta_{\min} = \min_{a\notin\mathcal{A}^\star(\nu)} \Delta_a$. Let us also denote $\Sigma_{\Delta^{-1}} = \sum_{a\notin\mathcal{A}^\star(\nu)} \Delta_a^{-1}$. With regard to Assumption 6 and 8, we introduce for convenience the quantity $\varepsilon_\nu = \frac{\delta_{\min}}{4} \wedge (1-\mu^\star)$, where $\delta_{\min} = \Delta_{\min}$ if Θ is a singleton, and $\delta_{\min} = \min_{a\neq a'}\{\Delta_{\min},\ \theta^\star_{a,a'}-(\mu_a-\mu_{a'})\}$ otherwise. In particular, these problem-dependent quantities are away from 0. We further introduce $\forall a \in \mathcal{A}, \forall t \geqslant 1$, the quantity $\varepsilon_a(t) = \frac{3}{2}\cdot\frac{\gamma_t}{1-\gamma_t}\cdot\Delta_a$, and note that $\varepsilon_a(t) < \Delta_a/4$ provided that $\gamma_t < 1/7$. Last, we recall that $\gamma_1 = 1/(7 \vee \sqrt{\Phi(2|\mathcal{A}|+1)})$, where $\Phi(x) = x\log(x)$. Our main result is a precise finite-time bound on the regret of `IMED-GS`, in which we meticulously fill out the details, highlighting the first-order, second-order and constant terms.

Theorem 3 (Regret upper bound)**.** *Let us consider a configuration $\nu \in \mathcal{D}_\Theta$ with means $\mu = (\mu_a)_{a\in\mathcal{A}} \in (0,1)^{\mathcal{A}}$. Under Assumptions 4-5-7-6-8-9, for all positive decreasing sequence $(\gamma_t)_{t\geqslant1} < \gamma_1$, for all $\Gamma \geqslant 1$, for all $\xi > 1$, for all $d \geqslant 0$, for all accuracy $0 < \varepsilon < \varepsilon_\nu$, for all time horizon $T \geqslant |\mathcal{A}|$,*

$$R(\nu,T) \leqslant (A) + (B) + C_{\xi,d,\varepsilon}\,,\ \textit{with}$$

$$(A) = \inf_{\tau\in[1,T]}\left(\lambda_\tau(\mu,\varepsilon)\cdot\mathfrak{C}_\Theta\big(\mu(\varepsilon,\tau)\big)\cdot\mathrm{f}_\xi(T) + \sum_{a\in\mathcal{A}}\Delta_a\cdot\tau\right)$$

$$(B) = \frac{\sum_{a\notin\mathcal{A}^\star}\Delta_a^{-1}\Gamma\gamma_T^{-2}}{2(\mu^\star-\varepsilon)(1-\mu^\star-\varepsilon)}\left[\mathrm{f}_\xi\left(\frac{98\,\Sigma_{\Delta^{-1}}}{\delta_{\min}}\mathrm{f}_\xi(T)+1\right)+\mathrm{f}_\xi(C_{\xi,d,\varepsilon})+1\right],$$

where for the first term (A), we have introduced $\forall t \geqslant 1$, $\mu(\varepsilon,t) = (\mu_a(\varepsilon,t))_{a\in\mathcal{A}}$ with

$$\mu_a(\varepsilon,t) = \begin{cases} \mu_a + \varepsilon + \varepsilon_a(t) & ,\ \text{if } a \notin \mathcal{A}^\star \\ \mu_a - \varepsilon & ,\ \text{if } a \in \mathcal{A}^\star \end{cases}$$

$$\textit{and}\quad \lambda_t(\mu,\varepsilon) = \max_{a\notin\mathcal{A}^\star}\left(\frac{\Delta_a+2\varepsilon+\varepsilon_a(t)}{\Delta_{a'}-2\varepsilon-\varepsilon_a(t)}\right)^2.$$

Furthermore, $(\lambda_t(\mu,\varepsilon))_{t\geqslant1}$ and $\big(\mathfrak{C}_\Theta\big(\mu(\varepsilon,t)\big)\big)_{t\geqslant1}$ are non-increasing sequences such that for $t \geqslant 1$,

$$1 \leqslant \lambda_t(\mu,\varepsilon) \leqslant 49 \qquad \mathfrak{C}_\Theta(\mu) \leqslant \mathfrak{C}_\Theta\big(\mu(\varepsilon,t)\big) \leqslant 2\Sigma_{\Delta^{-1}}$$

and, provided that $\lim_{t\to\infty}\gamma_t = 0$,

$$\lim_{\varepsilon\to0}\lim_{t\to\infty}\lambda_t(\mu,\varepsilon) = 1 \qquad \lim_{\varepsilon\to0}\lim_{t\to\infty}\mathfrak{C}_\Theta\big(\mu(\varepsilon,t)\big) = \mathfrak{C}_\Theta(\mu)\,.$$

Finally, the term $C_{\xi,d,\varepsilon} = O(\varepsilon^{-2})$ does not depend on T and is made explicit in Section 4.6.4.

Asymptotically when $T \to \infty$, provided that $\lim_{t\to\infty} \gamma_t = 0$, the first term is the leading term, and scales with $\mathfrak{C}_\Theta(\mu)\log(T)$. Indeed, the infimum involved in the regret bound is bounded above by $\mathfrak{C}_\Theta(\mu)\log(T)+o(\log(T))$ when $\tau = \sqrt{\log(T)}$ for instance. Note that besides the fact the terms $\lambda_\tau(\mu,\varepsilon)$ and $\mathfrak{C}_\Theta(\mu(\varepsilon,\tau))$ before $\mathrm{f}_\xi(T)$ are asymptotically optimal, these terms are decreasing and provably upper-bounded. The second term scales with $O(\Gamma\gamma_T^{-2}\log\log(T))$, and the third one is a constant. Further, ε can be chosen such that $\varepsilon = o(\log(T)^{-1/2})$ in order to obtain an asymptotically optimal bound on the regret that only depends on horizon T, parameters ξ, d, Γ and $(\gamma_t)_{t\geqslant 1}$ of `IMED-GS` and the intrisinc characteristics of the considered structured bandit. In particular, we deduce the asymptotic optimality of `IMED-GS`, provided that $\lim_{t\to\infty} \gamma_t \sqrt{\frac{\log(t)}{\log\log(t)}} = \infty$.

Corollary 3 (Asymptotic optimality)**.** *Let us consider a set of Bernoulli distributions $\nu \in \mathcal{D}_\Theta$. Then under `IMED-GS` algorithm, for $(\gamma_t)_{t\geqslant 1}$ such that $\lim_{t\to\infty} \gamma_t = 0$ and $\lim_{t\to\infty} \gamma_t \sqrt{\frac{\log(t)}{\log\log(t)}} = \infty$, it holds*

$$\limsup_{T\to\infty} \frac{R(\nu,T)}{\log(T)} \leqslant \mathfrak{C}_\Theta(\mu)\,.$$

Discussion. Let us remark that Theorem 3 is a non-asymptotic result, were all terms are fully explicit (see Section 4.6 for the precise value of the remaining constant term). This is not very common in structured bandits and contrasts with regret bounds previously obtained for alternative algorithms in the literature. Further, we highlight that we only require the parameter ξ to exceed 1 thanks to a refinement of concentration inequalities that is of independent interest. We detail this result in Theorem 4 (Chapter 5) as we believe it can benefit other regret analysis. The main innovation is to take into account the fact that for small values of the number of observations, the regret can be handled with other tools than concentration. Hence, concentration only needs to be handled after some burn-in phase and not for all time steps. Now, the use of the reduced indexes with parameter d is interesting as it enables to decrease to value of the constant $C_{\xi,d,\varepsilon}$. Indeed, this enables to to make appear an exponential dependency in the parameter d instead of an exponential dependency of the number of arms $|\mathcal{A}|$. In contrast, in Magureanu et al. (2014) the authors required $\xi > 3\,|\mathcal{A}|$ (and so in Degenne et al. (2020b), as they employ same concentration results): there exists $t_0 \geqslant 1$ such that for all $t \geqslant t_0$,

$$\mathbb{P}_\nu\left(\sum_{a\in\mathcal{A}} N_a(t)\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a) \geqslant \mathrm{f}_{3|\mathcal{A}|+1}(t)\right) \leqslant \frac{1}{t\log(t)}\,.$$

A second point is that our concentration result (Theorem 4) does not directly come from stochastic orderings alone due to the presence of the random term $N_{\mathcal{A}'}(t) = \sum_{a\in\mathcal{A}'} N_a(t)$ for $\mathcal{A}' \subset \mathcal{A}$. This random term causes the analysis in Magureanu et al. (2014) to break, hence, we needed to derive a novel analysis to handle this difficulty. Lastly, some terms, e.g. that $C_{\xi,d,\varepsilon}$, are reminiscent of other analysis, such that appearing in the optimality bounds for `KLUCB` or `TS` (Cappé et al. (2013); Kaufmann et al. (2012)). Indeed these terms are due to concentration inequalities, hence it is expected that they appear here as well. We tighten their control with respect to the previous work. Indeed it is sometimes argued that the non-first order terms are large, making the bound unpractical even for large T. While this phenomenon cannot be completely avoided, we made specific efforts to control the terms tightly, which is of independent interest.

4.4 Numerical experiments

In this section, we illustrate the performance of `IMED-GS` algorithm when specialized to Unimodal, Lipschitz structures and Aggregates of bandits. These examples enable to compare the regret performance of this algorithm to existing state-of-the-art. We detail these structures in Section A.1.

In all considered bandit configuration, we naturally compare each time IMED-GS to IMED algorithm for unstructured bandits. We also compare IMED-GS to OSSB algorithm for generic structured bandits. For Unimodal structure we add specific comparison with OSUB, UTS and IMED-UB that are specialized to this structure, and for Lipschitz structure we add numerical comparison with CKL-UCB. We further report the IMED-GS run with setting $d = |\mathcal{A}| - 1$ (hence without the downsizing in the burn-in phase), in order to show that this parameter can be chosen large in practice without hindering much performance (the parameter $d = 3$ is suggested by theory). Here the parameter Γ is set to $|\mathcal{A}|^{1.5}$ and $\xi = 1$. In each experiment, we considered a time horizon of $T = 3000$, and results averaged over 300 independent experiments, reporting 10% and 90% quantiles of the regret on top of its average. The results are reported in Figure 4.1, and show the potential benefit of the algorithm. The distributions and the structures used for Figure 4.1 are provided below. Additional details and complementary experiments are provided in Section A.4.

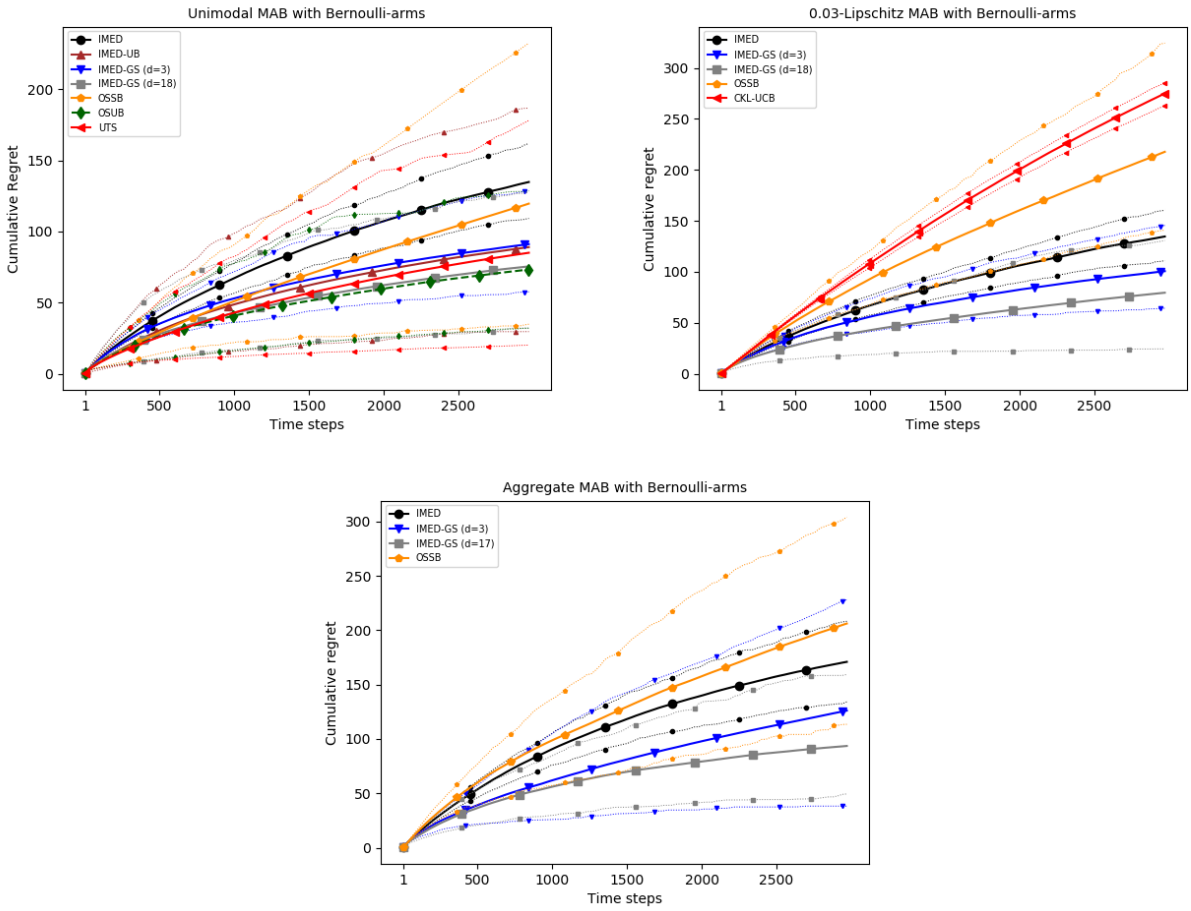


Figure 4.1: Comparison of IMED-GS to other algorithms on several structured bandit instances.

In Figure 4.2 below, we successively represent the vectors of means used for the experiments of Figure 4.1 to respectively illustrate the Unimodal and Lipschitz structures and the Aggregate of bandits. Note that for the Lipschitz structure, the vector of means is 0.03-Lipschitz as assumed in the corresponding experiment. For the Aggregate of bandits, the set of arms is decomposed as $\mathcal{A} = \mathcal{X} \times \mathcal{K}$ and the means $\mu = (\mu_{x,k})_{x \in \mathcal{X}, k \in \mathcal{K}}$ are represented in lexicographical order, with $\mathcal{X} = [\![1\,,6]\!]$ and $\mathcal{K} = [\![1\,,3]\!]$. This means we assume an aggregate of 6 bandits with 3 arms each. We set the relationship matrix equal to $\omega_{x,x'} = 0.07\,|x - x'|$ for $x, x' \in \mathcal{X}$ and each

arm is represented with a specific marker in Figure 4.2. For the Aggregate of Bandits, the first three arms are the arms of the first bandit, the arms $\{4, 5, 6\}$ are the arms of the second bandit, and so on.

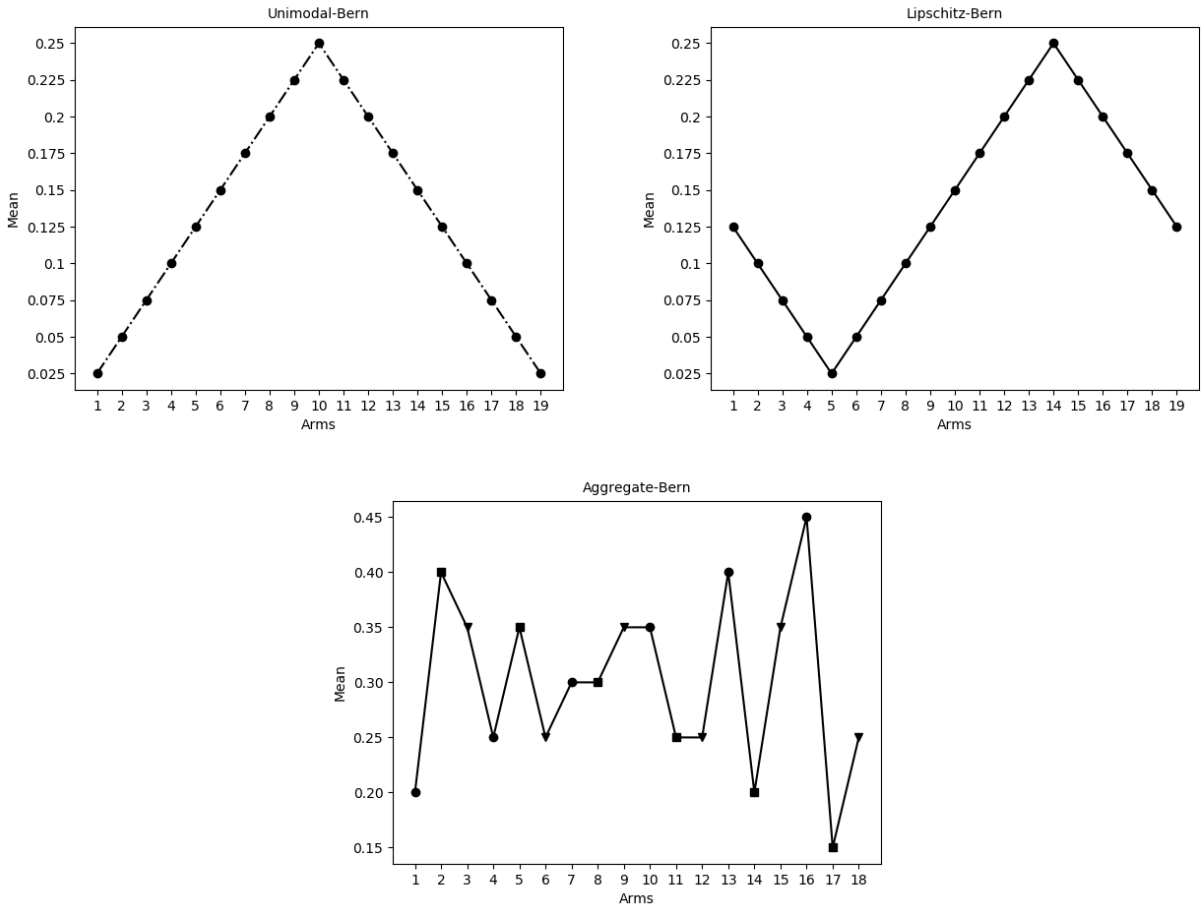


Figure 4.2: Means used for the experiments of Figure 4.1 to illustrate the Unimodal and Lipschitz structures and the Aggregate of bandits

4.5 Finite time properties of `IMED-GS` algorithm

In Section 4.5.1, we state the additional assumptions used in Theorem 3. We prove in Section 4.5.2 algorithm-based empirical bounds on the numbers of pulls. Then we introduce in Section 4.5.3 several notions of deviation of the empirical means and define the subsets of times where these deviations occur before providing upper bounds on the size of these subsets. Then in Section 4.5.4, we consider subset of times where undesirable events occur and establish, using empirical bounds from Section 4.5.2, relations between these subsets and the subsets of times where the empirical means deviate (subsets previously introduced in Section 4.5.3). Lastly, in Section 4.5.5 we show that we have nice reliability properties during most of time steps.

4.5.1 Additional assumptions

In this subsection, we state the technical assumptions allowing a simultaneous finite time analysis of the considered structures under the formalism of the graph structure. All theses assumptions are satisfied the structure

Unimodal, Lipschitz and Aggregate of Bandits.

First, we assume that the instance ν has a unique optimal arm, which is a folklore assumption.

Assumption 7 (Unique optimal arm)**.** *The instance ν has a unique optimal arm $a^\star \in \mathcal{A}$, that is $\mathcal{A}^\star = \{a^\star\}$.*

The next assumption is motivated by technical concerns and is only used when Θ is not a singleton. For Unimodal structure, Assumption 8 is trivially satisfied as it reads ($a'' \in [\![a', a]\!]$ or $a'' \in [\![a', a]\!]$) is equivalent to ($a'' \in [\![a, a']\!]$ or $a'' \in [\![a, a']\!]$) and that for all $a < a' \leqslant a^\star$ or $a^\star \leqslant a' < a$, $\mu_a \neq \mu_{a'}$. Please refer to Section A.1 for more details.

Assumption 8 (Symmetry, No adherence)**.** *If Θ is not a singleton, then for all $a, a', a'' \in \mathcal{A}$, $\theta^{(a)}_{a',a''} = \theta^{(a')}_{a,a''}$.*

Assumption 9 (No adherence)**.** *For all arms $a \neq a'$, $\mu_a - \mu_{a'} \neq \theta^\star_{a,a'}$.*

4.5.2 Algorithm-based empirical bounds

Lemma 6 (Empirical lower bounds)**.** *Under* `IMED-GS`*, at each step time $t \geqslant |\mathcal{A}|$, for all current sub-optimal arm $a \notin \widehat{\mathcal{A}}^\star(t)$,*

$$\mathrm{f}_\xi\big(N_{a_{t+1}}(t)\big) \leqslant \sum_{a' \in \widehat{\mathcal{A}}^{(d)}_a(t)} N_{a'}(t)\, \mathrm{kl}(\widehat{\mu}_{a'}(t)\Big|\widehat{\mu}^\star(t) - \widehat{\theta}^\star_{a,a'}(t)\Big) + \mathrm{f}_\xi\Big(N_{\widehat{\mathcal{A}}^{(d)}_a(t)}(t)\Big)\,, \tag{4.16}$$

and, for all current optimal arm $\widehat{a}^\star \in \widehat{\mathcal{A}}^\star(t)$,

$$N_{a_{t+1}}(t) \leqslant N_{\widehat{a}^\star}(t)\,. \tag{4.17}$$

Furthermore, if time step $t \geqslant |\mathcal{A}|$ corresponds to exploration, that is $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$, for all current sub-optimal arm $a' \in \mathcal{A}$,

$$\mathrm{f}_\xi\big(N_{\underline{a}_t}(t)\big) \leqslant N_{a'}(t)\, \widehat{\mathrm{k}}_{a'}(t) + \mathrm{f}_\xi(N_{a'}(t))\,. \tag{4.18}$$

Proof. First we note that

$$I_a(t) \leqslant \overline{I}^{(d)}_a(t)\,, \quad \forall a \in \mathcal{A}\,, \tag{4.19}$$

implies

$$\min_{a \in \mathcal{A}} I_a(t) \leqslant \min_{a \in \mathcal{A}} \overline{I}^{(d)}_a(t)\,, \tag{4.20}$$

that is,

$$I_{\dot{a}_t}(t) \leqslant \overline{I}^{(d)}_{\overline{a}_t}(t) \tag{4.21}$$

From Equations (4.9), (4.6), (4.10) and (4.15) defining indexes $\Big(\overline{I}^{(d)}_a(t)\Big)_{a \in \mathcal{A}}$, $(I_a(t))_{a \in \mathcal{A}}$, $\overline{a}_t$ and $\dot{a}_t$, we have

$$\begin{aligned} \forall a \notin \widehat{\mathcal{A}}^\star(t), \forall a' \in \overline{\mathcal{A}}_a(t), \quad & \mathrm{f}_\xi(N_a(t)) \leqslant I_a(t) \leqslant \overline{I}^{(d)}_a(t) \\ & \mathrm{f}_\xi(N_{a'}(t)) \leqslant \overline{I}^{(d)}_a(t) \end{aligned} \tag{4.22}$$

$$\forall \widehat{a}^\star \in \widehat{\mathcal{A}}^\star(t), \quad \mathrm{f}_\xi(N_{\widehat{a}^\star}(t)) = I_{\widehat{a}^\star}(t) = \overline{I}^{(d)}_{\widehat{a}^\star}(t)\,,$$

and

$$\forall a \notin \widehat{\mathcal{A}}^\star(t), \quad \overline{I}^{(d)}_{\overline{a}_t}(t) \leqslant \overline{I}^{(d)}_a(t) \leqslant \sum_{a' \in \widehat{\mathcal{A}}^{(d)}_a(t)} N_{a'}(t)\, \mathrm{kl}(\widehat{\mu}_{a'}(t)\Big|\widehat{\mu}^\star(t) - \widehat{\theta}^\star_{a,a'}(t)\Big) + \mathrm{f}_\xi\Big(N_{\widehat{\mathcal{A}}^{(d)}_a(t)}(t)\Big)\,, \tag{4.23}$$

$$\forall \widehat{a}^\star \in \widehat{\mathcal{A}}^\star(t), \quad \overline{I}^{(d)}_{\overline{a}_t}(t) \leqslant \overline{I}^{(d)}_{\widehat{a}^\star}(t) = \mathrm{f}_\xi(N_{\widehat{a}^\star}(t))\,.$$

Combining Equations (4.21), (4.22) and (4.23) yields

$$\begin{aligned}
&\forall a'' \in \overline{\mathcal{A}}_{\overline{a}_t}(t) \cup \{\overline{a}_t\}\,, \forall a \notin \widehat{\mathcal{A}}^\star(t),\\
&\mathrm{f}_\xi(N_{\dot{a}_t}(t))\,,\ \mathrm{f}_\xi(N_{\overline{a}_t}(t))\,,\ \mathrm{f}_\xi(N_{a''}(t)) \leqslant \sum_{a' \in \widehat{\mathcal{A}}_a^{(d)}(t)} N_{a'}(t)\,\mathrm{kl}(\widehat{\mu}_{a'}(t)\Big|\widehat{\mu}^\star(t) - \widehat{\theta}^\star_{a,a'}(t)\Big) + \mathrm{f}_\xi\Big(N_{\widehat{\mathcal{A}}_a^{(d)}(t)}(t)\Big)\,,\\
&\forall a'' \in \overline{\mathcal{A}}_{\overline{a}_t}(t) \cup \{\overline{a}_t\}\,, \forall \widehat{a}^\star \in \widehat{\mathcal{A}}^\star(t),\\
&N_{\dot{a}_t}(t),\ N_{\overline{a}_t}(t),\ N_{a''}(t) \leqslant N_{\widehat{a}^\star}(t)\,,
\end{aligned} \tag{4.24}$$

where, by convention, $\overline{\mathcal{A}}_{\widehat{a}^\star}(t) = \emptyset$ for all $\widehat{a}^\star \in \widehat{\mathcal{A}}^\star(t)$. We note that $\{\overline{a}_t\} \subset \widehat{\mathcal{A}}_{\overline{a}_t}(t)$ when $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$. In particular, by considering Equation (4.12) that defines a_t^{opt}, when $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$ we have

$$a_t^{\mathrm{opt}} \in \overline{\mathcal{A}}_{\overline{a}_t}(t)$$

and Equation (4.24) then implies

$$\begin{aligned}
&\forall a \notin \widehat{\mathcal{A}}^\star(t),\\
&\qquad \mathrm{f}_\xi(N_{\dot{a}_t}(t))\,,\ \mathrm{f}_\xi(N_{\overline{a}_t}(t))\,,\ \mathbb{I}_{\left\{\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)\right\}} \mathrm{f}_\xi\Big(N_{a_t^{\mathrm{opt}}}(t)\Big)\\
&\leqslant \sum_{a' \in \widehat{\mathcal{A}}_a^{(d)}(t)} N_{a'}(t)\,\mathrm{kl}(\widehat{\mu}_{a'}(t)\Big|\widehat{\mu}^\star(t) - \widehat{\theta}^\star_{a,a'}(t)\Big) + \mathrm{f}_\xi\Big(N_{\widehat{\mathcal{A}}_a^{(d)}(t)}(t)\Big)\,,\\
&\forall \widehat{a}^\star \in \widehat{\mathcal{A}}^\star(t),\\
&\qquad N_{\dot{a}_t}(t),\ N_{\overline{a}_t}(t),\ \mathbb{I}_{\left\{\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)\right\}} N_{a_t^{\mathrm{opt}}}(t) \leqslant N_{\widehat{a}^\star}(t)\,.
\end{aligned} \tag{4.25}$$

If $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$, from Equations (4.13) and (4.14) defining indexes $(\underline{I}_{a'}(t))_{a' \in \mathcal{A}}$ and $\underline{a}_t$, we have

$$\forall a' \notin \widehat{\mathcal{A}}^\star(t), \quad \mathrm{f}_\xi(N_{a'}(t)) \leqslant \underline{I}_{a'}(t)\,, \tag{4.26}$$

and

$$\forall a' \notin \widehat{\mathcal{A}}^\star(t), \quad \underline{I}_{\underline{a}_t}(t) \leqslant \underline{I}_{a'}(t) \leqslant N_{a'}(t)\,\widehat{\mathrm{k}}^\star_{a'}(t) + \mathrm{f}_\xi(N_{a'}(t))\,. \tag{4.27}$$

□

Lemma 7 (Empirical upper bounds)**.** *Under* `IMED-GS`*, at each step time* $t \geqslant |\mathcal{A}|$ *such that* $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$*,*

$$N_{\dot{a}_t}(t) \leqslant \frac{\mathrm{f}_\xi(t)}{\mathrm{kl}(\widehat{\mu}_{\dot{a}_t}(t)|\widehat{\mu}^\star(t))}\,, \tag{4.28}$$

$$N_{\overline{a}_t}(t) \leqslant \frac{\mathrm{f}_\xi(t)}{\mathrm{kl}(\widehat{\mu}_{\overline{a}_t}(t)|\widehat{\mu}^\star(t))}\,, \tag{4.29}$$

$$N_{a_t^{\mathrm{opt}}}(t) \leqslant N^{\mathrm{opt}}_{a_t^{\mathrm{opt}}}(t)\,, \tag{4.30}$$

$$\mathbb{I}_{\left\{\underline{a}_t \neq a_t^{\mathrm{opt}}\right\}} N_{\dot{a}_t}(t) \leqslant \frac{\mathrm{f}_\xi\Big(N_{a_t^{\mathrm{opt}}}(t)\Big)}{(\gamma_t)^2 \widehat{\mu}^\star(t)(1-\widehat{\mu}^\star(t))\,\mathrm{kl}(\widehat{\mu}_{\dot{a}_t}(t)|\widehat{\mu}^\star(t))}\,, \tag{4.31}$$

$$\mathbb{I}_{\{\underline{a}_t \neq a_t^{\text{opt}}\}} N_{\overline{a}_t}(t) \leqslant \frac{\overline{I}^{(d)}_{\overline{a}_t}(t)}{I_{\dot{a}_t}(t)} \frac{\mathrm{f}_\xi\Big(N_{a_t^{\text{opt}}}(t)\Big) + 1}{(\gamma_t)^2 \widehat{\mu}^\star(t)(1-\widehat{\mu}^\star(t))\,\mathrm{kl}(\widehat{\mu}_{\overline{a}_t}(t)|\widehat{\mu}^\star(t))}\,, \tag{4.32}$$

$$N_{\underline{a}_t}(t) \leqslant \max\left\{ N^{\text{opt}}_{\underline{a}_t}(t)\,,\, \frac{\mathrm{f}_\xi\Big(N_{a_t^{\text{opt}}}(t)\Big)}{\widehat{\mathrm{k}}^\star_{\underline{a}_t}(t)} \right\} \leqslant N^{\text{opt}}_{\underline{a}_t}(t) + \max_{a \notin \widehat{\mathcal{A}}^\star(t)} \frac{\mathrm{f}_\xi(N_a(t))}{\widehat{\mathrm{k}}^\star_{\underline{a}_t}(t)}\,, \tag{4.33}$$

with the convention $0/0 = 0$ *and* $I/0 = \infty$ *for* $I > 0$.

Proof. From Equations (4.9), (4.10) and Equations (4.6), (4.15) defining $\big(\overline{I}_a(t)\big)_{a\in\mathcal{A}}$, $\overline{a}_t$ and $(I_a(t))_{a\in\mathcal{A}}$, $\dot{a}_t$, we have for $\widehat{a}^\star \in \widehat{\mathcal{A}}^\star(t)$,

$$N_{\dot{a}_t}(t)\,\mathrm{kl}(\widehat{\mu}_{\dot{a}_t}(t)|\widehat{\mu}^\star(t)) \leqslant I_{\dot{a}_t}(t) \leqslant I_{\widehat{a}^\star}(t) \leqslant \mathrm{f}_\xi(t) \tag{4.34}$$

and

$$N_{\overline{a}_t}(t)\,\mathrm{kl}(\widehat{\mu}_{\overline{a}_t}(t)|\widehat{\mu}^\star(t)) \leqslant \overline{I}^{(d)}_{\overline{a}_t}(t) \leqslant \overline{I}_{\widehat{a}^\star}(t) \leqslant \mathrm{f}_\xi(t)\,. \tag{4.35}$$

Thus, we deduce Equations (4.28) and (4.29) from Equations (4.34) and (4.35).

From Equation (4.12) that defines a_t^{opt}, proving Equation (4.30) from Lemma 7 amounts to prove

$$\max_{a' \in \overline{\mathcal{A}}_{\overline{a}_t}(t)} N^{\text{opt}}_{a'}(t) - N_{a'}(t) \geqslant 0\,.$$

Since $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$, from Equation (4.11) that defines $(n^{\text{opt}}_a(t))_{a\in\mathcal{A}}$ as solution of current minimization problem, we must have

$$\sum_{a' \in \overline{\mathcal{A}}_{\overline{a}_t}(t)} \mathrm{kl}\Big(\widehat{\mu}_{a'}(t)\Big|\widehat{\mu}^\star(t) - \widehat{\theta}^\star_{\overline{a}_t,a'}(t)\Big) n^{\text{opt}}_{a'}(t) \geqslant 1\,. \tag{4.36}$$

Since $(N^{\text{opt}}_a(t))_{a\in\mathcal{A}} = \big(n^{\text{opt}}_a(t)\,\overline{I}_{\overline{a}_t}(t)\big)_{a\in\mathcal{A}}$, from Equation (4.36) we have

$$\sum_{a' \in \overline{\mathcal{A}}_{\overline{a}_t}(t)} \mathrm{kl}\Big(\widehat{\mu}_{a'}(t)\Big|\widehat{\mu}^\star(t) - \widehat{\theta}^\star_{\overline{a}_t,a'}(t)\Big)\, N^{\text{opt}}_{a'}(t) \geqslant \overline{I}_{\overline{a}_t}(t)\,. \tag{4.37}$$

Since $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$, from Equation (4.9) that explains $\overline{I}_{\overline{a}}(t)$ we have

$$\overline{I}_{\overline{a}_t}(t) \geqslant \sum_{a' \in \overline{\mathcal{A}}_{\overline{a}_t}(t)} \mathrm{kl}\Big(\widehat{\mu}_{a'}(t)\Big|\widehat{\mu}^\star(t) - \widehat{\theta}^\star_{\overline{a}_t,a'}(t)\Big)\, N_{a'}(t)\,. \tag{4.38}$$

By combining Equations (4.37) and (4.38), we get

$$\sum_{a' \in \overline{\mathcal{A}}_{\overline{a}_t}(t)} \mathrm{kl}\Big(\widehat{\mu}_{a'}(t)\Big|\widehat{\mu}^\star(t) - \widehat{\theta}^\star_{\overline{a}_t,a'}(t)\Big)\, \big(N^{\text{opt}}_{a'}(t) - N_{a'}(t)\big) \geqslant 0\,. \tag{4.39}$$

Since $\mathrm{kl}\Big(\widehat{\mu}_{a'}(t)\Big|\widehat{\mu}^\star(t) - \widehat{\theta}^\star_{\overline{a}_t,a'}(t)\Big) \geqslant 0$ for all $a' \in \overline{\mathcal{A}}_{\overline{a}_t}(t)$, Equation (4.39) implies

$$\max_{a' \in \overline{\mathcal{A}}_{\overline{a}_t}(t)} N^{\text{opt}}_{a'}(t) - N_{a'}(t) \geqslant 0$$

which ends the proof of Equation (4.30).

From Lemma 23, we have

$$(\gamma_t)^2\widehat{\mu}^\star(t)(1-\widehat{\mu}^\star(t))\,\mathrm{kl}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) \leqslant \widehat{\mathrm{k}}_a(t)\,, \quad \forall a \in \mathcal{A}\,. \tag{4.40}$$

From previous Equation (4.40) and Equations (4.13) and (4.6) defining $(\underline{I}_a(t))_{a\in\mathcal{A}}$ and $(I_a(t))_{a\in\mathcal{A}}$ we have

$$(\gamma_t)^2\widehat{\mu}^\star(t)(1-\widehat{\mu}^\star(t))\,I_{a'}(t) \leqslant \underline{I}_{a'}(t)\,, \quad \forall a' \neq a_t^{\mathrm{opt}}\,. \tag{4.41}$$

Since $I_{\dot{a}_t}(t) = \min_{a\in\mathcal{A}} I_a(t)$, previous Equation (4.41) implies

$$(\gamma_t)^2\widehat{\mu}^\star(t)(1-\widehat{\mu}^\star(t))\,I_{\dot{a}_t}(t) \leqslant \min_{a'\neq a_t^{\mathrm{opt}}} \underline{I}_{a'}(t)\,. \tag{4.42}$$

Since $\underline{I}_{\underline{a}_t}(t) = \min_{a'\in\mathcal{A}} \underline{I}_{a'}(t)$, previous Equation (4.42) implies

$$\mathbb{I}_{\left\{\underline{a}_t\neq a_t^{\mathrm{opt}}\right\}}(\gamma_t)^2\widehat{\mu}^\star(t)(1-\widehat{\mu}^\star(t))\,I_{\dot{a}_t}(t) \leqslant \mathbb{I}_{\left\{\underline{a}_t\neq a_t^{\mathrm{opt}}\right\}}\underline{I}_{\underline{a}_t}(t) \leqslant \underline{I}_{a_t^{\mathrm{opt}}}(t) = \mathrm{f}_\xi\Big(N_{a_t^{\mathrm{opt}}}(t)\Big)\,. \tag{4.43}$$

Combining Equation (4.34) and (4.43), we get

$$\mathbb{I}_{\left\{\underline{a}_t\neq a_t^{\mathrm{opt}}\right\}}(\gamma_t)^2\widehat{\mu}^\star(t)(1-\widehat{\mu}^\star(t))\,N_{\dot{a}_t}(t)\,\mathrm{kl}(\widehat{\mu}_{\dot{a}_t}(t)|\widehat{\mu}^\star(t)) \leqslant \mathrm{f}_\xi\Big(N_{a_t^{\mathrm{opt}}}(t)\Big)\,. \tag{4.44}$$

We deduce Equation (4.31) from previous Equation (4.44).

We note that

$$\overline{I}^{(d)}_{\overline{a}_t}(t) \leqslant \mathbb{I}_{\left\{I_{\dot{a}_t}(t)\neq 0\right\}}\frac{\overline{I}^{(d)}_{\overline{a}_t}(t)}{I_{\dot{a}_t}(t)}I_{\dot{a}_t}(t) + \mathbb{I}_{\left\{I_{\dot{a}_t}(t)=0\right\}}\overline{I}^{(d)}_{\overline{a}_t}(t)\,. \tag{4.45}$$

Combining previous Equation (4.45) and Equations (4.35), (4.43), we get

$$\mathbb{I}_{\left\{\underline{a}_t\neq a_t^{\mathrm{opt}}\right\}}(\gamma_t)^2\widehat{\mu}^\star(t)(1-\widehat{\mu}^\star(t))\,N_{\overline{a}_t}(t)\,\mathrm{kl}(\widehat{\mu}_{\overline{a}_t}(t)|\widehat{\mu}^\star(t)) \leqslant \frac{\overline{I}^{(d)}_{\overline{a}_t}(t)}{I_{\dot{a}_t}(t)}\,\Big(\mathrm{f}_\xi\Big(N_{a_t^{\mathrm{opt}}}(t)\Big)+1\Big)\,. \tag{4.46}$$

We deduce Equation (4.32) from previous Equation (4.46). □

4.5.3 Non-reliable current means

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

$$\mathcal{E}^+_{a,a'}(\varepsilon) \coloneqq \{t \geqslant 1 :\ a_{t+1} = a',\ N_{a'}(t) \leqslant N_a(t),\ \widehat{\mu}_a(t) \geqslant \mu_a + \varepsilon\}\,. \tag{4.47}$$

We similarly define

$$\mathcal{E}^-_{a,a'}(\varepsilon) \coloneqq \{t \geqslant 1 :\ a_{t+1} = a',\ N_{a'}(t) \leqslant N_a(t),\ \widehat{\mu}_a(t) \leqslant \mu_a - \varepsilon\}\,. \tag{4.48}$$

We also define

$$\mathcal{E}_{a,a'}(\varepsilon) = \mathcal{E}^+_{a,a'}(\varepsilon) \cup \mathcal{E}^-_{a,a'}(\varepsilon)\,. \tag{4.49}$$

Definition 2 (kl-f_ξ deviation)**.** *For $\varepsilon>0$, the couple of arms $(a,a')\in\mathcal{A}^2$ shows ε^+-kl-f_ξ deviation at time step $t\geqslant 1$ if the following conditions are satisfied,*

$$\begin{aligned}&(1)\quad a_{t+1}=a'\\&(2)\quad \widehat{\mu}_a(t)\geqslant \mu_a+\varepsilon\\&(3)\quad \mathrm{f}_\xi(N_{a'}(t))\leqslant N_a(t)\,\mathrm{kl}(1-\widehat{\mu}_a(t)|1-\mu_a-\varepsilon)+\mathrm{f}_\xi(N_a(t))\,.\end{aligned}$$

For $\varepsilon>0$, the couple of arms $(a,a')\in\mathcal{A}^2$ shows ε^--kl-f_ξ deviation at time step $t\geqslant 1$ if the following conditions are satisfied,

$$\begin{aligned}&(1)\quad a_{t+1}=a'\\&(2)\quad \widehat{\mu}_a(t)\leqslant \mu_a-\varepsilon\\&(3)\quad \mathrm{f}_\xi(N_{a'}(t))\leqslant N_a(t)\,\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a-\varepsilon)+\mathrm{f}_\xi(N_a(t))\,.\end{aligned}$$

For all couple of arms $(a,a')\in\mathcal{A}^2$ and for all accuracy $\varepsilon>0$, let $\mathcal{K}^+_{a,a'}(\varepsilon)$ be the set of times where couple of arms (a,a') shows ε^+-kl-f_ξ deviation, that is

$$\mathcal{K}^+_{a,a'}(\varepsilon)\coloneqq\left\{t\geqslant 1:\ \begin{array}{ll}(1)&a_{t+1}=a'\\(2)&\widehat{\mu}_a(t)\geqslant\mu_a+\varepsilon\\(3)&\mathrm{f}_\xi(N_{a'}(t))\leqslant N_a(t)\,\mathrm{kl}(1-\widehat{\mu}_a(t)|1-\mu_a-\varepsilon)+\mathrm{f}_\xi(N_a(t))\end{array}\right\}. \tag{4.50}$$

For all couple of arms $(a,a')\in\mathcal{A}^2$ and for all accuracy $\varepsilon>0$, let $\mathcal{K}^-_{a,a'}(\varepsilon)$ be the set of times where couple of arms (a,a') shows ε^--kl-f_ξ deviation, that is

$$\mathcal{K}^-_{a,a'}(\varepsilon)\coloneqq\left\{t\geqslant 1:\ \begin{array}{ll}(1)&a_{t+1}=a'\\(2)&\widehat{\mu}_a(t)\leqslant\mu_a-\varepsilon\\(3)&\mathrm{f}_\xi(N_{a'}(t))\leqslant N_a(t)\,\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a-\varepsilon)+\mathrm{f}_\xi(N_a(t))\end{array}\right\}. \tag{4.51}$$

We note that,

$$\mathcal{E}^+_{a,a'}(\varepsilon)\subset\mathcal{K}^+_{a,a'}(\varepsilon)\qquad\qquad\mathcal{E}^-_{a,a'}(\varepsilon)\subset\mathcal{K}^-_{a,a'}(\varepsilon)\,.$$

We also define

$$\mathcal{K}_{a,a'}(\varepsilon)=\mathcal{K}^+_{a,a'}(\varepsilon)\cup\mathcal{K}^-_{a,a'}(\varepsilon)\,. \tag{4.52}$$

Finally, for accuracy $\varepsilon>0$, we denote by

$$\mathcal{K}^\star(\varepsilon)\coloneqq\left\{t\geqslant 1:\ \begin{array}{ll}(1)&\widehat{\mu}_a(t)\leqslant\mu_a-\varepsilon,\quad\forall a\in\widehat{\mathcal{A}}_{a^\star}(t)\\(2)&1\leqslant N_{\widehat{\mathcal{A}}_{a^\star}(t)}(t)\leqslant\left|\widehat{\mathcal{A}}_{a^\star}(t)\right|\mathrm{f}_\xi\big(N_{a_{t+1}}(t)\big)/2\varepsilon^2\\(3)&\displaystyle\sum_{a\in\widehat{\mathcal{A}}_{a^\star}(t)}N_a(t)\,\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a-\varepsilon)\\&\geqslant\mathbb{I}_{\left\{|\widehat{\mathcal{A}}_{a^\star}(t)|>d+1\right\}}\Phi\Big(2\left|\widehat{\mathcal{A}}_{a^\star}(t)\right|+1\Big))\vee\Big(\mathrm{f}_\xi\big(N_{a_{t+1}}(t)\big)-\mathrm{f}_\xi\Big(N_{\widehat{\mathcal{A}}_{a^\star}(t)}(t)\Big)\Big)\\&+\mathbb{I}_{\left\{|\widehat{\mathcal{A}}_{a^\star}(t)|\leqslant d+1\right\}}(d+1)\vee\Big(\mathrm{f}_\xi\big(N_{a_{t+1}}(t)\big)-\mathrm{f}_\xi\Big(N_{\widehat{\mathcal{A}}_{a^\star}(t)}(t)\Big)\Big)\end{array}\right\} \tag{4.53}$$

$$\mathcal{K}^\star_{a',\mathcal{A}'}(\varepsilon)=\Big\{t\in\mathcal{K}^\star(\varepsilon):\ a_{t+1}=a',\ \widehat{\mathcal{A}}_{a^\star}(t)=\mathcal{A}'\Big\},\quad a'\in\mathcal{A},\ \mathcal{A}'\subset\mathcal{A} \tag{4.54}$$

and

$$\mathcal{L}^{(d)}(\varepsilon)=\bigcup_{a\in\mathcal{A}}\mathcal{L}^{(d)}_a(\varepsilon)\qquad\mathcal{L}^{(d)}_a(\varepsilon)=\left\{t\geqslant 1:\ \begin{array}{ll}(1)&\widehat{\mu}_a(t)\leqslant\mu_a-\varepsilon\\(2)&N_a(t)\geqslant\dfrac{\mathrm{f}_\xi\big(N_{a_{t+1}}(t)\big)}{2\varepsilon^2}\wedge\dfrac{e^{-d-1}N_{a_{t+1}}(t)}{d+1}\end{array}\right\} \tag{4.55}$$

Lemma 8 (Bounded subsets of times). *For $\varepsilon>0$, for $(a,a')\in\mathcal{A}^2$,*

$$\mathbb{E}_\nu\left[\left|\mathcal{E}^+_{a,a'}(\varepsilon)\right|\right],\ \mathbb{E}_\nu\left[\left|\mathcal{E}^-_{a,a'}(\varepsilon)\right|\right]\leqslant\frac{e^{2\varepsilon^2}}{2\varepsilon^2} \tag{4.56}$$

$$\mathbb{E}_\nu\left[\left|\mathcal{L}^{(d)}(\varepsilon)\right|\right]\leqslant|\mathcal{A}|\,\frac{(d+1)e^{d+1}}{2\varepsilon^2}\left(\mathrm{f}_\xi\left(\frac{(d+1)e^{d+1}}{2\varepsilon^2}\right)\right)^2+\sum_{n\geqslant3}\frac{|\mathcal{A}|^2}{n\left(\log(n)\right)^\xi} \tag{4.57}$$

$$\mathbb{E}_\nu\left[\left|\mathcal{K}^+_{a,a'}(\varepsilon)\right|\right],\ \mathbb{E}_\nu\left[\left|\mathcal{K}^-_{a,a'}(\varepsilon)\right|\right]\leqslant K_{\xi,1}+\sum_{n\geqslant3}\frac{1}{n\left(\log(n)\right)^\xi} \tag{4.58}$$

$$\mathbb{E}_\nu[|\mathcal{K}^\star(\varepsilon)|]\leqslant K^\star_\xi\,, \tag{4.59}$$

where for $\xi>1$,

$$\begin{aligned}
K^\star_\xi &= |\mathcal{A}|\left(e^2\vee\frac{|\mathcal{A}|}{2\varepsilon^2}\right)\left(\mathrm{f}_\xi\left(e^2\vee\frac{|\mathcal{A}|}{2\varepsilon^2}\right)\right)^2\\
&+\sum_{n\geqslant3}\sum_{\mathcal{A}'\subset\mathcal{A}}\frac{e^{|\mathcal{A}'|+2}}{|\mathcal{A}'|^{|\mathcal{A}'|}}\log(n)^{|\mathcal{A}'|+1}\left[\Phi(2\,|\mathcal{A}'|+1)\vee\mathrm{f}_\xi(n)+2\varepsilon^2\underline{N}\right]^{|\mathcal{A}'|+1}e^{-\left[\Phi(2|\mathcal{A}'|+1)\vee\mathrm{f}_\xi(n)+2\varepsilon^2\underline{N}\right]}\\
&+\sum_{n\geqslant3}\sum_{\substack{\mathcal{A}'\subset\mathcal{A}\\|\mathcal{A}'|\leqslant d+1}}\frac{e^{|\mathcal{A}'|+2}}{|\mathcal{A}'|^{|\mathcal{A}'|}}\log(n)^{|\mathcal{A}'|+1}\left[(d+1)\vee\mathrm{f}_\xi(n)+2\varepsilon^2\underline{N}\right]^{|\mathcal{A}'|+1}e^{-\left[(d+1)\vee\mathrm{f}_\xi(n)+2\varepsilon^2\underline{N}\right]},
\end{aligned}$$

$$K_{\xi,1}=\left(e^2\vee\frac{|\mathcal{A}|}{2\varepsilon^2}\right)\left(\mathrm{f}_\xi\left(e^2\vee\frac{|\mathcal{A}|}{2\varepsilon^2}\right)\right)^2+\sum_{n\geqslant3}e^3\log(n)^3\left[\mathrm{f}_\xi(n)+2\varepsilon^2\underline{N}\right]^3e^{-\left[\mathrm{f}_\xi(n)+2\varepsilon^2\underline{N}\right]},$$

in which we introduce $\underline{N}=1\vee\dfrac{\mathrm{f}_\xi(n)}{1-\log(1-\max_{a\in\mathcal{A}'}\mu_a)}$.

Proof. *We start by* proving $\mathbb{E}_\nu\left[\left|\mathcal{E}^-_{a,a'}(\varepsilon)\right|\right]\leqslant e^{2\varepsilon^2}/2\varepsilon^2$. The proof that $\mathbb{E}_\nu\left[\left|\mathcal{E}^+_{a,a'}(\varepsilon)\right|\right]\leqslant e^{2\varepsilon^2}/2\varepsilon^2$ is similar.

We write

$$\left|\mathcal{E}^-_{a,a'}(\varepsilon)\right|=\sum_{t\geqslant1}\mathbb{I}_{\{a_{t+1}=a',N_{a'}(t)\leqslant N_a(t),\ \mu_a-\widehat{\mu}_a(t)\geqslant\varepsilon\}}\,. \tag{4.60}$$

Considering the stopped stopping times $\tau_n=\inf\{t\geqslant1,N_{a'}(t)=n\}$ we will rewrite the previous sum of indicators and use Lemma 32.

$$\begin{aligned}
\left|\mathcal{E}^+_{a,a'}(\varepsilon)\right| &\leqslant \sum_{t\geqslant1}\mathbb{I}_{\{a_{t+1}=a',\ N_{a'}(t)\leqslant N_a(t),\ \mu_a-\widehat{\mu}_a(t)\geqslant\varepsilon\}} \qquad (4.61)\\
&\leqslant \sum_{n\geqslant1}\mathbb{I}_{\{n-1\leqslant N_a(\tau_n-1),\ \mu_a-\widehat{\mu}_a(\tau_n-1)\geqslant\varepsilon\}}\\
&\leqslant 1+\sum_{n\geqslant2}\mathbb{I}_{\{n-1\leqslant N_a(\tau_n-1),\ \mu_a-\widehat{\mu}_a(\tau_n-1)\geqslant\varepsilon\}}\,.
\end{aligned}$$

Taking the expectation of Equation (4.61), it comes

$$\mathbb{E}_\nu\left[\left|\mathcal{E}^+_{a,a'}(\varepsilon)\right|\right]\leqslant1+\sum_{n\geqslant1}\mathbb{P}_\nu\left(\bigcup_{\substack{t\geqslant|\mathcal{A}|\\\widehat{\mu}_a(t)<\mu_a\\N_a(t)\geqslant n}}\mu_a-\widehat{\mu}_a(t)\geqslant\varepsilon\right). \tag{4.62}$$

By Pinsker's inequality, previous Equation (4.62) implies

$$\mathbb{E}_\nu\big[\big|\mathcal{E}^+_{a,a'}(\varepsilon)\big|\big] \leqslant 1 + \sum_{n\geqslant 1} \mathbb{P}_\nu\left(\bigcup_{\substack{t\geqslant|\mathcal{A}|\\ \widehat{\mu}_a(t)<\mu_a\\ N_a(t)\geqslant n}} \mathrm{kl}(\widehat{\mu}_a(t)|\mu_a) \geqslant 2\varepsilon^2\right). \tag{4.63}$$

From Lemma 32, previous Equation (4.63) implies

$$\mathbb{E}_\nu\big[\big|\mathcal{E}^+_{a,a'}(\varepsilon)\big|\big] \leqslant \sum_{n\geqslant 0} \exp\big(-2n\varepsilon^2\big) = \frac{1}{1-e^{-2\varepsilon^2}}\,. \tag{4.64}$$

Finally we note that

$$\frac{1}{1-e^{-2\varepsilon^2}} = \frac{e^{2\varepsilon^2}}{e^{2\varepsilon^2}-1} \leqslant \frac{e^{2\varepsilon^2}}{2\varepsilon^2}\,,$$

which ends the proof.

We then prove Equation (4.57). By using in particular Pinsker's inequality, we have

$$\begin{aligned}
\big|\mathcal{L}^{(d)}(\varepsilon)\big| &\leqslant \left|\left\{t\geqslant 1:\ N_{a_{t+1}}(t) < \frac{(d+1)e^{d+1}}{2\varepsilon^2}\left(\mathrm{f}_\xi\left(\frac{(d+1)e^{d+1}}{2\varepsilon^2}\right)\right)^2\right\}\right| \\
&\quad + \sum_{a\in\mathcal{A}} \big|\big\{t\geqslant 1:\ N_a(t)\geqslant \mathrm{f}_\xi\big(N_{a_{t+1}}(t)\big)/2\varepsilon^2,\ \widehat{\mu}_a(t)<\mu_a,\ \mathrm{kl}\big(\widehat{\mu}_a(t)\,\big|\mu_a\big)\geqslant 2\varepsilon^2\big\}\big| \\
&\leqslant |\mathcal{A}|\,\frac{(d+1)e^{d+1}}{2\varepsilon^2}\left(\mathrm{f}_\xi\left(\frac{(d+1)e^{d+1}}{2\varepsilon^2}\right)\right)^2 \\
&\quad + \sum_{a\in\mathcal{A}} |\mathcal{A}| \sum_{n\geqslant 3} \big|\big\{t\geqslant 1:\ N_a(t)\geqslant \mathrm{f}_\xi(n)/2\varepsilon^2,\ \widehat{\mu}_a(t)<\mu_a,\ \mathrm{kl}\big(\widehat{\mu}_a(t)\,\big|\mu_a\big)\geqslant 2\varepsilon^2\big\}\big|\,.
\end{aligned}$$

By taking the expectation on both sides of previous inequality, it comes

$$\begin{aligned}
\mathbb{E}_\nu\big[\big|\mathcal{L}^{(d)}(\varepsilon)\big|\big] &\leqslant |\mathcal{A}|\,\frac{(d+1)e^{d+1}}{2\varepsilon^2}\left(\mathrm{f}_\xi\left(\frac{(d+1)e^{d+1}}{2\varepsilon^2}\right)\right)^2 \\
&\quad + \sum_{a\in\mathcal{A}} |\mathcal{A}| \sum_{n\geqslant 3} \mathbb{P}_\nu\left(\bigcup_{\substack{t\geqslant 1\\ \widehat{\mu}_a(t)<\mu_a\\ N_a(t)\geqslant \mathrm{f}_\xi(n)/2\varepsilon^2}} \mathrm{kl}\big(\widehat{\mu}_a(t)\,\big|\mu_a\big)\geqslant 2\varepsilon^2\right).
\end{aligned} \tag{4.65}$$

From Lemma 32, previous Equation (4.65) implies

$$\begin{aligned}
\mathbb{E}_\nu\big[\big|\mathcal{L}^{(d)}(\varepsilon)\big|\big] &\leqslant |\mathcal{A}|\,\frac{(d+1)e^{d+1}}{2\varepsilon^2}\left(\mathrm{f}_\xi\left(\frac{(d+1)e^{d+1}}{2\varepsilon^2}\right)\right)^2 \\
&\quad + \sum_{a\in\mathcal{A}} |\mathcal{A}| \sum_{n\geqslant 3} e^{-\mathrm{f}_\xi(n)}\,,
\end{aligned} \tag{4.66}$$

which ends the proof.

We now prove Equation (4.59). The remaining inequalities are proven similarly. We simply note that

$$
\begin{aligned}
& |\mathcal{K}^\star(\varepsilon)| \\
\leqslant & \sum_{\substack{a'\in\mathcal{A}\\ \mathcal{A}'\subset\mathcal{A}}} \left|\mathcal{K}^\star_{a',\mathcal{A}'}(\varepsilon)\right| \\
\leqslant & \sum_{a'\in\mathcal{A}} n_{\xi,\mathcal{A},\varepsilon} \\
& + \sum_{\substack{a'\in\mathcal{A}\\ \mathcal{A}'\subset\mathcal{A}\\ n\geqslant n_{\xi,\mathcal{A},\varepsilon}}} \left| \left\{ t\geqslant 1 : \begin{array}{c} \widehat{\mu}_a(t) < \mu_a - \varepsilon, \quad \forall a\in\mathcal{A}' \\ 1\leqslant N_{\mathcal{A}'}(t) \leqslant |\mathcal{A}'|\,\mathrm{f}_\xi(n)/2\varepsilon^2 \\ \sum_{a\in\mathcal{A}'} N_a(t)\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a-\varepsilon) \geqslant \Phi(2\,|\mathcal{A}'|+1)\vee(\mathrm{f}_\xi(n)-\mathrm{f}_\xi(N_{\mathcal{A}'}(t))) \end{array} \right\} \right| \\
& + \sum_{\substack{a'\in\mathcal{A}\\ \mathcal{A}'\subset\mathcal{A},|\mathcal{A}'|\leqslant d+1\\ n\geqslant n_{\xi,\mathcal{A},\varepsilon}}} \left| \left\{ t\geqslant 1 : \begin{array}{c} \widehat{\mu}_a(t) < \mu_a - \varepsilon, \quad \forall a\in\mathcal{A}' \\ 1\leqslant N_{\mathcal{A}'}(t) \leqslant |\mathcal{A}'|\,\mathrm{f}_\xi(n)/2\varepsilon^2 \\ \sum_{a\in\mathcal{A}'} N_a(t)\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a-\varepsilon) \geqslant (d+1)\vee(\mathrm{f}_\xi(n)-\mathrm{f}_\xi(N_{\mathcal{A}'}(t))) \end{array} \right\} \right|,
\end{aligned}
$$

where $n_{\xi,\mathcal{A},\varepsilon} = \left(e^2\vee\frac{|\mathcal{A}|}{2\varepsilon^2}\right)\left(\mathrm{f}_\xi\left(e^2\vee\frac{|\mathcal{A}|}{2\varepsilon^2}\right)\right)^2$. The proof ends by taking the expectation on both sides of previous inequality and by applying Theorem 4. □

4.5.4 Non-reliable current best arm and current informative sets of arms

For accuracy $\varepsilon > 0$, let $\mathcal{M}^\star(\varepsilon)$ be the set of times $t\geqslant 1$ that do not belong to $\cup_{\widehat{a}^\star\in\widehat{\mathcal{A}}^\star(t)}\mathcal{E}^+_{\widehat{a}^\star,a_{t+1}}(\varepsilon)$ and where some of the current best arms do not belong to $\mathcal{A}^\star$,

$$
\mathcal{M}^\star(\varepsilon) := \left\{ t\geqslant 1 : \begin{array}{ll} (1) & t\notin \bigcup\limits_{\widehat{a}^\star\in\widehat{\mathcal{A}}^\star(t)} \mathcal{E}^+_{\widehat{a}^\star,a_{t+1}}(\varepsilon) \\ (2) & \widehat{\mathcal{A}}^\star(t)\neq\{a^\star\} \end{array} \right\}. \tag{4.67}
$$

For accuracy $\varepsilon>0$, let $\mathcal{M}(\varepsilon)$ be the set of times $t\geqslant 1$ during exploration phases where the current means of current informative arms are not well ε-estimated,

$$
\mathcal{M}(\varepsilon) := \left\{ t\geqslant 1 : \begin{array}{ll} (1) & \overline{a}_t\notin\widehat{\mathcal{A}}^\star(t),\ a_{t+1}=\underline{a}_t=a_t^{\mathrm{opt}} \\ (2) & \exists a'\notin\widehat{\mathcal{A}}^\star(t), \quad |\widehat{\mu}_{a'}(t)-\mu_{a'}|\geqslant\varepsilon+\widehat{\varepsilon}^\star_{a'}(t) \end{array} \right\}. \tag{4.68}
$$

Lemma 9 (Relation between subsets of times)**.** *For accuracy* $0<\varepsilon<\varepsilon_0 := \delta_{\min}/3$,

$$
\mathcal{M}^\star(\varepsilon) \subset \mathcal{K}^\star(\varepsilon_0)\cup\mathcal{L}^{(d)}(\varepsilon_0)\,, \tag{4.69}
$$

$$
\mathcal{M}(\varepsilon) \subset \bigcup_{a\in\mathcal{A},\ t\geqslant 1} \mathcal{K}_{a,a_{t+1}}(\varepsilon)\,. \tag{4.70}
$$

Proof. We start by proving Equation (4.69).

Let us consider $t\in\mathcal{M}^\star(\varepsilon)$. Then there exists $\widehat{a}^\star\in\widehat{\mathcal{A}}^\star(t)\setminus\{a^\star\}$ and

$$
\widehat{\mu}_{\widehat{a}^\star}(t) = \widehat{\mu}^\star(t) \geqslant \widehat{\mu}_{a^\star}(t)\,. \tag{4.71}
$$

Since $t \in \mathcal{M}^\star(\varepsilon)$, $t \notin \mathcal{E}^+_{\widehat{a}^\star, a_{t+1}}(\varepsilon)$. By considering empirical lower bounds (4.17) and Equation (4.47), we have

$$\mu_{\widehat{a}^\star} + \varepsilon \geqslant \widehat{\mu}_{\widehat{a}^\star}(t)\,. \tag{4.72}$$

By combining Equations (4.71) and (4.72), it comes

$$\mu_{\widehat{a}^\star} + \varepsilon \geqslant \widehat{\mu}_{a^\star}(t)\,. \tag{4.73}$$

Since $\widehat{a}^\star \notin \mathcal{A}^\star$, $\varepsilon < \varepsilon_0 \leqslant \Delta_{\widehat{a}^\star}/2$ and

$$\mu^\star - \varepsilon_0 > \mu_{\widehat{a}^\star} + \varepsilon \geqslant \widehat{\mu}_{\widehat{a}^\star}(t)\,. \tag{4.74}$$

Then Equation (4.73) implies

$$\mu_{a^\star} - \varepsilon_0 > \widehat{\mu}_{a^\star}(t)\,. \tag{4.75}$$

Since $t \in \mathcal{M}^\star(\varepsilon)$, $t \notin \cup_{\widehat{a}^\star \in \widehat{\mathcal{A}}^\star(t)} \mathcal{E}^+_{\widehat{a}^\star, a_{t+1}}(\varepsilon)$. By considering empirical lower bounds (4.17) and Equation (4.47), Equation (4.75) implies

$$a^\star \notin \widehat{\mathcal{A}}^\star(t)\,. \tag{4.76}$$

Let us then consider $a \in \widehat{\mathcal{A}}_{a^\star}(t)$. From Equation (4.8), this means

$$\begin{aligned} \widehat{\mu}_a(t) &\leqslant \widehat{\mu}^\star(t) - \widehat{\theta}^\star_{a^\star, a} \\ &= \widehat{\mu}_{\widehat{a}^\star}(t) - \theta^{(\widehat{a}^\star)}_{a^\star, a} \end{aligned}\,. \tag{4.77}$$

From Assumption 8, previous Equation (4.77) implies

$$\widehat{\mu}_a(t) \leqslant \widehat{\mu}_{\widehat{a}^\star}(t) - \theta^{(a^\star)}_{\widehat{a}^\star, a} = \widehat{\mu}_{\widehat{a}^\star}(t) - \theta^\star_{\widehat{a}^\star, a} \tag{4.78}$$

By combining Equations (4.73) and (4.78), it comes

$$\widehat{\mu}_a(t) \leqslant \widehat{\mu}^\star(t) - \widehat{\theta}^\star_{a^\star, a} < \mu_{\widehat{a}^\star} - \theta^\star_{\widehat{a}^\star, a} + \varepsilon\,. \tag{4.79}$$

From Equation (4.1) and Assumption 8, we have

$$\mu_{\widehat{a}^\star} - \theta^\star_{\widehat{a}, a} \leqslant \mu_a - 2\varepsilon_0\,. \tag{4.80}$$

By combining Equations (4.79) and (4.80), we show that

$$\widehat{\mu}_a(t) \leqslant \widehat{\mu}^\star(t) - \widehat{\theta}^\star_{a^\star, a} \leqslant \mu_a - \varepsilon_0, \quad \forall a \in \widehat{\mathcal{A}}_{a^\star}(t) \supset \widehat{\mathcal{A}}^{(d)}_{a^\star}(t)\,. \tag{4.81}$$

Since $a^\star \notin \widehat{\mathcal{A}}^\star(t)$, from empirical lower bounds (4.16) we have

$$\mathrm{f}_\xi\big(N_{a_{t+1}}(t)\big) \leqslant \sum_{a \in \widehat{\mathcal{A}}^{(d)}_{a^\star}(t)} N_a(t)\,\mathrm{kl}(\widehat{\mu}_a(t) | \widehat{\mu}^\star(t) - \theta_{a^\star, a}) + \mathrm{f}_\xi\Big(N_{\widehat{\mathcal{A}}^{(d)}_{a^\star}(t)}(t)\Big)\,. \tag{4.82}$$

The monotony of $\mathrm{kl}(\widehat{\mu}_a(t) | \cdot)$, for $a \in \widehat{\mathcal{A}}^{(d)}_{a^\star}(t)$ and Equation (4.81) imply

$$\mathrm{kl}(\widehat{\mu}_a(t) | \widehat{\mu}^\star(t) - \theta_{a^\star, a}) \leqslant \mathrm{kl}(\widehat{\mu}_a(t) | \mu_a - \varepsilon_0)\,, \quad \forall a \in \widehat{\mathcal{A}}^{(d)}_{a^\star}(t)\,. \tag{4.83}$$

By combining Equations (4.82) and (4.83), we get

$$\mathrm{f}_\xi\big(N_{a_{t+1}}(t)\big) \leqslant \sum_{a \in \widehat{\mathcal{A}}^{(d)}_{a^\star}(t)} N_a(t)\,\mathrm{kl}(\widehat{\mu}_a(t) | \mu_a - \varepsilon_0) + \mathrm{f}_\xi\Big(N_{\widehat{\mathcal{A}}^{(d)}_{a^\star}(t)}(t)\Big) \tag{4.84}$$

or equivalently

$$\sum_{a\in\widehat{\mathcal{A}}^{(d)}_{a^\star}(t)} N_a(t)\,\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a-\varepsilon_0) \geqslant \mathrm{f}_\xi\big(N_{a_{t+1}}(t)\big) - \mathrm{f}_\xi\Big(N_{\widehat{\mathcal{A}}^{(d)}_{a^\star}(t)}(t)\Big)\,. \tag{4.85}$$

According to the definition of $\widehat{\mathcal{A}}^{(d)}_{a^\star}(t)$, if $\left|\widehat{\mathcal{A}}^{(d)}_{a^\star}(t)\right| > d+1$ we have $\sum_{a\in\widehat{\mathcal{A}}_{a^\star}(t)} 2(\widehat{\mu}^\star(t)-\widehat{\theta}^\star_{a^\star,a}-\widehat{\mu}_a(t))^2 N_a(t) \geqslant \Phi(2|\widehat{\mathcal{A}}^{(d)}_{a^\star}(t)|+1)$. Then, from Equations (4.81), (4.85) and Pinsker's inequality (Lemma 24), this implies $\sum_{a\in\widehat{\mathcal{A}}^{(d)}_{a^\star}(t)} N_a(t)\,\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a-\varepsilon_0) \geqslant \Phi(2|\widehat{\mathcal{A}}^{(d)}_{a^\star}(t)|+1)\vee\Big(\mathrm{f}_\xi\big(N_{a_{t+1}}(t)\big)-\mathrm{f}_\xi\Big(N_{\widehat{\mathcal{A}}^{(d)}_{a^\star}(t)}(t)\Big)\Big)$. Furthermore, if $\left|\widehat{\mathcal{A}}^{(d)}_{a^\star}(t)\right| \leqslant d+1$ and $N_{\widehat{\mathcal{A}}^{(d)}_{a^\star}(t)}(t) \leqslant e^{-d-1}N_{a_{t+1}}(t)$, we have $\sum_{a\in\widehat{\mathcal{A}}^{(d)}_{a^\star}(t)} N_a(t)\,\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a-\varepsilon_0) \geqslant \mathrm{f}_\xi\big(N_{a_{t+1}}(t)\big)-\mathrm{f}_\xi\Big(N_{\widehat{\mathcal{A}}^{(d)}_{a^\star}(t)}(t)\Big) \geqslant \log\big(N_{a_{t+1}}(t)\big)-\log\Big(N_{\widehat{\mathcal{A}}^{(d)}_{a^\star}(t)}(t)\Big) \geqslant d+1$.

Now we prove Equation (4.70).

Since $t\in\mathcal{M}(\varepsilon)$, there exists $a'\notin\widehat{\mathcal{A}}^\star(t)$, such that

$$|\widehat{\mu}_{a'}(t)-\mu_{a'}| \geqslant \varepsilon+\widehat{\varepsilon}^\star_{a'}(t)\,. \tag{4.86}$$

From Equation (4.86), by triangle inequality we have

$$|\widehat{\mu}_{a'}(t)-\mu_{a'}-\varepsilon| \geqslant \widehat{\varepsilon}^\star_{a'}(t)\,. \tag{4.87}$$

Let us introduce

$$\widehat{\mathrm{k}}_{a'}(\varepsilon,t) := \mathbb{I}_{\{\widehat{\mu}_a(t)<\mu_{a'}-\varepsilon\}}\mathrm{kl}(\widehat{\mu}_{a'}(t)|\mu_{a'}-\varepsilon)+\mathbb{I}_{\{1-\widehat{\mu}_{a'}(t)<1-\mu_{a'}-\varepsilon\}}\mathrm{kl}(1-\widehat{\mu}_{a'}(t)|1-\mu_{a'}-\varepsilon) \tag{4.88}$$

$$\widehat{\mathrm{k}}_{a'}(t) := 2\big(\widehat{\varepsilon}^\star_{a'}(t)\big)^2\,. \tag{4.89}$$

From Equation (4.88), by Pinsker's inequality we have

$$\widehat{\mathrm{k}}_{a'}(\varepsilon,t) \geqslant 2\left(\widehat{\mu}_{a'}(t)-\mu_{a'}-\varepsilon\right)^2\,. \tag{4.90}$$

By combining Equations (4.87) and (4.90) and Equation (4.89) that defines $\widehat{\mathrm{k}}_{a'}(t)$, we obtain

$$\widehat{\mathrm{k}}_{a'}(\varepsilon,t) \geqslant 2\big(\widehat{\varepsilon}^\star_{a'}(t)\big)^2 = \widehat{\mathrm{k}}^\star_{a'}(t)\,. \tag{4.91}$$

Since $a'\in\widehat{\mathcal{A}}^\star(t)$, by combining empirical lower bounds (4.18) and Equation (4.91), we have

$$\mathrm{f}_\xi\big(N_{\underline{a}_t}(t)\big) \leqslant N_{a'}(t)\,\widehat{\mathrm{k}}_{a'}(\varepsilon,t)+\mathrm{f}_\xi(N_{a'}(t))\,. \tag{4.92}$$

Since $t\in\mathcal{M}(\varepsilon)$, $\overline{a}_t\notin\widehat{\mathcal{A}}^\star(t)$ and

$$a_{t+1} = \underline{a}_t\,. \tag{4.93}$$

By combining Equations (4.88), (4.92) and (4.93), we have

$$t\in\mathcal{K}^+_{a',a_{t+1}}(\varepsilon)\cup\mathcal{K}^-_{a',a_{t+1}}(\varepsilon) = \mathcal{K}_{a',a_{t+1}}(\varepsilon)\,.$$

□

4.5.5 Reliable current means of current informative sets of Arms

For accuracy $0<\varepsilon<\varepsilon_0=\delta_{\min}/3$, we define the union of the previous considered subsets of times

$$\mathcal{U}(\varepsilon) := \mathcal{M}(\varepsilon)\bigcup\mathcal{M}^\star(\varepsilon)\bigcup\mathcal{K}^\star(\varepsilon_0)\bigcup\mathcal{L}^{(d)}(\varepsilon_0)\bigcup_{a\in\mathcal{A},\ t\geqslant 1}\mathcal{K}_{a,a_{t+1}}(\varepsilon)\bigcup_{a\in\mathcal{A},\ t\geqslant 1}\mathcal{E}_{a,a_{t+1}}(\varepsilon)\,. \tag{4.94}$$

From Lemma 9, we have

$$\mathcal{U}(\varepsilon) = \mathcal{K}^\star(\varepsilon_0)\bigcup\mathcal{L}^{(d)}(\varepsilon_0)\bigcup_{a\in\mathcal{A},\ t\geqslant 1}\mathcal{K}_{a,a_{t+1}}(\varepsilon)\bigcup_{a\in\mathcal{A},\ t\geqslant 1}\mathcal{E}_{a,a_{t+1}}(\varepsilon)\,. \tag{4.95}$$

Lemma 10 (Reliable current mean of current pulled arm)**.** *For accuracy $0<\varepsilon<\varepsilon_0=\delta_{\min}/3$, for all $t\notin\mathcal{U}(\varepsilon)$,*

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

Proof. Let us consider $t\notin\mathcal{U}(\varepsilon)$. Then $t\notin\mathcal{E}_{a_{t+1},a_{t+1}}(\varepsilon)$. □

Lemma 11 (Reliable current best mean)**.** *For accuracy $0<\varepsilon<\varepsilon_0=\delta_{\min}/3$, for all $t\notin\mathcal{U}(\varepsilon)$,*

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

$$|\widehat{\mu}^\star(t)-\mu^\star|<\varepsilon\,.$$

Proof. Let us consider $t\notin\mathcal{U}(\varepsilon)$. Then $t\notin\mathcal{M}^\star(\varepsilon)\bigcup_{\widehat{a}^\star\in\widehat{\mathcal{A}}^\star(t)}\mathcal{E}^+_{\widehat{a}^\star,a_{t+1}}(\varepsilon)$ and Equation (4.67) implies

$$\widehat{\mathcal{A}}^\star(t)=\{a^\star\}\ . \tag{4.96}$$

Since $t\notin\mathcal{U}(\varepsilon)$, $t\notin\bigcup_{\widehat{a}^\star\in\widehat{\mathcal{A}}^\star(t)}\mathcal{E}_{\widehat{a}^\star,a_{t+1}}(\varepsilon)$. By considering empirical lower bounds (4.17) and Equation (4.48), we have

$$|\widehat{\mu}_{\widehat{a}^\star}(t)-\mu_{\widehat{a}^\star}|<\varepsilon,\quad\forall\widehat{a}^\star\in\widehat{\mathcal{A}}^\star(t)\,. \tag{4.97}$$

Since $\widehat{\mathcal{A}}^\star(t)=\underset{a\in\mathcal{A}}{\operatorname{argmax}}\,\widehat{\mu}_a(t)$, by combining Equations (4.96) and (4.97), we have

$$|\widehat{\mu}^\star(t)-\mu^\star|<\varepsilon\,. \tag{4.98}$$

□

Lemma 12 (Reliable current means)**.** *For accuracy $0<\varepsilon<\varepsilon_0=\delta_{\min}/3$, for all time step $t\notin\mathcal{U}(\varepsilon)$ that corresponds to an exploration phase (that is $\overline{a}_t\notin\widehat{\mathcal{A}}^\star(t)$) such that $a_{t+1}=\underline{a}_t=a_t^{\text{opt}}$, for all current sub-optimal arm $a'\notin\widehat{\mathcal{A}}^\star(t)$,*

$$|\widehat{\mu}_{a'}(t)-\mu_{a'}|\leqslant\varepsilon+\widehat{\varepsilon}^\star_{a'}(t)$$

and

$$\widehat{\varepsilon}^\star_{a'}(t)\leqslant\varepsilon_{a'}(t)\,,$$

where

$$\widehat{\varepsilon}^\star_{a'}(t)\leqslant\frac{5}{3}\cdot\frac{\gamma_t}{1-\gamma_t}\cdot\Delta_{a'}$$

$$\varepsilon_{a'}(t)=\frac{5}{3}\cdot\frac{\gamma_t}{1-\gamma_t}\cdot\Delta_{a'}\leqslant\frac{\Delta_{a'}}{3}\,.$$

Proof. Let us consider $t \notin \mathcal{U}(\varepsilon)$ such that $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$ and $a_{t+1} = \underline{a}_t = a_t^{\text{opt}}$. Let us consider a current sub-optimal arm $a' \notin \widehat{\mathcal{A}}^\star(t)$. Then $t \notin \mathcal{M}(\varepsilon)$, $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$ and Equation (4.68) that defines $\mathcal{M}(\varepsilon)$ implies

$$|\widehat{\mu}_{a'}(t) - \mu_{a'}| < \varepsilon + \widehat{\varepsilon}^\star_{a'}(t) \,. \tag{4.99}$$

Since $t \notin \mathcal{U}(\varepsilon)$, from Lemma 11 we have

$$|\widehat{\mu}^\star(t) - \mu^\star| < \varepsilon \,. \tag{4.100}$$

We have by definition

$$\widehat{\varepsilon}^\star_{a'}(t) = \gamma_t \cdot (\widehat{\mu}^\star(t) - \widehat{\mu}_{a'}(t)) \ . \tag{4.101}$$

By combining Equations (4.99), (4.100) and (4.101), we have

$$\widehat{\varepsilon}^\star_{a'}(t) \leqslant \gamma_t \cdot (\mu^\star - \mu_{a'} + 2\varepsilon + \widehat{\varepsilon}^\star_{a'}(t)) = \gamma_t \cdot (\Delta_{a'} + 2\varepsilon + \widehat{\varepsilon}^\star_{a'}(t)) \ . \tag{4.102}$$

Equation (4.102) implies

$$\widehat{\varepsilon}^\star_{a'}(t) \leqslant \frac{\gamma_t}{1-\gamma_t} \cdot (\Delta_{a'} + 2\varepsilon) \ . \tag{4.103}$$

Since $0 < \varepsilon < \varepsilon_0 = \delta_{\min}/3$, Equation (4.103) implies

$$\widehat{\varepsilon}^\star_{a'}(t) \leqslant \frac{5}{3} \cdot \frac{\gamma_t}{1-\gamma_t} \cdot \Delta_{a'} \,. \tag{4.104}$$

Finally we note that γ_t is set so that

$$\frac{\gamma_t}{1-\gamma_t} < 1/5 \,. \tag{4.105}$$

□

Lemma 13 (Reliable current informative sets of arms)**.** *For accuracy $0 < \varepsilon < \varepsilon_0 = \delta_{\min}/3$, for all time step $t \notin \mathcal{U}(\varepsilon)$ that corresponds to an exploration phase, that is $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$, such that $a_{t+1} = \underline{a}_t = a_t^{\text{opt}}$, for all current sub-optimal arm $a \notin \widehat{\mathcal{A}}^\star(t)$,*

$$\{a\} \subset \mathcal{A}_a(\varepsilon, t) \subset \widehat{\mathcal{A}}_a(t) \,,$$

where

$$\mathcal{A}_a(\varepsilon, t) := \left\{ a' \notin \widehat{\mathcal{A}}^\star(t) : \ 2\varepsilon + \varepsilon_{a'}(t) \leqslant \delta_{a,a'}(\theta^\star) \right\} .$$

Proof. Since $t \notin \mathcal{U}(\varepsilon)$, $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$ and $a_{t+1} = \underline{a}_t = a_t^{\text{opt}}$, from Lemma 12 we have

$$\varepsilon_a(t) \leqslant \frac{\Delta_a}{3} \,. \tag{4.106}$$

Since $\varepsilon < \varepsilon_0 \leqslant \Delta_a/3$, from Equation (4.106) we have

$$2\varepsilon + \varepsilon_a(t) < \Delta_a \,. \tag{4.107}$$

Since $\delta_{a,a}(\theta^\star) = \Delta_a$, Equation (4.107) implies $a \in \mathcal{A}_a(\varepsilon, t) \neq \emptyset$.

Let us consider $a' \in \mathcal{A}_a(\varepsilon, t)$. Then we have

$$2\varepsilon + \varepsilon_{a'}(t) < \delta_{a,a'}(\theta) = \mu^\star - \mu_{a'} - \theta^\star_{a,a'}(t) \,, \tag{4.108}$$

that is

$$\mu_{a'} + \varepsilon + \varepsilon_{a'}(t) < \mu^\star - \varepsilon - \theta^\star_{a,a'}(t) \,. \tag{4.109}$$

Since $t \notin \mathcal{U}(\varepsilon)$, from Lemma 11 we have

$$\begin{aligned} \theta^\star &= \theta^{(a^\star)} = \widehat{\theta}^\star(t) \\ \mu^\star - \varepsilon &< \widehat{\mu}^\star(t)\,. \end{aligned} \tag{4.110}$$

Since $t \notin \mathcal{U}(\varepsilon)$, $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$ and $a_{t+1} = \underline{a}_t = a_t^{\text{opt}}$. From Lemma 12 then we have

$$\mu_{a'} + \varepsilon + \varepsilon_{a'}(t) > \widehat{\mu}_{a'}(t)\,. \tag{4.111}$$

By combining Equations (4.109), (4.110) and (4.111), we have

$$\widehat{\mu}_{a'}(t) \leqslant \widehat{\mu}^\star(t) - \widehat{\theta}^\star_{a,a'}(t)\,, \tag{4.112}$$

that is $a' \in \widehat{\mathcal{A}}_a(t)$. □

4.6 Upper bounds under `IMED-GS` algorithm

In this section, we now provide the final results related to the control of the numbers of pulls of sub-optimal arms. We start by providing in Section 4.6.1 an upper-bound on the number of pulls $N_a^{\text{opt}}(t)$ solution to the empirical optimization problem. Then, in Section 4.6.2, we provide an almost sure upper bound on the number of pulls of sub-optimal arms, making appear the size of the random set of times $\mathcal{U}(\varepsilon)$ introduced in the previous section. This is further used in Section 4.6.3 to provide a fully explicit, finite-time upper bound on the regret valid almost surely. Last, in Section 4.6.4, we handle the size of the random set thanks to the results of Section 4.5 (in particular, Lemma 8), and derive the main regret bound.

4.6.1 Upper bounds on the optimal numbers of pulls

Lemma 14 (Upper bounds on optimal numbers of pulls)**.** *For accuracy $0 < \varepsilon < \varepsilon_0 = \delta_{\min}/3$, for all time step $t \notin \mathcal{U}(\varepsilon)$ that corresponds to an exploration phase, that is $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$, such that $a_{t+1} = \underline{a}_t = a_t^{\text{opt}}$,*

$$\mathfrak{C}_{\{\widehat{\theta}^\star(t)\}}\big(\widehat{\mu}(t)\big) \leqslant \max_{a \notin \mathcal{A}^\star} \frac{\Delta_a + 2\varepsilon + \varepsilon_a(t)}{\Delta_a - 2\varepsilon - \varepsilon_a(t)} \cdot \mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,t)\big)\,,$$

$$\sum_{a \notin \mathcal{A}^\star} N_a^{\text{opt}}(t)\,\Delta_a \leqslant \max_{a \notin \mathcal{A}^\star} \left(\frac{\Delta_a + 2\varepsilon + \varepsilon_a(t)}{\Delta_a - 2\varepsilon - \varepsilon_a(t)}\right)^2 \cdot\ \mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,t)\big) \cdot \mathrm{f}_\xi(t)\,,$$

where for all $a \in \mathcal{A}$,

$$\mu_a(\varepsilon,t) = \begin{cases} \mu_a + \varepsilon + \varepsilon_a(t) & \text{, if } a \notin \mathcal{A}^\star \\ \mu_a - \varepsilon & \text{, if } a \in \mathcal{A}^\star\,. \end{cases}$$

Proof. Let us consider $t \notin \mathcal{U}(\varepsilon)$ such that $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$, $a_{t+1} = \underline{a}_t = a_t^{\text{opt}}$ and $n \in \mathbb{R}_+^{\mathcal{A}}$ such that

$$\forall a \notin \operatorname{argmax}\big(\mu(\varepsilon,t)\big), \quad \sum_{\substack{a' \in \mathcal{A} \\ \mu_{a'}(\varepsilon,t) \leqslant \max((\mu(\varepsilon,t)) - \theta^\star_{a,a'}}} \mathrm{kl}\big(\mu_{a'}(\varepsilon,t) \big| \max(\mu(\varepsilon,t)) - \theta^\star_{a,a'}\big)\, n_{a'} \geqslant 1\,. \tag{4.113}$$

Since $t \notin \mathcal{U}(\varepsilon)$, from Lemma 11 and Assumptions 7-6 we have

$$\begin{aligned} \widehat{\mathcal{A}}^\star(t) &= \mathcal{A}^\star = \{a^\star\} \\ \theta^\star &= \theta^{(a^\star)} = \widehat{\theta}^\star(t) \end{aligned} \tag{4.114}$$

$$|\widehat{\mu}^\star(t) - \mu^\star| < \varepsilon\,. \tag{4.115}$$

Since $t \notin \mathcal{U}(\varepsilon)$ is such that $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$ and $a_{t+1} = \underline{a}_t = a_t^{\text{opt}}$, from Lemma 12 and Equation (4.114) we have

$$|\widehat{\mu}_a(t) - \mu_a| \leqslant \varepsilon + \varepsilon_a(t)\,, \quad \forall a \notin \mathcal{A}^\star\,, \tag{4.116}$$

$$\varepsilon_a(t) \leqslant \frac{\Delta_a}{3}\,, \quad \forall a \notin \mathcal{A}^\star\,. \tag{4.117}$$

From Equations (4.114), (4.115) and (4.117) we have

$$\operatorname{argmax}\big(\mu(\varepsilon,t)\big) = \mathcal{A}^\star = \{a^\star\}\,. \tag{4.118}$$

Since $t \notin \mathcal{U}(\varepsilon)$ is such that $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$ and $a_{t+1} = \underline{a}_t = a_t^{\text{opt}}$, from Lemma 13 and Equation (4.114) we have for all sub-optimal arm $a \notin \widehat{\mathcal{A}}^\star(t)$,

$$\{a\} \subset \mathcal{A}_a(\varepsilon,t) \subset \widehat{\mathcal{A}}_a(t)\,, \tag{4.119}$$

where

$$\begin{aligned} \mathcal{A}_a(\varepsilon,t) &= \{a' \notin \mathcal{A}^\star :\ 2\varepsilon + \varepsilon_{a'}(t) \leqslant \delta_{a,a'}(\theta^\star)\} \\ &= \Big\{a' \notin \mathcal{A}^\star :\ \mu_{a'}(\varepsilon,t) \leqslant \max\left((\mu(\varepsilon,t)) - \widehat{\theta}^\star_{a,a'}(t)\right\}\,. \end{aligned} \tag{4.120}$$

Then, Equation (4.113) rewrites

$$\forall a \notin \mathcal{A}^\star, \quad \sum_{a' \in \mathcal{A}_a(\varepsilon,t)} \mathrm{kl}\big(\mu_{a'} + \varepsilon + \varepsilon_{a'}(t) \big| \mu^\star - \varepsilon - \widehat{\theta}^\star_{a,a'}(t)\big)\, n_{a'} \geqslant 1\,. \tag{4.121}$$

From Equations (4.114), (4.116) and standard monotonic properties of $\mathrm{kl}(\cdot\,|\cdot)$, we have

$$\begin{aligned} &\forall a \notin \mathcal{A}^\star,\ \forall a' \in \mathcal{A}_a(\varepsilon,t)\,, \\ &\qquad \mathrm{kl}\big(\mu_{a'} + \varepsilon + \varepsilon_{a'}(t) \big| \mu^\star - \varepsilon - \widehat{\theta}^\star_{a,a'}(t)\big) \leqslant \mathrm{kl}\Big(\widehat{\mu}_{a'}(t) \Big| \widehat{\mu}^\star(t) - \widehat{\theta}^\star_{a,a'}(t)\Big)\,. \end{aligned} \tag{4.122}$$

By combining Equations (4.121) and (4.122) we have

$$\forall a \notin \mathcal{A}^\star, \quad \sum_{a' \in \mathcal{A}_a(\varepsilon,t)} \mathrm{kl}\Big(\widehat{\mu}_{a'}(t) \Big| \widehat{\mu}^\star(t) - \widehat{\theta}^\star_{a,a'}(t)\Big)\, n_{a'} \geqslant 1\,. \tag{4.123}$$

Then, by combining Equations (4.119) and (4.123) we have

$$\forall a \notin \mathcal{A}^\star, \quad \sum_{a' \in \widehat{\mathcal{A}}_a(t)} \mathrm{kl}\Big(\widehat{\mu}_{a'}(t) \Big| \widehat{\mu}^\star(t) - \widehat{\theta}^\star_{a,a'}(t)\Big)\, n_{a'} \geqslant 1\,. \tag{4.124}$$

Since

$$\begin{aligned} \mathfrak{C}_{\{\widehat{\theta}^\star(t)\}}\big(\widehat{\mu}(t)\big) = \min_{n \in \mathbb{R}_+^{\mathcal{A}}} &\sum_{a \in \mathcal{A}} n_a \left(\widehat{\mu}^\star(t) - \widehat{\mu}_a(t)\right) \\ s.t. \quad &\forall a \notin \widehat{\mathcal{A}}^\star(t) = \mathcal{A}^\star, \quad \sum_{a' \in \widehat{\mathcal{A}}_a(t)} \mathrm{kl}\Big(\widehat{\mu}_{a'}(t) \Big| \widehat{\mu}^\star(t) - \widehat{\theta}^\star_{a,a'}(t)\Big)\, n_{a'} \geqslant 1 \end{aligned}\,, \tag{4.125}$$

Equation (4.124) implies

$$\mathfrak{C}_{\{\widehat{\theta}^{\star}(t)\}}\big(\widehat{\mu}(t)\big) \leqslant \sum_{a\in\mathcal{A}} n_a \left(\widehat{\mu}^{\star}(t) - \widehat{\mu}_a(t)\right). \tag{4.126}$$

By combining Equations (4.114), (4.115), (4.116), (4.117) and (4.126) we have

$$\mathfrak{C}_{\{\widehat{\theta}^{\star}(t)\}}\big(\widehat{\mu}(t)\big) \leqslant \sum_{a\notin\mathcal{A}^{\star}} n_a \left(\mu^{\star} - \mu_a + 2\varepsilon + \varepsilon_a(t)\right) = \sum_{a\notin\mathcal{A}^{\star}} n_a \left(\Delta_a + 2\varepsilon + \varepsilon_a(t)\right). \tag{4.127}$$

Then Equations (4.118) and (4.127) imply

$$\mathfrak{C}_{\{\widehat{\theta}^{\star}(t)\}}\big(\widehat{\mu}(t)\big) \leqslant \max_{a\notin\mathcal{A}^{\star}} \frac{\Delta_a + 2\varepsilon + \varepsilon_a(t)}{\Delta_a - 2\varepsilon - \varepsilon_a(t)} \cdot \sum_{a\in\mathcal{A}} n_a \left(\max(\mu(\varepsilon,t)) - \mu_a(\varepsilon,t)\right). \tag{4.128}$$

Since previous Equation (4.128) is satisfied for all $n\in\mathbb{R}_{+}^{\mathcal{A}}$ satisfying Equation (4.113), we have

$$\mathfrak{C}_{\{\widehat{\theta}^{\star}(t)\}}\big(\widehat{\mu}(t)\big) \leqslant \max_{a\notin\mathcal{A}^{\star}} \frac{\Delta_a + 2\varepsilon + \varepsilon_a(t)}{\Delta_a - 2\varepsilon - \varepsilon_a(t)} \cdot \mathfrak{C}_{\{\theta^{\star}\}}\big(\mu(\varepsilon,t)\big). \tag{4.129}$$

Furthermore, by combining Equations (4.114), (4.115), (4.116) and (4.117) we have

$$\begin{aligned}
\sum_{a\notin\mathcal{A}^{\star}} N_a^{\mathrm{opt}}(t)\,\Delta_a &\leqslant \max\nolimits_{a\notin\mathcal{A}^{\star}} \left(\frac{\Delta_a}{\Delta_a - 2\varepsilon - \varepsilon_a(t)}\right) \sum_{a\notin\widehat{\mathcal{A}}^{\star}(t)} N_a^{\mathrm{opt}}(t)\,\widehat{\Delta}_a(t) \\
&\leqslant \max\nolimits_{a\notin\mathcal{A}^{\star}} \left(\frac{\Delta_a + 2\varepsilon + \varepsilon_a(t)}{\Delta_a - 2\varepsilon - \varepsilon_a(t)}\right) \sum_{a\notin\widehat{\mathcal{A}}^{\star}(t)} N_a^{\mathrm{opt}}(t)\,\widehat{\Delta}_a(t).
\end{aligned} \tag{4.130}$$

Since $\big(N_a^{\mathrm{opt}}(t)\big)_{a\in\mathcal{A}} = \big(\overline{I}_{\bar{a}_t}(t)\, n_a^{\mathrm{opt}}(t)\big)_{a\in\mathcal{A}}$, we have

$$\sum_{a\notin\widehat{\mathcal{A}}^{\star}(t)} N_a^{\mathrm{opt}}(t)\,\widehat{\Delta}_a(t) = \min_{a\in\mathcal{A}} \overline{I}_a(t) \sum_{a\notin\widehat{\mathcal{A}}^{\star}(t)} n_a^{\mathrm{opt}}(t)\,\widehat{\Delta}_a(t). \tag{4.131}$$

From Equation (4.9) that defines indexes $\big(\overline{I}_a(t)\big)_{a\in\mathcal{A}}$ and previous Equation (4.131) we have

$$\sum_{a\notin\widehat{\mathcal{A}}^{\star}(t)} N_a^{\mathrm{opt}}(t)\,\widehat{\Delta}_a(t) \leqslant \mathrm{f}_{\xi}(t) \sum_{a\notin\widehat{\mathcal{A}}^{\star}(t)} n_a^{\mathrm{opt}}(t)\,\widehat{\Delta}_a(t). \tag{4.132}$$

From Equation (4.11) that defines $\big(n_a^{\mathrm{opt}}(t)\big)_{a\in\mathcal{A}}$ we have

$$\sum_{a\notin\widehat{\mathcal{A}}^{\star}(t)} n_a^{\mathrm{opt}}(t)\,\widehat{\Delta}_a(t) = \mathfrak{C}_{\{\widehat{\theta}^{\star}(t)\}}\big(\widehat{\mu}(t)\big). \tag{4.133}$$

By combining Equations (4.132) and (4.133) we have

$$\sum_{a\notin\widehat{\mathcal{A}}^{\star}(t)} N_a^{\mathrm{opt}}(t)\,\widehat{\Delta}_a(t) \leqslant \mathfrak{C}_{\{\widehat{\theta}^{\star}(t)\}}\big(\widehat{\mu}(t)\big) \cdot \mathrm{f}_{\xi}(t). \tag{4.134}$$

Combining Equations (4.129), (4.130) and (4.134) ends the proof. □

4.6.2 Upper bounds on the numbers of pulls

We recall that what we call optimal numbers of pulls at time step $t \geqslant 1$ are the numbers of pulls $(N_a^{\mathrm{opt}}(t))_{a\in\mathcal{A}}$ solution of the optimisation problem.

Lemma 15 (Upper bounds on current pulled sub-optimal arm)**.** *For accuracy $0<\varepsilon<\varepsilon_0$, for all sub-optimal $a\notin\mathcal{A}^\star$, for all time step $t\notin\mathcal{U}(\varepsilon)$ such that $a_{t+1}=\underline{a}_t=a_t^{\mathrm{opt}}=a$,*

$$N_a(t) \leqslant N_a^{\mathrm{opt}}(t) \leqslant \frac{\lambda_t(\mu,\varepsilon)\mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,t)\big)}{\Delta_a}\cdot \mathrm{f}_\xi(t)$$

where

$$\lambda_t(\mu,\varepsilon) \coloneqq \max_{a'\notin\mathcal{A}^\star}\left(\frac{\Delta_{a'}+2\varepsilon+\varepsilon_{a'}(t)}{\Delta_{a'}-2\varepsilon-\varepsilon_{a'}(t)}\right)^2 .$$

For accuracy $0<\varepsilon<\varepsilon_0$, for all sub-optimal $a\notin\mathcal{A}^\star$, for all time step $t\notin\mathcal{U}(\varepsilon)$ such that $a_{t+1}=a\neq\underline{a}_t$,

$$N_a(t) \leqslant \frac{\mathrm{f}_\xi(t)}{\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)} .$$

For accuracy $0<\varepsilon<\varepsilon_0$, for all sub-optimal $a\notin\mathcal{A}^\star$,

$$\forall t\geqslant 1,\quad N_a(t) \leqslant \max_{1\leqslant s\leqslant t}\left\{\frac{\lambda_t(\mu,\varepsilon)\mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,s)\big)}{\Delta_a}, \frac{1}{\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)}\right\}\cdot \mathrm{f}_\xi(t) + |\mathcal{U}(\varepsilon)| + 1 .$$

For accuracy $0<\varepsilon<\varepsilon_0\vee(1-\mu^\star)$, for all sub-optimal $a\notin\mathcal{A}^\star$, for all time step $t\notin\mathcal{U}(\varepsilon)$ such that $a_{t+1}=a\neq\underline{a}_t$,

$$N_a(t) \leqslant \frac{\Gamma\left[\mathrm{f}_\xi\left(\max_{1\leqslant s\leqslant t}\left\{\frac{\lambda_t(\mu,\varepsilon)\mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,s)\big)}{\Delta_a}, \frac{1}{\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)}\right\}\mathrm{f}_\xi(t)+|\mathcal{U}(\varepsilon)|+1\right)+1\right]}{(\gamma_t)^2(\mu^\star-\varepsilon)(1-\mu^\star-\varepsilon)\,\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)} .$$

Proof. Let us consider a time step $t\notin\mathcal{U}(\varepsilon)$ such that $a_{t+1}=a$.

Since $t\notin\mathcal{U}(\varepsilon)$, from Lemma 11 and Assumption 7 we have

$$\widehat{\mathcal{A}}^\star(t) = \mathcal{A}^\star = \{a^\star\} . \tag{4.135}$$

In particular, we have

$$a_{t+1} = a \notin \widehat{\mathcal{A}}^\star(t) . \tag{4.136}$$

and Algorithm 9 implies

$$\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t) . \tag{4.137}$$

Since $0<\varepsilon<(\mu^\star-\mu_a)/3$, $t\notin\mathcal{U}(\varepsilon)$ and $a_{t+1}=a$, Lemma 12, Lemma 11 and Equation (4.135) imply

$$\widehat{\mu}_a(t) \leqslant \mu_a+\varepsilon < \mu^\star-\varepsilon \leqslant \widehat{\mu}^\star(t) \leqslant \mu^\star+\varepsilon . \tag{4.138}$$

Let us consider a time step $t\notin\mathcal{U}(\varepsilon)$ such that $a_{t+1}=\underline{a}_t=a_t^{\mathrm{opt}}=a$. Then, Lemma 7 implies

$$N_a(t) \leqslant N_a^{\mathrm{opt}}(t) . \tag{4.139}$$

Since $t \notin \mathcal{U}(\varepsilon)$, we have $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t) = \mathcal{A}^\star$ as shown in Equation (4.137) and $a_{t+1} = \underline{a}_t = a_t^{\text{opt}}$. Lemma 14 then implies

$$\forall a' \notin \widehat{\mathcal{A}}^\star(t) = \mathcal{A}^\star, \quad N_{a'}^{\text{opt}}(t) \leqslant \frac{\lambda_t(\mu,\varepsilon)\mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,t)\big)}{\Delta_{a'}} \cdot \mathrm{f}_\xi(t)\,. \tag{4.140}$$

By combining Equations (4.139) and (4.140), we get

$$N_a(t) \leqslant N_a^{\text{opt}}(t) \leqslant \frac{\lambda_t(\mu,\varepsilon)\mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,t)\big)}{\Delta_a} \cdot \mathrm{f}_\xi(t)\,.$$

Let us consider a time step $t \notin \mathcal{U}(\varepsilon)$ such that $a_{t+1} = a \neq \underline{a}_t$. From Equation (4.137), we have $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$. Then Algorithm 9 implies

$$a_{t+1} \in \{\overline{a}_t, \dot{a}_t\}\,. \tag{4.141}$$

By combining previous Equation (4.141) and Lemma 7, we have

$$N_a(t) \leqslant \frac{\mathrm{f}_\xi(t)}{\mathrm{kl}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t))}\,. \tag{4.142}$$

From Equations (4.138) and (4.142), the classical monotonic properties of $\mathrm{kl}(\cdot\,|\cdot)$ imply

$$N_a(t) \leqslant \frac{\mathrm{f}_\xi(t)}{\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)}\,.$$

Let us consider a time step $t \geqslant 1$. We write the number of pulls $N_a(t)$ as follows

$$\begin{aligned} N_a(t) &= 1 + \sum_{s\geqslant|\mathcal{A}|}^{t-1} \mathbb{I}_{\{a_{t+1}=a\}} \\ &= 1 + \sum_{s\geqslant|\mathcal{A}|}^{t-1} \mathbb{I}_{\{a_{s+1}=a,\ s\in\mathcal{U}(\varepsilon)\}} + \sum_{s\geqslant|\mathcal{A}|}^{t-1} \mathbb{I}_{\left\{a_{s+1}=\underline{a}_s=a_s^{\text{opt}}=a,\ s\notin\mathcal{U}(\varepsilon)\right\}} + \mathbb{I}_{\{a_{s+1}=a\neq\underline{a}_s,\ s\notin\mathcal{U}(\varepsilon)\}}. \end{aligned} \tag{4.143}$$

We first note that

$$\sum_{s\geqslant|\mathcal{A}|}^{t-1} \mathbb{I}_{\{a_{s+1}=a,\ s\in\mathcal{U}(\varepsilon)\}} \leqslant |\mathcal{U}(\varepsilon)|\,. \tag{4.144}$$

Then, we denote by

$$\tau_a(t) = \max\left\{s \in [\![|\mathcal{A}|+1\,;t]\!] :\ (a_{s+1} = \underline{a}_s = a_s^{\text{opt}} = a \text{ or } a_{s+1} = a \neq \underline{a}_s) \quad \text{and} \quad s \notin \mathcal{U}(\varepsilon)\right\} \tag{4.145}$$

the last time step before time step t that does not belong to $\mathcal{U}(\varepsilon)$ such that we pull arm a. Then, we have

$$N_a(\tau_a(t)) \leqslant \max_{1\leqslant s\leqslant t} \left\{ \frac{\lambda_t(\mu,\varepsilon)\mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,s)\big)}{\Delta_a}\,, \frac{1}{\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)} \right\} \times \mathrm{f}_\xi(t)\,. \tag{4.146}$$

Furthermore, we note that

$$\begin{aligned} &\sum_{s\geqslant|\mathcal{A}|}^{t-1} \mathbb{I}_{\left\{a_{s+1}=\underline{a}_s=a_s^{\text{opt}}=a,\ s\notin\mathcal{U}(\varepsilon)\right\}} + \mathbb{I}_{\{a_{s+1}=a\neq\underline{a}_s,\ s\notin\mathcal{U}(\varepsilon)\}} \\ \leqslant\ & N_a(\tau_a(t)) \\ \leqslant\ & \max_{1\leqslant s\leqslant t} \left\{ \frac{\lambda_t(\mu,\varepsilon)\mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,s)\big)}{\Delta_a}\,, \frac{1}{\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)} \right\} \times \mathrm{f}_\xi(t)\,. \end{aligned} \tag{4.147}$$

By combining Equations (4.143), (4.144) and (4.147), we get

$$\forall a \neq a^\star,\ \forall t \geqslant 1,\quad N_a(t) \leqslant 1+|\mathcal{U}(\varepsilon)|+\max_{1\leqslant s\leqslant t}\left\{\frac{\lambda_t(\mu,\varepsilon)\mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,s)\big)}{\Delta_a}\,,\frac{1}{\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)}\right\}\times \mathrm{f}_\xi(t)\,. \tag{4.148}$$

Let us consider a time step $t \notin \mathcal{U}(\varepsilon)$ such that $a_{t+1} = a \neq \underline{a}_t$. From Equations (4.135) and (4.137), we have $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$. Then Algorithm 9 implies

$$\overline{a}_t, a_t^{\mathrm{opt}} \notin \widehat{\mathcal{A}}^\star(t) = \{a^\star\}\,,\quad a_{t+1} \in \{\overline{a}_t, \dot{a}_t\}\,. \tag{4.149}$$

By combining previous Equation (4.149) and Lemma 7, we have

$$N_a(t) \leqslant \left(\mathbb{I}_{\left\{\overline{a}_t\notin\widehat{\mathcal{A}}^\star(t)\text{ and }\underline{a}_t\neq a_t^{\mathrm{opt}}\right\}}\frac{\overline{I}_{\overline{a}_t}(t)}{I_{\dot{a}_t}(t)}\right)\frac{\mathrm{f}_\xi\Big(N_{a_t^{\mathrm{opt}}}(t)\Big)+1}{(\gamma_t)^2\widehat{\mu}^\star(t)(1-\widehat{\mu}^\star(t))\,\mathrm{kl}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t))}\,. \tag{4.150}$$

From Algorithm 9, we have

$$\mathbb{I}_{\left\{\overline{a}_t\notin\widehat{\mathcal{A}}^\star(t)\text{ and }\underline{a}_t\neq a_t^{\mathrm{opt}}\right\}}\frac{\overline{I}_{\overline{a}_t}(t)}{I_{\dot{a}_t}(t)} \leqslant \Gamma\,. \tag{4.151}$$

From Equations (4.138) and (4.142) and the classical monotonic properties of $\mathrm{kl}(\cdot\,|\,\cdot)$, we have

$$\widehat{\mu}^\star(t)(1-\widehat{\mu}^\star(t))\,\mathrm{kl}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) \geqslant (\mu^\star-\varepsilon)(1-\mu^\star-\varepsilon)\,\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)\,. \tag{4.152}$$

By combining Equations (4.148), (4.150), (4.151) and (4.152), we then get

$$N_a(t) \leqslant \frac{\Gamma\left[\mathrm{f}_\xi\left(1+|\mathcal{U}(\varepsilon)|+\max_{1\leqslant s\leqslant t}\left\{\frac{\lambda_t(\mu,\varepsilon)\mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,s)\big)}{\Delta_a}\,,\frac{1}{\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)}\right\}\mathrm{f}_\xi(t)\right)+1\right]}{(\gamma_t)^2(\mu^\star-\varepsilon)(1-\mu^\star-\varepsilon)\,\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)}\,. \tag{4.153}$$

□

Remark 14. *Proof of Lemma 15 highlights that for accuracy $0<\varepsilon<\varepsilon_0$, for all time step $t\notin\mathcal{U}(\varepsilon)$, if $a_{t+1}\notin\mathcal{A}^\star$ then $\overline{a}_t \notin \widehat{\mathcal{A}}^\star(t)$.*

4.6.3 Almost-sure upper bound on the cumulative regret

Proposition 4 (Upper bound on the non-averaged cumulative regret)**.** *For accuracy $0<\varepsilon<\varepsilon_0$, for all time horizon $T\geqslant|\mathcal{A}|$,*

$$\begin{aligned}\textstyle\sum_{a\notin\mathcal{A}^\star} N_a(T)\,\Delta_a \;\leqslant\;& \inf_{\tau\geqslant 1}\left\{\Lambda_\tau(\mu,\varepsilon)\cdot C_\tau(\mu,\varepsilon)\cdot \mathrm{f}_\xi(T)+\Sigma_\Delta\cdot\,\tau\right\}\\ &+\mathrm{G}_T(\mu,\varepsilon)\cdot\mathrm{F}_\xi(\mu,\varepsilon,T)+\mathrm{G}_T(\mu,\varepsilon)\cdot\mathrm{f}_\xi(|\mathcal{U}(\varepsilon)|)+|\mathcal{U}(\varepsilon)|\end{aligned}$$

where

$$\Lambda_\tau(\mu,\varepsilon)=\sup_{t\geqslant\tau}\lambda_t(\mu,\varepsilon)\qquad C_\tau(\mu,\varepsilon)=\sup_{t\geqslant\tau}\mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,t)\big)$$

$$\lambda_t(\mu,\varepsilon)=\max_{a'\notin\mathcal{A}^\star}\left(\frac{\Delta_{a'}+2\varepsilon+\varepsilon_{a'}(t)}{\Delta_{a'}-2\varepsilon-\varepsilon_{a'}(t)}\right)^2$$

$$\mathrm{F}_{\xi}(\mu,\varepsilon,T) = \max_{a\notin\mathcal{A}^\star} \mathrm{f}_{\xi}\left(\max_{t\geqslant 1}\left\{\frac{\lambda_t(\mu,\varepsilon)\mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,t)\big)}{\Delta_a}\,,\frac{1}{\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)}\right\}\mathrm{f}_{\xi}(T)+1\right)+1$$

$$\mathrm{G}_T(\mu,\varepsilon) = \sum_{a\notin\mathcal{A}^\star}\frac{\Delta_a\,\Gamma}{(\gamma_{_T})^2(\mu^\star-\varepsilon)(1-\mu^\star-\varepsilon)\,\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)}$$

$$\Sigma_\Delta = \sum_{a\in\mathcal{A}}\Delta_a\,.$$

Proof. Let us consider $\tau\geqslant 1$. For $a\notin\mathcal{A}^\star$, let us consider

$$\overline{\tau}_a(\varepsilon,T) = \max\left\{t\in[\![1,T-1]\!]:\ t\notin\mathcal{U}(\varepsilon),\ a_{t+1}=\underline{a}_t=a_t^{\mathrm{opt}}=a\right\} \tag{4.154}$$

$$\tau_a(\varepsilon,T) = \max\left\{t\in[\![1,T-1]\!]:\ t\notin\mathcal{U}(\varepsilon),\ a_{t+1}=a\neq\underline{a}_t\right\}. \tag{4.155}$$

Then, the number of pulls of sub-optimal arm $a\notin\mathcal{A}^\star$ satisfies

$$\begin{aligned}N_a(T) &\leqslant 1+\sum_{t=|\mathcal{A}|}^{T-1}\mathbb{I}_{\{t\notin\mathcal{U}(\varepsilon),\ a_{t+1}=a\neq\underline{a}_t\}}+\sum_{t=|\mathcal{A}|}^{T-1}\mathbb{I}_{\{t\notin\mathcal{U}(\varepsilon),\ a_{t+1}=a\neq\underline{a}_t\}}+\sum_{t\geqslant 1}\mathbb{I}_{\{t\in\mathcal{U}(\varepsilon),\ a_{t+1}=a\}}\\ &\leqslant N_a(\overline{\tau}_a(\varepsilon,T))+N_a(\tau_a(\varepsilon,T))+\sum_{t\geqslant 1}\mathbb{I}_{\{t\in\mathcal{U}(\varepsilon),\ a_{t+1}=a\}}\,.\end{aligned} \tag{4.156}$$

In particular, previous Equation (4.156) implies

$$N_a(T)\leqslant\mathbb{I}_{\{\tau\leqslant\overline{\tau}_a(\varepsilon,T)\}}\cdot N_a(\overline{\tau}_a(\varepsilon,T))+\tau+N_a(\tau_a(\varepsilon,T))+\sum_{t\geqslant 1}\mathbb{I}_{\{a_{t+1}=a,\ t\in\mathcal{U}(\varepsilon)\}}\,. \tag{4.157}$$

Since $\overline{\tau}_a(\varepsilon,T)\notin\mathcal{U}(\varepsilon)$ and $a_{\overline{\tau}_a(\varepsilon,T)+1}=\underline{a}_{\overline{\tau}_a(\varepsilon,T)}=a^{\mathrm{opt}}_{\overline{\tau}_a(\varepsilon,T)}=a$, Lemma 15 implies

$$N_a(\overline{\tau}_a(\varepsilon,T))\leqslant N_a^{\mathrm{opt}}(\overline{\tau}_a(\varepsilon,T))\,. \tag{4.158}$$

Since $\overline{\tau}_a(\varepsilon,T)\notin\mathcal{U}(\varepsilon)$ and $a_{\overline{\tau}_a(\varepsilon,T)+1}=a\notin\mathcal{A}^\star$, Remark 14 implies

$$\overline{\tau}_a(\varepsilon,T)\leqslant\overline{\tau}(\varepsilon,T):=\max\left\{t\in[\![1,T-1]\!]:\ t\notin\mathcal{U}(\varepsilon),\ a_{t+1}=\underline{a}_t=a_t^{\mathrm{opt}}\notin\widehat{\mathcal{A}}^\star(t)\right\}. \tag{4.159}$$

By combining Equations (4.158) and (4.159) we get

$$\overline{\tau}_a(\varepsilon,T)\leqslant\overline{\tau}(\varepsilon,T)\qquad\qquad N_a(\overline{\tau}_a(\varepsilon,T))\leqslant N_a^{\mathrm{opt}}(\overline{\tau}(\varepsilon,T))\,. \tag{4.160}$$

Since $\overline{\tau}(\varepsilon,T)\notin\mathcal{U}(\varepsilon)$ and $\overline{a}_{\overline{\tau}(\varepsilon,T)}\notin\widehat{\mathcal{A}}^\star(\overline{\tau}(\varepsilon,T)))$, from Lemma 14 we have

$$\mathbb{I}_{\{\tau\leqslant\overline{\tau}(\varepsilon,T)\leqslant T\}}\sum_{a\notin\mathcal{A}^\star}N_a^{\mathrm{opt}}(\overline{\tau}(\varepsilon,T))\ \Delta_a\leqslant\Lambda_\tau(\mu,\varepsilon)\cdot C_\tau(\mu,\varepsilon)\cdot\mathrm{f}_{\xi}(T)\,. \tag{4.161}$$

Since $\tau_a(\varepsilon,T)\notin\mathcal{U}(\varepsilon)$ and $a_{\tau_a(\varepsilon,T)+1}=a\neq\underline{a}_{\tau_a(\varepsilon,T)}$, Lemma 15 plus the monotony of $(\gamma_t)_{t\geqslant 1}$,$(\mathrm{f}_{\xi}(t))_{t\geqslant 1}$ imply

$$\begin{aligned}&N_a(\tau_a(\varepsilon,T))\\ \leqslant\ &\frac{\Gamma\left[\mathrm{f}_{\xi}\left(\max_{t\geqslant 1}\left\{\frac{\lambda_t(\mu,\varepsilon)\mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,t)\big)}{\Delta_a}\,,\frac{1}{\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)}\right\}\mathrm{f}_{\xi}(T)+|\mathcal{U}(\varepsilon)|+1\right)+1\right]}{(\gamma_{_T})^2(\mu^\star-\varepsilon)(1-\mu^\star-\varepsilon)\,\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)}\\ \leqslant\ &\frac{\Gamma\left[\mathrm{f}_{\xi}\left(\max_{t\geqslant 1}\left\{\frac{\lambda_t(\mu,\varepsilon)\mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,t)\big)}{\Delta_a}\,,\frac{1}{\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)}\right\}\mathrm{f}_{\xi}(T)+1\right)+1\right]}{(\gamma_{_T})^2(\mu^\star-\varepsilon)(1-\mu^\star-\varepsilon)\,\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)}\\ &+\frac{\Gamma}{(\gamma_{_T})^2(\mu^\star-\varepsilon)(1-\mu^\star-\varepsilon)\,\mathrm{kl}(\mu_a+\varepsilon|\mu^\star-\varepsilon)}\,\mathrm{f}_{\xi}(|\mathcal{U}(\varepsilon)|)\ .\end{aligned} \tag{4.162}$$

By combining Equations (4.157), (4.158), (4.160) and (4.162), we have for all sub-optimal arm $a \notin \mathcal{A}^\star$,

$$
\begin{aligned}
& \Delta_a \, N_a(T) \\
\leqslant \; & \mathbb{I}_{\{\tau \leqslant \overline{\tau}(\varepsilon,T) \leqslant T\}} \Delta_a \, N_a^{\mathrm{opt}}(\overline{\tau}(\varepsilon,T)) \\
& + \frac{\Delta_a \, \Gamma \left[\mathrm{f}_\xi \left(\max\limits_{t \geqslant 1} \left\{ \dfrac{\lambda_t(\mu,\varepsilon) \mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,t)\big)}{\Delta_a} \, , \dfrac{1}{\mathrm{kl}(\mu_a + \varepsilon | \mu^\star - \varepsilon)} \right\} \mathrm{f}_\xi(T) + 1 \right) + 1 \right]}{(\gamma_T)^2 (\mu^\star - \varepsilon)(1 - \mu^\star - \varepsilon) \, \mathrm{kl}(\mu_a + \varepsilon | \mu^\star - \varepsilon)} \\
& + \Delta_a \, \tau \\
& + \frac{\Delta_a \, \Gamma}{(\gamma_T)^2 (\mu^\star - \varepsilon)(1 - \mu^\star - \varepsilon) \, \mathrm{kl}(\mu_a + \varepsilon | \mu^\star - \varepsilon)} \, \mathrm{f}_\xi(|\mathcal{U}(\varepsilon)|) + \textstyle\sum_{t \geqslant 1} \mathbb{I}_{\{a_{t+1} = a, \, t \in \mathcal{U}(\varepsilon)\}} \, .
\end{aligned}
\tag{4.163}
$$

By summing previous Equation (4.163) over the sub-optimal arms, we obtain

$$
\begin{aligned}
& \sum_{a \notin \mathcal{A}^\star} \Delta_a \, N_a(T) \\
\leqslant \; & \mathbb{I}_{\{\tau \leqslant \overline{\tau}(\varepsilon,T) \leqslant T\}} \sum_{a \notin \mathcal{A}^\star} \Delta_a \, N_a^{\mathrm{opt}}(\overline{\tau}(\varepsilon,T)) \\
& + \sum_{a \notin \mathcal{A}^\star} \frac{\Delta_a \, \Gamma \left[\mathrm{f}_\xi \left(\max\limits_{t \geqslant 1} \left\{ \dfrac{\lambda_t(\mu,\varepsilon) \mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,t)\big)}{\Delta_a} \, , \dfrac{1}{\mathrm{kl}(\mu_a + \varepsilon | \mu^\star - \varepsilon)} \right\} \mathrm{f}_\xi(T) + 1 \right) + 1 \right]}{(\gamma_T)^2 (\mu^\star - \varepsilon)(1 - \mu^\star - \varepsilon) \, \mathrm{kl}(\mu_a + \varepsilon | \mu^\star - \varepsilon)} \\
& + \sum_{a \notin \mathcal{A}^\star} \Delta_a \cdot \tau \\
& + \sum_{a \notin \mathcal{A}^\star} \frac{\Delta_a \, \Gamma}{(\gamma_T)^2 (\mu^\star - \varepsilon)(1 - \mu^\star - \varepsilon) \, \mathrm{kl}(\mu_a + \varepsilon | \mu^\star - \varepsilon)} \, \mathrm{f}_\xi(|\mathcal{U}(\varepsilon)|) + \sum_{t \geqslant 1} \sum_{a \notin \mathcal{A}^\star} \mathbb{I}_{\{a_{t+1} = a, \, t \in \mathcal{U}(\varepsilon)\}} \, .
\end{aligned}
\tag{4.164}
$$

Since

$$
\sum_{t \geqslant 1} \sum_{a \notin \mathcal{A}^\star} \mathbb{I}_{\{a_{t+1} = a, \, t \in \mathcal{U}(\varepsilon)\}} = \sum_{t \geqslant 1} \mathbb{I}_{\{a_{t+1} \notin \mathcal{A}^\star, \, t \in \mathcal{U}(\varepsilon)\}} \leqslant \sum_{t \geqslant 1} \mathbb{I}_{\{t \in \mathcal{U}(\varepsilon)\}} = |\mathcal{U}(\varepsilon)| \tag{4.165}
$$

and

$$
\begin{aligned}
\mathrm{F}_\xi(\mu,\varepsilon,T) &= \max_{a \notin \mathcal{A}^\star} \mathrm{f}_\xi \left(\max_{t \geqslant 1} \left\{ \frac{\lambda_t(\mu,\varepsilon) \mathfrak{C}_{\{\theta^\star\}}\big(\mu(\varepsilon,t)\big)}{\Delta_a} \, , \frac{1}{\mathrm{kl}(\mu_a + \varepsilon | \mu^\star - \varepsilon)} \right\} \mathrm{f}_\xi(T) + 1 \right) + 1 \\
\mathrm{G}_T(\mu,\varepsilon) &= \sum_{a \notin \mathcal{A}^\star} \frac{\Delta_a \, \Gamma}{(\gamma_T)^2 (\mu^\star - \varepsilon)(1 - \mu^\star - \varepsilon) \, \mathrm{kl}(\mu_a + \varepsilon | \mu^\star - \varepsilon)} \\
\Sigma_\Delta &= \sum_{a \notin \mathcal{A}^\star} \Delta_a \, ,
\end{aligned}
\tag{4.166}
$$

Equation (4.164) implies

$$
\begin{aligned}
& \sum_{a \notin \mathcal{A}^\star} \Delta_a \, N_a(T) \\
\leqslant \; & \mathbb{I}_{\{\tau \leqslant \overline{\tau}(\varepsilon,T) \leqslant T\}} \sum_{a \notin \mathcal{A}^\star} \Delta_a \, N_a^{\mathrm{opt}}(\overline{\tau}(\varepsilon,T)) \\
& + \mathrm{G}_T(\mu,\varepsilon) \cdot \mathrm{F}_\xi(\mu,\varepsilon,T) \\
& + \Sigma_\Delta \cdot \tau \\
& + \mathrm{G}_T(\mu,\varepsilon) \cdot \mathrm{f}_\xi(|\mathcal{U}(\varepsilon)|) + |\mathcal{U}(\varepsilon)| \, .
\end{aligned}
\tag{4.167}
$$

Combining Equations (4.161) and (4.167) ends the proof. □

4.6.4 Upper bound on the regret

In this section, we prove Theorem 3.

Proof. Theorem 3 is obtained as a corollary of Proposition 4. More precisely, the monotony of $\lambda_\cdot(\mu,\cdot)$ and $\mathfrak{C}_{\{\theta^\star\}}(\cdot)=\mathfrak{C}_\Theta(\cdot)$ (see Lemma 16) simplify the expressions of $\Lambda_\tau(\mu,\varepsilon)$ and $C_\tau(\mu,\varepsilon)$. The upper bounds $\gamma_1\leqslant 1/7$, $\mathfrak{C}_\Theta(\cdot)\leqslant\mathfrak{C}_0(\cdot)$ (see Lemma 16) and Pinsker's inequality simplify the terms $\mathrm{F}_\xi(\mu,\varepsilon,T)$ and $\mathrm{G}_T(\mu,\varepsilon)$. Then, Theorem 3 is obtained by taking the expectation in both sides of the inequality. Indeed, since $\mathrm{f}_\xi(\cdot)$ is a concave function, we have $\mathbb{E}_\nu[\mathrm{f}_\xi(|\mathcal{U}(\varepsilon)|)]\leqslant\mathrm{f}_\xi(\mathbb{E}_\nu[|\mathcal{U}(\varepsilon)|])$ and the constant $C_{\xi,d,\varepsilon}=\mathbb{E}_\nu[|\mathcal{U}(\varepsilon)|]$. From Lemma 8 and Equation (4.95), we have

$$C_{\xi,d,\varepsilon}=K_\xi^\star+2\,|\mathcal{A}|^2\,K_{\xi,1}+\frac{(d+1)e^{d+1}}{2\varepsilon^2}\left(\mathrm{f}_\xi\left(\frac{(d+1)e^{d+1}}{2\varepsilon_\nu^2}\right)\right)^2+\sum_{n\geqslant 3}\frac{3\,|\mathcal{A}|^2}{n\,(\log(n))^\xi}+|\mathcal{A}|^2\,\frac{e^{2\varepsilon^2}}{\varepsilon^2}\,,$$

where $K_\xi^\star$ and $K_{\xi,1}$ are explained in Lemma 8. □

We end this subsection by proving the following lemma that ensures a nice control of $\mathfrak{C}_\Theta(\mu(\cdot,\cdot))$ appearing in the first order term of the upper bound on the regret.

Lemma 16 (Monotonic properties of $\mathfrak{C}_\Theta(\mu(\cdot,\cdot))$)**.** *For $0<\varepsilon<\varepsilon_0$, $(\mathfrak{C}_\Theta(\mu(\varepsilon,t)))_{t\geqslant 1}$ and $\mathfrak{C}_\Theta(\mu(\cdot,\infty))$ are non-increasing and $\lim_{\varepsilon\to 0}\lim_{t\to\infty}\mathfrak{C}_\Theta(\mu(\varepsilon,t))=\lim_{\varepsilon\to 0}\mathfrak{C}_\Theta(\mu(\varepsilon,\infty))=\mathfrak{C}_\Theta(\mu)$. Finally, we denote by*

$$\mathfrak{C}_0(\mu)\coloneqq\min_{n\in\mathbb{R}_+^{\mathcal{A}}}\left\{\sum_{a\in\mathcal{A}}n_a\left(\max(\mu)-\mu_a\right)\ \text{ s.t. }\ \forall a\notin\operatorname{argmax}(\mu),\ \mathrm{kl}\big(\mu_a|\max(\mu)\big)n_a\geqslant 1\right\}=\sum_{a\notin\operatorname{argmax}(\mu)}\frac{\max(\mu)-\mu_a}{\mathrm{kl}\big(\mu_a|\max(\mu)\big)}$$

the minimization problem $\mathfrak{C}_{[-1,1]^{\mathcal{A}\times\mathcal{A}}}(\mu)$ when there is no structure, that is when $\Theta=[-1,1]^{\mathcal{A}\times\mathcal{A}}$, for all $\mu\in(0,1)^{\mathcal{A}}$. Then we have $\mathfrak{C}_\Theta(\cdot)\leqslant\mathfrak{C}_0(\cdot)$.

Proof. We first note for all $\mu\in(0,1)^{\mathcal{A}}$, by considering the change of variables $n_a\leftarrow\Delta_a\,n_a$ if $a\notin\operatorname{argmax}\mu$ and $n_a\leftarrow n_a$ if $a\in\operatorname{argmax}\mu$, we have

$$\mathfrak{C}_\Theta(\mu)\coloneqq\min_{n\in\mathbb{R}_+^{\mathcal{A}}}\ \left\{\sum_{a\in\mathcal{A}}n_a\ \text{s.t.}\ \forall a\notin\operatorname{argmax}\mu,\ \min_{\theta\in\overline{\Theta}_a}\sum_{\substack{a'\in\mathcal{A}\\ \mu_{a'}\leqslant\max\mu-\theta_{a,a'}}}\frac{\mathrm{kl}\big(\mu_{a'}|\max\mu-\theta_{a,a'}\big)}{\max\mu-\mu_{a'}}\,n_{a'}\geqslant 1\right\}. \tag{4.168}$$

For $0<\varepsilon'\leqslant\varepsilon<\varepsilon_0$ and $t'\geqslant t\geqslant 1$, we show that

$$\mathfrak{C}_\Theta(\mu(\varepsilon',t'))\leqslant\mathfrak{C}_\Theta(\mu(\varepsilon,t))\,. \tag{4.169}$$

Then, since $\lim_{t\to\infty}\mu(\varepsilon,t)=\mu(\varepsilon,\infty)$ for $0<\varepsilon<\varepsilon_0$ and $\lim_{\varepsilon\to 0}\mu(\varepsilon,\infty)=\mu$, previous Equation (4.169) and Lemma 27 will imply

$$\lim_{\varepsilon\to 0}\lim_{t\to\infty}\mathfrak{C}_\Theta(\mu(\varepsilon,t))=\lim_{\varepsilon\to 0}\mathfrak{C}_\Theta(\mu(\varepsilon,\infty))=\mathfrak{C}_\Theta(\mu)\,. \tag{4.170}$$

Let $0<\varepsilon'\leqslant\varepsilon<\varepsilon_0$ and $t'\geqslant t\geqslant 1$. Let $n\in\mathbb{R}_+^{\mathcal{A}}$ such that

$$\forall a\notin\operatorname{argmax}\mu(\varepsilon,t),\quad\min_{\theta\in\overline{\Theta}_a}\sum_{\substack{a'\in\mathcal{A}\\ \mu_{a'}(\varepsilon,t)\leqslant\max\mu(\varepsilon,t)-\theta_{a,a'}}}\frac{\mathrm{kl}\big(\mu_{a'}(\varepsilon,t)|\max\mu(\varepsilon,t)-\theta_{a,a'}\big)}{\max\mu(\varepsilon,t)-\mu_a(\varepsilon,t)}\,n_{a'}\geqslant 1\,. \tag{4.171}$$

We first note that

$$\operatorname{argmax} \mu(\varepsilon, t) = \mu(\varepsilon', t') = \operatorname{argmax} \mu \,, \tag{4.172}$$

$$\max \mu(\varepsilon, t) = \mu^\star - \varepsilon \leqslant \mu^\star - \varepsilon' = \max \mu(\varepsilon', t') \,, \tag{4.173}$$

and

$$\forall a \notin \operatorname{argmax} \mu, \quad \mu_a(\varepsilon, t) = \mu_a + \varepsilon + \varepsilon(t) \geqslant \mu + \varepsilon' + \varepsilon(t') = \mu_a(\varepsilon', t') \,. \tag{4.174}$$

In particular, from Equations (4.172), (4.173) and (4.174), Lemma 25 implies

$$\forall a \notin \operatorname{argmax} \mu, \ \forall \theta \in \overline{\Theta}_a, \ \frac{\operatorname{kl}\big(\mu_{a'}(\varepsilon, t) | \max \mu(\varepsilon, t) - \theta_{a,a'}\big)}{\max \mu(\varepsilon, t) - \mu_a(\varepsilon, t)} \leqslant \frac{\operatorname{kl}\big(\mu_{a'}(\varepsilon', t') | \max \mu(\varepsilon', t') - \theta_{a,a'}\big)}{\max \mu(\varepsilon', t') - \mu_a(\varepsilon', t')} \,. \tag{4.175}$$

Combining all together Equations (4.171), (4.172), (4.173), (4.174), (4.175), we have

$$\forall a \notin \operatorname{argmax} \mu(\varepsilon', t'), \quad \min_{\theta \in \overline{\Theta}_a} \sum_{\substack{a' \in \mathcal{A} \\ \mu_{a'}(\varepsilon', t') \leqslant \max \mu(\varepsilon', t') - \theta_{a,a'}}} \frac{\operatorname{kl}\big(\mu_{a'}(\varepsilon', t') | \max \mu(\varepsilon', t') - \theta_{a,a'}\big)}{\max \mu(\varepsilon', t') - \mu_a(\varepsilon', t')} n_{a'} \geqslant 1 \,. \tag{4.176}$$

Previous Equation 4.176 implies

$$\mathfrak{C}_\Theta(\mu(\varepsilon', t')) \leqslant \sum_{a \in \mathcal{A}} n_a \,. \tag{4.177}$$

Since previous Equation (4.177) holds for all $n \in \mathbb{R}_+^{\mathcal{A}}$ satisfying Equation (4.171), we have

$$\mathfrak{C}_\Theta(\mu(\varepsilon', t')) \leqslant \mathfrak{C}_\Theta(\mu(\varepsilon, t)) \ . \tag{4.178}$$

□

Chapter 5

Concentration Inequalities for Structured Bandits

In this chapter, we state and prove Theorem 4, the main concentration result used in Chapter 4 for `IMED-GS` analysis. This key result enables to tighten some terms and reduce the burn-in phase as it enables the use of a parameter $d < |\mathcal{A}| - 1$. See also the numerical experiments in Section A.4 showing the practical benefit of this contribution.

Theorem 4. *For all set of Bernoulli distributions $\nu = (\nu_a)_{a\in\mathcal{A}}$ with means $(\mu_a)_{a\in\mathcal{A}} \subset (0,1)^{\mathcal{A}}$, for all $\xi \geqslant 0$, for all $\underline{M} \geqslant 1$, for all $\mathcal{A}' \subset \mathcal{A}$, for all $\Phi \geqslant |\mathcal{A}'|+1$, for all $\varepsilon > 0$, for all $n \geqslant n_{\xi,\mathcal{A}',\varepsilon}$,*

$$\begin{aligned}
&\mathbb{P}_\nu\left(\bigcup_{t\geqslant 1}\left\{\sum_{a\in\mathcal{A}'} N_a(t)\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a-\varepsilon) \geqslant \Phi \vee (\mathrm{f}_\xi(n) - \mathrm{f}_\xi(N_{\mathcal{A}'}(t)))\right\} \cap \Omega_{\mathcal{A}',\varepsilon,n}(t)\right) \\
\leqslant\ & \frac{e^{|\mathcal{A}'|+2}}{|\mathcal{A}'|^{|\mathcal{A}'|}} \log(n)^{|\mathcal{A}'|+1} \left[\Phi \vee \mathrm{f}_\xi(n) + 2\varepsilon^2 \underline{N}_{\nu,\xi,\mathcal{A}'}(n)\right]^{2|\mathcal{A}'|+1} e^{-\left[\Phi\vee \mathrm{f}_\xi(n) + 2\varepsilon^2 \underline{N}_{\nu,\xi,\mathcal{A}'}(n)\right]},
\end{aligned}$$

where $\Omega_{\mathcal{A}',\varepsilon,n}(t) = \left\{\underline{M} \leqslant N_{\mathcal{A}'}(t) \leqslant |\mathcal{A}'|\,\mathrm{f}_\xi(n)/2\varepsilon^2\right\} \cap \left\{\forall a \in \mathcal{A}',\ \widehat{\mu}_a(t) < \mu_a - \varepsilon\right\}$,

$$\underline{N}_{\nu,\xi,\mathcal{A}'}(n) = \underline{M} \vee \frac{\mathrm{f}_\xi(n)}{1 - \log(1 - \max_{a\in\mathcal{A}'} \mu_a)} \quad \textit{and} \quad n_{\xi,\mathcal{A}',\varepsilon} = \left(e^2 \vee \frac{|\mathcal{A}'|}{2\varepsilon^2}\right)\left(\mathrm{f}_\xi\left(e^2 \vee \frac{|\mathcal{A}'|}{2\varepsilon^2}\right)\right)^2.$$

Although the statement looks a little intricate, it enables to obtain more precise control of the concentration terms compared to alternative tools present in the literature (e.g. in Magureanu et al. (2014)). Note the presence of the random term $N_{\mathcal{A}'}(t)$ in event $\Omega_{\mathcal{A}',\varepsilon,n}(t)$ in the left-hand side of the deviation inequality. Besides, it enables to show that setting $\xi = 1$ is enough for the regret to be controlled.

Proof. The five main steps of the proof are the following. First, we establishing a lower bound on the aggregated number of pulls $N_{\mathcal{A}'}(t)$ for $t \geqslant 1$. Second, we proceed to a change of measurement applying Lemma 24. Third, we use twice the peeling technique, for the aggregated number of pulls $N_{\mathcal{A}'}(t)$ and for each ratio $N_a(t)/N_{\mathcal{A}'}(t)$, $a \in \mathcal{A}'$. Fourth, we control each band of the peeling using Lemma 17 based on multivariate stochastic ordering, see Lemma 8 from Magureanu et al. (2014). Fifth, we apply the union bound.

Let us consider $t \geqslant 1$ such that

$$\begin{aligned}
&\forall a \in \mathcal{A}',\ \widehat{\mu}_a(t) < \mu_a - \varepsilon \\
&\underline{M} \leqslant N_{\mathcal{A}'}(t) \leqslant |\mathcal{A}'|\,\mathrm{f}_\xi(n)/2\varepsilon^2 \\
&\sum_{a\in\mathcal{A}'} N_a(t)\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a-\varepsilon) \geqslant \Phi \vee (\mathrm{f}_\xi(n) - \mathrm{f}_\xi(N_{\mathcal{A}'}(t)))\ .
\end{aligned} \tag{5.1}$$

1. Lower bound on $N_{\mathcal{A}'}(t)$:
From Equation (5.1), we have $\widehat{\mu}_a(t) < \mu_a - \varepsilon$ for all $a \in \mathcal{A}'$. Then, the monotonic properties of $\mathrm{kl}(\cdot\,|\,\cdot)$ imply

$$\forall a \in \mathcal{A}', \quad \mathrm{kl}(\widehat{\mu}_a(t)|\mu_a - \varepsilon) \leqslant \mathrm{kl}(0|\mu_a) = -\log(1-\mu_a) \tag{5.2}$$

and

$$\sum_{a\in\mathcal{A}'} N_a(t)\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a - \varepsilon) \leqslant -\log(1-\max_{a\in\mathcal{A}'}\mu_a) N_{\mathcal{A}'}(t)\,. \tag{5.3}$$

From Equation (5.1), since $N_{\mathcal{A}'}(t) \geqslant \mathrm{f}_\xi(N_{\mathcal{A}'}(t))$, Equation (5.1) and previous Equation (5.3) imply

$$N_{\mathcal{A}'}(t) \geqslant \underline{N} \coloneqq \underline{M} \vee \frac{\mathrm{f}_\xi(n)}{1 - \log(1-\max_{a\in\mathcal{A}'}\mu_a) \vee e}\,. \tag{5.4}$$

2. Change of measurement :
From Equation (5.1), we have $\widehat{\mu}_a(t) < \mu_a - \varepsilon < \mu_a$ for all $a \in \mathcal{A}'$. Then, Lemma 24 implies

$$\forall a \in \mathcal{A}', \quad \mathrm{kl}(\widehat{\mu}_a(t)|\mu_a - \varepsilon) \leqslant \mathrm{kl}(\widehat{\mu}_a(t)|\mu_a) - 2\varepsilon^2 \tag{5.5}$$

and

$$\sum_{a\in\mathcal{A}'} N_a(t)\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a - \varepsilon) \leqslant \sum_{a\in\mathcal{A}'} N_a(t)\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a) - 2\varepsilon^2 N_{\mathcal{A}'}(t)\,. \tag{5.6}$$

From Lemma 28, previous Equation (5.6) implies

$$\sum_{a\in\mathcal{A}'} N_a(t)\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a) \geqslant \Phi \vee (\mathrm{f}_\xi(n) - \mathrm{f}_\xi(N_{\mathcal{A}'}(t))) + 2\varepsilon^2 N_{\mathcal{A}'}(t) \tag{5.7}$$

or similarly

$$\sum_{a\in\mathcal{A}'} \frac{N_a(t)}{N_{\mathcal{A}'}(t)} \mathrm{kl}(\widehat{\mu}_a(t)|\mu_a) \geqslant \frac{\Phi \vee (\mathrm{f}_\xi(n) - \mathrm{f}_\xi(N_{\mathcal{A}'}(t))) + 2\varepsilon^2 N_{\mathcal{A}'}(t)}{N_{\mathcal{A}'}(t)}\,. \tag{5.8}$$

3. Peeling :
Let us consider

$$\rho = \sqrt{\frac{\Phi \vee \mathrm{f}_\xi(n) + 2\varepsilon^2\underline{N} - 1}{\Phi \vee \mathrm{f}_\xi(n) + 2\varepsilon^2\underline{N}}}\,. \tag{5.9}$$

Since $\Phi \vee \mathrm{f}_\xi(n) + 2\varepsilon^2\underline{N} - 1 > \Phi - 1 \geqslant |\mathcal{A}'|$, previous Equation (5.9) implies

$$\frac{|\mathcal{A}'|}{|\mathcal{A}'|+1} \leqslant \rho^2 = 1 - \frac{1}{\Phi \vee \mathrm{f}_\xi(n) + 2\varepsilon^2\underline{N}} < 1\,. \tag{5.10}$$

Now we introduce decreasing geometric sequences $(N_k)_{k\geqslant 0}$ and $(x_\ell)_{\ell\geqslant 0}$ for the peeling :

$$\left\{ \begin{array}{rcl} N_0 & = & \dfrac{|\mathcal{A}'|\,\mathrm{f}_\xi(n)}{2\varepsilon^2} \\ N_{k+1} & = & \rho\, N_k, \quad k \geqslant 0 \end{array} \right. \qquad \left\{ \begin{array}{rcl} x_0 & = & 1 \\ x_{\ell+1} & = & \rho\, x_\ell, \quad \ell \geqslant 0\,. \end{array} \right. \tag{5.11}$$

By considering

$$K = \left\lfloor \frac{\log(N_0/\underline{N})}{-\log(\rho)} \right\rfloor \qquad L = \left\lfloor \frac{\log(N_0)}{-\log(\rho)} \right\rfloor, \tag{5.12}$$

we have

$$N_{K+1} \leqslant \underline{N} \leqslant N_{\mathcal{A}'}(t) \qquad x_{L+1} \leqslant \frac{1}{N_0} \leqslant \frac{N_a(t)}{N_{\mathcal{A}'}(t)}\,,\ \forall a \in \mathcal{A}'\,. \tag{5.13}$$

Then, by considering the bands for $k \in [\![0, K]\!]$, for $(\ell_a)_{a\in\mathcal{A}'} \subset [\![0, L]\!]$,

$$B(k,(\ell_a)_{a\in\mathcal{A}'}) = \left\{ t \geqslant 1 : \begin{array}{l} \forall a \in \mathcal{A}',\ \widehat{\mu}_a(t) < \mu_a \\ N_{k+1} \leqslant N_{\mathcal{A}'}(t) \leqslant N_k \\ \forall a \in \mathcal{A}',\ x_{\ell_a+1} \leqslant \dfrac{N_a(t)}{N_{\mathcal{A}'}(t)} \leqslant x_{\ell_a} \\ \displaystyle\sum_{a\in\mathcal{A}'} Z_a \geqslant \dfrac{\Phi \vee (\mathrm{f}_\xi(n) - \mathrm{f}_\xi(N_{\mathcal{A}'}(t))) + 2\varepsilon^2 N_{\mathcal{A}'}(t)}{N_{\mathcal{A}'}(t)} \end{array} \right\}, \tag{5.14}$$

we have

$$t \in \bigcup_{\substack{k\in[\![0,K]\!] \\ (\ell_a)_{a\in\mathcal{A}'}\subset[\![0,L]\!]}} B(k,(\ell_a)_{a\in\mathcal{A}'}) \ , \tag{5.15}$$

where

$$\forall a \in \mathcal{A}', \quad Z_a = \frac{N_a(t)}{N_{\mathcal{A}'}(t)} \mathrm{kl}(\widehat{\mu}_a(t)|\mu_a) \ . \tag{5.16}$$

Monotony of both function $\mathrm{f}_\xi(\cdot)$ and sequence $(N_k)_{k\geqslant 0}$ imply for $k \in [\![0\,, K]\!]$, for $(\ell_a)_{a\in\mathcal{A}'} \subset [\![0\,, L]\!]$,

$$B(k,(\ell_a)_{a\in\mathcal{A}'}) \subset \overline{\mathcal{B}}(k,(\ell_a)_{a\in\mathcal{A}'}) \cap \overline{S}(k,(\ell_a)_{a\in\mathcal{A}'}) \, , \tag{5.17}$$

with

$$\overline{\mathcal{B}}(k,(\ell_a)_{a\in\mathcal{A}'}) \coloneqq \left\{ t \geqslant 1 : \begin{array}{l} \forall a \in \mathcal{A}',\ \widehat{\mu}_a(t) < \mu_a \\ N_{k+1} \leqslant N_{\mathcal{A}'}(t) \leqslant N_k \\ \forall a \in \mathcal{A}',\ x_{\ell_a+1} \leqslant \dfrac{N_a(t)}{N_{\mathcal{A}'}(t)} \leqslant x_{\ell_a} \end{array} \right\} \tag{5.18}$$

$$\overline{S}(k,(\ell_a)_{a\in\mathcal{A}'}) \coloneqq \left\{ t \geqslant 1 : \sum_{a\in\mathcal{A}'} \mathbb{I}_{\left\{t\in\overline{\mathcal{B}}\left(k,(\ell_a)_{a\in\mathcal{A}'}\right)\right\}} Z_a \geqslant \frac{\Phi \vee (\mathrm{f}_\xi(n) - \mathrm{f}_\xi(N_0)) + 2\varepsilon^2 \underline{N}}{N_k} \right\},$$

where, according to Lemma 28,

$$\mathrm{f}_\xi(n) - \mathrm{f}_\xi(N_0) \geqslant 0 \qquad\qquad n \geqslant N_0 \, , \tag{5.19}$$

since $n \geqslant n_{\nu,\mathcal{A}',\varepsilon} = (e^2 \vee (|\mathcal{A}'|/2\varepsilon^2))\,(\mathrm{f}_\xi(e^2 \vee (|\mathcal{A}'|/2\varepsilon^2)))^2$. In particular, Equations (5.15) and (5.17)-(5.18) imply

$$t \in \bigcup_{\substack{k\in[\![0,K]\!] \\ (\ell_a)_{a\in\mathcal{A}'}\subset[\![0,L]\!]}} \overline{\mathcal{B}}(k,(\ell_a)_{a\in\mathcal{A}'}) \cap \overline{S}(k,(\ell_a)_{a\in\mathcal{A}'}) \ . \tag{5.20}$$

4. Stochastic ordering :

The total probability rule implies for $k\in[\![0\,,K]\!]$, $(\ell_a)_{a\in\mathcal{A}'}\subset[\![0\,,L]\!]$, for all $\zeta\in\mathbb{R}_+^{\mathcal{A}'}$,

$$\begin{aligned}
&\mathbb{P}_\nu\left(\bigcap_{a\in\mathcal{A}'}\mathbb{I}_{\left\{t\in\overline{\mathcal{B}}\left(k,(\ell_a)_{a\in\mathcal{A}'}\right)\right\}}Z_a\geqslant\zeta_a\right)\\
=&\sum_{\substack{M\in\mathbb{N}\\(n_a)_{a\in\mathcal{A}'}\subset\mathbb{N}}}\mathbb{P}_\nu\left(N_{\mathcal{A}'}(t)=M,\,\bigcap_{a\in\mathcal{A}'}N_a(t)=n_a,\,\mathbb{I}_{\left\{t\in\overline{\mathcal{B}}\left(k,(\ell_a)_{a\in\mathcal{A}'}\right)\right\}}Z_a\geqslant\zeta_a\right)\\
=&\sum_{\substack{M\in\mathbb{N}\\(n_a)_{a\in\mathcal{A}'}\subset\mathbb{N}}}\prod_{a\in\mathcal{A}'}\mathbb{P}_\nu\Big(N_{\mathcal{A}'}(t)=M,\,(N_{a'}(t))_{a'\in\mathcal{A}'}=(n_{a'})_{a'\in\mathcal{A}'},\,\mathbb{I}_{\left\{t\in\overline{\mathcal{B}}\left(k,(\ell_a)_{a\in\mathcal{A}'}\right)\right\}}Z_a\geqslant\zeta_a\Big)\\
\leqslant&\prod_{a\in\mathcal{A}'}\sum_{\substack{M\in\mathbb{N}\\(n_a)_{a\in\mathcal{A}'}\subset\mathbb{N}}}\mathbb{P}_\nu\Big(N_{\mathcal{A}'}(t)=M,\,(N_{a'}(t))_{a'\in\mathcal{A}'}=(n_{a'})_{a'\in\mathcal{A}'},\,\mathbb{I}_{\left\{t\in\overline{\mathcal{B}}\left(k,(\ell_a)_{a\in\mathcal{A}'}\right)\right\}}Z_a\geqslant\zeta_a\Big)\\
=&\prod_{a\in\mathcal{A}'}\mathbb{P}_\nu\Big(\mathbb{I}_{\left\{t\in\overline{\mathcal{B}}\left(k,(\ell_a)_{a\in\mathcal{A}'}\right)\right\}}Z_a\geqslant\zeta_a\Big)\,.
\end{aligned}\tag{5.21}$$

By combining Previous Equation (5.21) and Lemma 32, we have for $k\in[\![0\,,K]\!]$, $(\ell_a)_{a\in\mathcal{A}'}\subset[\![0\,,L]\!]$, for all $\zeta\in\mathbb{R}_+^{\mathcal{A}'}$,

$$\begin{aligned}
\mathbb{P}_\nu\left(\bigcap_{a\in\mathcal{A}'}\mathbb{I}_{\left\{t\in\overline{\mathcal{B}}\left(k,(\ell_a)_{a\in\mathcal{A}'}\right)\right\}}Z_a\geqslant\zeta_a\right)&\leqslant\prod_{a\in\mathcal{A}'}\mathbb{P}_\nu\left(\bigcup_{\substack{t\geqslant1\\\widehat{\mu}_a(t)<\mu_a\\N_a(t)\geqslant x_{\ell_a+1}N_{k+1}}}\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a)\geqslant\zeta_a/x_{\ell_a}\right)\\
&\leqslant\prod_{a\in\mathcal{A}'}\exp\left(-\frac{x_{\ell_a+1}}{x_{\ell_a}}N_{k+1}\,\zeta_a\right)\\
&=\prod_{a\in\mathcal{A}'}\exp(-\rho\,N_{k+1}\,\zeta_a)\\
&=\exp\left(-\rho\,N_{k+1}\sum_{a\in\mathcal{A}'}\zeta_a\right)\,.
\end{aligned}\tag{5.22}$$

Furthermore, for $k\in[\![0,K]\!]$,

$$\rho\,N_{k+1}\frac{\Phi\vee(\mathrm{f}_\xi(n)-\mathrm{f}_\xi(N_0))+2\varepsilon^2\underline{N}}{N_k}=\rho^2\left[\Phi\vee(\mathrm{f}_\xi(n)-\mathrm{f}_\xi(N_0))+2\varepsilon^2\underline{N}\right]\,.\tag{5.23}$$

Using Equation (5.10) we have $\rho^2\geqslant|\mathcal{A}'|\,/\,|\mathcal{A}'|+1$, and previous Equation (5.23) implies

$$\rho\,N_{k+1}\frac{\mathrm{f}_\xi(n)-\mathrm{f}_\xi(N_0)+2\varepsilon^2\underline{N}}{N_k}\geqslant\frac{|\mathcal{A}'|}{|\mathcal{A}'|+1}\times\Phi\geqslant\frac{|\mathcal{A}'|}{|\mathcal{A}'|+1}\times(|\mathcal{A}'|+1)=|\mathcal{A}'|\,.\tag{5.24}$$

This Equation (5.24) is the key equation allowing the use of Lemma 17. From Lemma 17, Equations (5.22) and (5.24) imply for $k\in[\![0\,,K]\!]$, $(\ell_a)_{a\in\mathcal{A}'}\subset[\![0\,,L]\!]$,

$$\begin{aligned}
&\mathbb{P}_\nu\left(\sum_{a\in\mathcal{A}'}\mathbb{I}_{\left\{t\in\overline{\mathcal{B}}\left(k,(\ell_a)_{a\in\mathcal{A}'}\right)\right\}}Z_a\geqslant\frac{\Phi\vee(\mathrm{f}_\xi(n)-\mathrm{f}_\xi(N_0))+2\varepsilon^2\underline{N}}{N_k}\right)\\
\leqslant&\frac{e^{|\mathcal{A}'|}}{|\mathcal{A}'|^{|\mathcal{A}'|}}\left[\rho^2\left(\Phi\vee(\mathrm{f}_\xi(n)-\mathrm{f}_\xi(N_0))+2\varepsilon^2\underline{N}\right)\right]^{|\mathcal{A}'|}e^{-\rho^2\left(\Phi\vee\left(\mathrm{f}_\xi(n)-\mathrm{f}_\xi(N_0)\right)+2\varepsilon^2\underline{N}\right)}\,,
\end{aligned}\tag{5.25}$$

that is, according to Equation (5.17),

$$\begin{aligned}
&\mathbb{P}_\nu\big(t\in\overline{\mathcal{B}}(k,(\ell_a)_{a\in\mathcal{A}'})\cap\overline{S}(k,(\ell_a)_{a\in\mathcal{A}'})\big)\\
\leqslant\ & \frac{e^{|\mathcal{A}'|}}{|\mathcal{A}'|^{|\mathcal{A}'|}}\left[\rho^2\left(\Phi\vee(\mathrm{f}_\xi(n)-\mathrm{f}_\xi(N_0))+2\varepsilon^2\underline{N}\right)\right]^{|\mathcal{A}'|}e^{-\rho^2\left(\Phi\vee\left(\mathrm{f}_\xi(n)-\mathrm{f}_\xi(N_0)\right)+2\varepsilon^2\underline{N}\right)}\\
\leqslant\ & \frac{e^{|\mathcal{A}'|}}{|\mathcal{A}'|^{|\mathcal{A}'|}}\left[\Phi\vee\mathrm{f}_\xi(n)+2\varepsilon^2\underline{N}\right]^{|\mathcal{A}'|}e^{-\rho^2\left(\Phi\vee\left(\mathrm{f}_\xi(n)-\mathrm{f}_\xi(N_0)\right)+2\varepsilon^2\underline{N}\right)}. \qquad (5.26)
\end{aligned}$$

5. Union bound :
The union bound writes

$$\begin{aligned}
&\mathbb{P}_\nu\left(t\in\bigcup_{\substack{k\in[\![0,K]\!]\\(\ell_a)_{a\in\mathcal{A}'}\subset[\![0,L]\!]}}\overline{\mathcal{B}}(k,(\ell_a)_{a\in\mathcal{A}'})\cap\overline{S}(k,(\ell_a)_{a\in\mathcal{A}'})\right)\\
\leqslant\ & \sum_{\substack{k\in[\![0,K]\!]\\(\ell_a)_{a\in\mathcal{A}'}\subset[\![0,L]\!]}}\mathbb{P}_\nu\big(t\in\overline{\mathcal{B}}(k,(\ell_a)_{a\in\mathcal{A}'})\cap\overline{S}(k,(\ell_a)_{a\in\mathcal{A}'})\big). \qquad (5.27)
\end{aligned}$$

By combining both previous Equations (5.26) and (5.27), we have

$$\begin{aligned}
&\mathbb{P}_\nu\left(t\in\bigcup_{\substack{k\in[\![0,K]\!]\\(\ell_a)_{a\in\mathcal{A}'}\subset[\![0,L]\!]}}\overline{\mathcal{B}}(k,(\ell_a)_{a\in\mathcal{A}'})\cap\overline{S}(k,(\ell_a)_{a\in\mathcal{A}'})\right) \qquad (5.28)\\
\leqslant\ & \sum_{\substack{k\in[\![0,K]\!]\\(\ell_a)_{a\in\mathcal{A}'}\subset[\![0,L]\!]}}\frac{e^{|\mathcal{A}'|}}{|\mathcal{A}'|^{|\mathcal{A}'|}}\left[\Phi\vee\mathrm{f}_\xi(n)+2\varepsilon^2\underline{N}\right]^{|\mathcal{A}'|}e^{-\rho^2\left(\Phi\vee\left(\mathrm{f}_\xi(n)-\mathrm{f}_\xi(N_0)\right)+2\varepsilon^2\underline{N}\right)},
\end{aligned}$$

which implies, according to Equations (5.1) and (5.20),

$$\begin{aligned}
&\mathbb{P}_\nu\left(\bigcup_{\substack{t\geqslant1\\\forall a\in\mathcal{A}',\,\widehat{\mu}_a(t)<\mu_a-\varepsilon\\\underline{M}\leqslant N_{\mathcal{A}'}(t)\leqslant|\mathcal{A}'|\mathrm{f}_\xi(n)/2\varepsilon^2}}\sum_{a\in\mathcal{A}'}N_a(t)\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a-\varepsilon)+\mathrm{f}_\xi(N_{\mathcal{A}'}(t))\geqslant\mathrm{f}_\xi(n)\right)\\
\leqslant\ & \sum_{\substack{k\in[\![0,K]\!]\\(\ell_a)_{a\in\mathcal{A}'}\subset[\![0,L]\!]}}\frac{e^{|\mathcal{A}'|}}{|\mathcal{A}'|^{|\mathcal{A}'|}}\left[\Phi\vee\mathrm{f}_\xi(n)+2\varepsilon^2\underline{N}\right]^{|\mathcal{A}'|}e^{-\rho^2\left(\Phi\vee\left(\mathrm{f}_\xi(n)-\mathrm{f}_\xi(N_0)\right)+2\varepsilon^2\underline{N}\right)}\\
=\ & \frac{e^{|\mathcal{A}'|}}{|\mathcal{A}'|^{|\mathcal{A}'|}}(K+1)(L+1)^{|\mathcal{A}'|}\left[\Phi\vee\mathrm{f}_\xi(n)+2\varepsilon^2\underline{N}\right]^{|\mathcal{A}'|}e^{-\rho^2\left(\Phi\vee\left(\mathrm{f}_\xi(n)-\mathrm{f}_\xi(N_0)\right)+2\varepsilon^2\underline{N}\right)}.
\end{aligned} \qquad (5.29)$$

We now simplify these terms in order to optimize the obtained upper bound. We note that Equations (5.10) and (5.19) plus Lemma (29) imply

$$K,\ L\leqslant\log(n)\left[\Phi\vee\mathrm{f}_\xi(n)+2\varepsilon^2\underline{N}\right] \qquad (5.30)$$

and

$$\begin{aligned}&\rho < 1\\&(K+1)\,(L+1)^{|\mathcal{A}'|} \leqslant \log(n)^{|\mathcal{A}'|+1}\left[\Phi \vee \mathrm{f}_\xi(n) + 2\varepsilon^2\underline{N}\right]^{|\mathcal{A}'|+1}\,.\end{aligned} \tag{5.31}$$

Furthermore, Equations (5.9) and (5.19) imply

$$\begin{aligned}
& -\rho^2\left(\Phi \vee (\mathrm{f}_\xi(n) - \mathrm{f}_\xi(N_0)) + 2\varepsilon^2\underline{N}\right)\\
\leqslant\ & -\rho^2\left(\Phi \vee \mathrm{f}_\xi(n) - \Phi \vee \mathrm{f}_\xi(N_0) + 2\varepsilon^2\underline{N}\right)\\
=\ & -\frac{\Phi \vee \mathrm{f}_\xi(n) + 2\varepsilon^2\underline{N} - 1}{\Phi \vee \mathrm{f}_\xi(n) + 2\varepsilon^2\underline{N}}\left(\Phi \vee \mathrm{f}_\xi(n) - \Phi \vee \mathrm{f}_\xi(N_0) + 2\varepsilon^2\underline{N}\right)\\
=\ & -\left[\Phi \vee \mathrm{f}_\xi(n) + 2\varepsilon^2\underline{N} - 1\right] + \frac{\Phi \vee \mathrm{f}_\xi(N_0)}{\Phi \vee \mathrm{f}_\xi(n) + 2\varepsilon^2\underline{N}}\\
=\ & -\left[\Phi \vee \mathrm{f}_\xi(n) + 2\varepsilon^2\underline{N}\right] + 1 + \frac{\Phi \vee \mathrm{f}_\xi(N_0)}{\Phi \vee \mathrm{f}_\xi(n) + 2\varepsilon^2\underline{N}}\\
\leqslant\ & -\left[\Phi \vee \mathrm{f}_\xi(n) + 2\varepsilon^2\underline{N}\right] + 2\,.
\end{aligned} \tag{5.32}$$

Thus, by combining Equations (5.29), (5.31) and (5.32) we get

$$\begin{aligned}
&\mathbb{P}_\nu\left(\bigcup_{\substack{t\geqslant 1\\ \forall a\in\mathcal{A}',\ \widehat{\mu}_a(t)<\mu_a-\varepsilon\\ \underline{M}\leqslant N_{\mathcal{A}'}(t)\leqslant|\mathcal{A}'|\mathrm{f}_\xi(n)/2\varepsilon^2}}\ \sum_{a\in\mathcal{A}'} N_a(t)\mathrm{kl}(\widehat{\mu}_a(t)|\mu_a-\varepsilon) + \mathrm{f}_\xi(N_{\mathcal{A}'}(t)) \geqslant \mathrm{f}_\xi(n)\right)\\
\leqslant\ & \frac{e^{|\mathcal{A}'|+2}}{|\mathcal{A}'|^{|\mathcal{A}'|}}\log(n)^{|\mathcal{A}'|+1}\left[\Phi \vee \mathrm{f}_\xi(n) + 2\varepsilon^2\underline{N}\right]^{|\mathcal{A}'|+1} e^{-\left[\Phi\vee\mathrm{f}_\xi(n)+2\varepsilon^2\underline{N}\right]}\,,
\end{aligned} \tag{5.33}$$

where $\underline{N} = \underline{M} \vee \dfrac{\mathrm{f}_\xi(n)}{1-\log(1-\max_{a\in\mathcal{A}'}\mu_a)}$. □

We remind in the following lemma a result from Magureanu et al. (2014) based on multivariate stochastic ordering.

Lemma 17 (Stochastic ordering)**.** *Let $M>0$, $A\geqslant 2$. Let $Z\in\mathbb{R}^A$ be a random variable such that for all $\zeta\in\mathbb{R}_+^A$,*

$$\mathbb{P}(Z \geqslant \zeta) \leqslant \exp\left(-M\sum_{a=1}^{A}\zeta_a\right).$$

Then for all $z\geqslant A/M$,

$$\mathbb{P}\left(\sum_{a=1}^{A} Z_a \geqslant z\right) \leqslant \left(\frac{Mze}{A}\right)^A e^{-Mz}\,.$$

Chapter 6

Routine Bandits

6.1 The Routine Bandit Setting

A routine bandit problem is specified by a time horizon $T \geqslant 1$ and a finite set of distributions $\nu = (\nu_b)_{b\in\mathcal{B}}$ with means $(\mu_{a,b})_{a\in\mathcal{A},b\in\mathcal{B}}$, where $\mathcal{A}$ is a finite set of arms and $\mathcal{B}$ is a finite set of bandit configurations. Each $b \in \mathcal{B}$ can be seen as a classical multi-armed bandit problem defined by $\nu_b = (\nu_{a,b})_{a\in\mathcal{A}}$. At each period $h \geqslant 1$ and for all time steps $t \in [\![1, T]\!]$, the learner deals with a bandit $b^h_\star \in \mathcal{B}$ and chooses an arm $a^h_t \in \mathcal{A}$, based only on the past. The learner then receives and observes a reward $X^h_t \sim \nu_{a^h_t,b^h_\star}$. The goal of the learner is to maximize the expected sum of rewards received over time (up to some unknown number of periods $H \geqslant 1$). The distributions are unknown, which makes the problem non-trivial. The optimal strategy therefore consists in playing repeatedly on each period h, an optimal arm $a^h_\star \in \operatorname{argmax}_{a\in\mathcal{A}} \mu_{a,b^h_\star}$, which has mean $\mu^h_\star = \mu_{a^h_\star,b^h_\star}$. The goal of the learner is equivalent to minimizing the cumulative *regret* with respect to an optimal strategy:

$$R(\nu, H, T) = \mathbb{E}_\nu \left[\sum_{h=1}^{H} \sum_{t=1}^{T} \left(\mu^h_\star - X^h_t \right) \right] . \tag{6.1}$$

Related works One of the closest setting to routine bandits is the sequential transfer scenario Ghesh-laghi Azar et al. (2013), where the cardinality $|\mathcal{B}|$ and quantities H and T are known ahead of time, and the instances in $\mathcal{B}$ are either known perfectly or estimated with known confidence. Routine bandits also bear similarity with clustering bandits Gentile et al. (2014), a contextual bandit setting Langford and Zhang (2007) where contexts can be clustered into finite (unknown) clusters. While both settings are recurring bandit problems, routine bandits assume no information on users (including their number) but users are recurring for several iterations of interaction, while clustering bandits assume that each user is seen only once, but is characterized by features such that they can be associated with previously seen users. Finally, latent bandits Maillard and Mannor (2014) consider the less structured situation when the learner faces a possibly different user at every time.

Assumptions and working conditions The configuration ν, the set of bandits $\mathcal{B}$, and the sequence of bandits $(b^h_\star)_{h\geqslant 1}$ are *unknown* (in particular $|\mathcal{B}|$ and the identity of user $b^h_\star$ are unknown to the learner at time t). The learner only knows that $\nu \in \mathcal{D}$, where $\mathcal{D}$ is a given set of bandit configurations. In order to leverage information from the bandit instances encountered, we should consider that bandits reoccur. We denote by $\beta^h_b = \sum_{h'=1}^{h} \mathbb{I}_{\{b^{h'}_\star = b\}}/h$ the frequency of bandit $b \in \mathcal{B}$ at period $h \geqslant 1$ and assume $\beta^H_b > 0$. The next two assumptions respectively allow for two bandit instances b and b' to be distinguishable from their means when $b \neq b'$ and show consistency in their optimal strategy when $b = b'$.

Assumption 10 (Separation). *Let us consider* $\gamma_\nu := \min_{b\neq b'} \min_{a\in\mathcal{A}} \{|\mu_{a,b} - \mu_{a,b'}|, 1\}$. *We assume* $\gamma_\nu > 0$.

Assumption 11 (Unique optimal arm)**.** *Each bandit $b \in \mathcal{B}$ has a unique optimal arm $a_b^\star$.*

Assumption 11 is standard. Finally, we consider normally-distributed rewards. Although most of our analysis (e.g., concentration) would extend to exponential families of dimension 1, Assumption 12 increases readability of the statements.

Assumption 12 (Gaussian arms)**.** *The set $\mathcal{D}$ is the set of bandit configurations such that for all bandit $b \in \mathcal{B}$, for all arm $a \in \mathcal{A}$, $\nu_{a,b}$ is a one-dimensional Gaussian distribution with mean $\mu_{a,b} \in \mathbb{R}$ and variance $\sigma^2 = 1$.*

For $\nu \in \mathcal{D}$, we define for an arm $a \in \mathcal{A}$ and a bandit $b \in \mathcal{B}$ their gap $\Delta_{a,b} = \mu_b^\star - \mu_{a,b}$ and their total number of pulls over H periods $N_{a,b}(H,T) = \sum_{h=1}^{H} \sum_{t=1}^{T} \mathbb{I}_{\{a_t^h = a, b_\star^h = b\}}$. An arm is optimal for a bandit if their gap is equal to zero and sub-optimal if it is positive. Thanks to the chain rule, the regret rewrites as

$$R(\nu, H, T) = \sum_{b \in \mathcal{B}} \sum_{a \neq a_b^\star} \mathbb{E}_\nu[N_{a,b}(H,T)]\, \Delta_{a,b}\,. \tag{6.2}$$

Remark 15 (Fixed horizon time)**.** *We assume the time horizon T to be the same for all periods $h \in [\![1\,,H]\!]$ out of clarity of exposure of the results and simplified definition of consistency (Definition 3). Considering a different time T_h for each h would indeed require a substantial rewriting of the statements (e.g. think of the regret lower bound), which we believe hinders readability and comparison to classical bandits.*

We conclude this section by adapting for completeness the known lower bound on the regret Lai and Robbins (1985); Agrawal et al. (1989); Graves and Lai (1997) for *consistent* strategies to the routine bandit setting. We defer the proof to Appendix C.1.

Definition 3 (Consistent strategy)**.** *A strategy is H-consistent on $\mathcal{D}$ if for all configuration $\nu \in \mathcal{D}$, for all bandit $b \in \mathcal{B}$, for all sub-optimal arm $a \neq a_\star^b$, for all $\alpha > 0$,*

$$\lim_{T \to \infty} \mathbb{E}_\nu \left[\frac{N_{a,b}(H,T)}{N_b(H,T)^\alpha} \right] = 0\,,$$

where $N_b(H,T) = \beta_b^H HT$ is the number of time steps the learner has dealt with bandit b.

Proposition 5 (Lower bounds on the regret)**.** *Let us consider a consistent strategy. Then, for all configuration $\nu \in \mathcal{D}$, it must be that*

$$\liminf_{T \to \infty} \frac{R(\nu, H, T)}{\log(T)} \geqslant c_\nu^\star := \sum_{b \in \mathcal{B}} \sum_{a \neq a_b^\star} \frac{\Delta_{a,b}}{\mathrm{KL}(\mu_{a,b} | \mu_b^\star)}\,,$$

where $\mathrm{KL}(\mu|\mu') = (\mu' - \mu)^2 / 2\sigma^2$ denotes the Kullback-Leibler divergence between one-dimensional Gaussian distributions with means $\mu, \mu' \in \mathbb{R}$ and variance $\sigma^2 = 1$.

This lower bound differs (it is larger) from structured lower bound that can exclude some set of arms, as in Agrawal et al. (1989); Maillard and Mannor (2014) using prior knowledge on $\mathcal{B}$, which here is not available. On the other hand, we remark that the right hand side of the bound does not depend on H, which suggests that one at least asymptotically, one can learn from the recurring bandits. In the classical bandit setting, lower bounds on the regret Lai and Robbins (1985) have inspired the design of the well-known `KLUCB` Garivier and Cappé (2011) algorithm. In the next section, we build on this optimal strategy to propose a variant for the routine bandit.

6.2 The `KLUCB-RB` Strategy

Given the current period h, the general idea of this optimistic strategy consists in aggregating observations acquired in previous periods $1 \ldots h-1$ where bandit instances are tested to be the same as the current bandit $b_\star^h$. To achieve this, `KLUCB-RB` relies both on concentration of observations gathered in previous periods and the consistency of the allocation strategy between different periods.

Notations The number of pulls, the sum of the rewards and the empirical mean of the rewards from the arm a in period $h \geqslant 1$ at time $t \geqslant 1$, are respectively denoted by $N_a^h(t) = \sum_{s=1}^t \mathbb{I}_{\{a_s^h = a\}}$, $S_a^h(t) = \sum_{s=1}^t \mathbb{I}_{\{a_s^h = a\}} X_s^h$ and $\widehat{\mu}_a^h(t) = S_a^h(t)/N_a^h(t)$ if $N_a^h(t) > 0$, 0 otherwise.

Strategy For each period $h \geqslant 1$ we compute an empirical best arm for bandit $b_\star^h$ as the arm with maximum number of pulls in this period: $\overline{a}_\star^h \in \operatorname*{argmax}_{a \in \mathcal{A}} N_a^h(T)$.[1] Similarly, in the current period $h \geqslant 1$, for each time step $t \in [\![1, T]\!]$, we consider an arm with maximum number of pulls: $\overline{a}_t^h \in \operatorname*{argmax}_{a \in \mathcal{A}} N_a^h(t)$ (arbitrarily chosen). At each period $h \in [\![2, H]\!]$ each arm is pulled once. Then at each time step $t \geqslant |\mathcal{A}| + 1$, in order to possibly identify the current bandit $b_\star^h$ with some bandits $b_\star^k$ from a previous period $k \in [\![1, h-1]\!]$, we introduce for all arm $a \in \mathcal{A}$, the test statistics

$$Z_a^{k,h}(t) = \infty \cdot \mathbb{I}_{\{\overline{a}_t^h \neq \overline{a}_\star^k\}} + \left|\widehat{\mu}_a^h(t) - \widehat{\mu}_a^k(T)\right| - \mathrm{d}\big(N_a^h(t), \delta^h(t)\big) - \mathrm{d}\big(N_a^k(T), \delta^h(t)\big)\,, \tag{6.3}$$

where the deviation for $n \geqslant 1$ pulls with probability $1-\delta$, for $\delta > 0$, and probability $\delta^h(t)$ are, respectively,

$$\mathrm{d}(n, \delta) = \sqrt{2\left(1 + \frac{1}{n}\right)\frac{\log\left(\sqrt{n+1}/\delta\right)}{n}} \qquad \delta^h(t) = \frac{1}{4\,|\mathcal{A}|} \times \frac{1}{h-1} \times \frac{1}{t(t+1)}\,.$$

The algorithm finally computes the test

$$\mathrm{T}^{k,h}(t) \coloneqq \max_{a \in \mathcal{A}} Z_a^{k,h}(t) \leqslant 0\,. \tag{6.4}$$

After t rounds in current period h, the previous bandit $b_\star^k$ is suspected of being the same as $b_\star^h$ if the test $\mathrm{T}^{k,h}(t)$ is true. From Eq. 6.3, we note that this requires the current mostly played arm to be the same as the arm that was mostly played in period k, which happens if there is consistency in the allocation strategy for both periods under Assumption 11. We then define aggregated numbers of pulls and averaged means: For all arm $a \in \mathcal{A}$, for all period $h \geqslant 1$, for all time step $t \geqslant 1$,

$$\overline{N}_a^h(t) \coloneqq N_a^h(t) + \sum_{k=1}^{h-1} \mathbb{I}_{\{\mathrm{T}^{k,h}(t)\}} N_a^k(T), \quad \overline{K}_t^h \coloneqq \sum_{k=1}^{h-1} \mathbb{I}_{\{\mathrm{T}^{k,h}(t)\}},$$
$$\overline{S}_a^h(t) \coloneqq S_a^h(t) + \sum_{k=1}^{h-1} \mathbb{I}_{\{\mathrm{T}^{k,h}(t)\}} S_a^k(T), \quad \overline{\mu}_a^h(t) = \overline{S}_a^h(t)/\overline{N}_a^h(t)\,.$$

and follow a `KLUCB` strategy by defining the index of arm $a \in \mathcal{A}$ in period $h \geqslant 1$ at time step $t \geqslant 1$ as

$$u_a^h(t) = \min\left\{U_a^h(t)\,, \overline{U}_a^h(t)\right\}, \tag{6.5}$$

where

$$U_a^h(t) \coloneqq \widehat{\mu}_a^h(t) + \sqrt{\frac{2f(t)}{N_a^h(t)}}\,, \tag{6.6}$$

$$\overline{U}_a^h(t) \coloneqq \overline{\mu}_a^h(t) + \sqrt{\frac{2f\left(\overline{K}_t^h T + t\right)}{\overline{N}_a^h(t)}}\,, \tag{6.7}$$

[1] ties are broken arbitrarily

with the function f being chosen, following Cappé et al. (2013) for classical bandits, as

$$f(x) \coloneqq \log(x) + 3\log\log(\max\{e, x\}), \forall x \geqslant 1\,.$$

One recognizes that Eq. 6.6 corresponds to the typical KLUCB upper bound for Gaussian distributions. The resulting KLUCB−RB strategy is summarized in Algorithm 10.

Algorithm 10 KLUCB−RB

Initialization (period $h=1$): follow a KLUCB strategy for bandit $b^1_\star$.
for period $h \geqslant 2$ **do**
  Pull each arm once
  for time step $t \in [\![|\mathcal{A}|, T-1]\!]$ **do**
    Compute for each previous period $k \in [\![1, h-1]\!]$ the test $\mathrm{T}^{k,h}(t) \coloneqq \max_{a\in\mathcal{A}} Z_a^{k,h}(t) \leqslant 0$
    Aggregate data from periods with positive test and compute for each arm $a \in \mathcal{A}$ the index $u_a^h(t)$ according to equations (6.5)-(6.6)-(6.7).
    Pull an arm with maximum index $a_{t+1}^h \in \mathrm{argmax}_{a\in\mathcal{A}}\, u_a^h(t)$
  end for
end for

Theoretical guarantees The next result shows that the number of sub-optimal pulls done by KLUCB−RB is upper-bounded in a near-optimal way.

Theorem 5 (Upper bounds)**.** *Let us consider a routine bandit problem specified by a set of Gaussian distributions $\nu \in \mathcal{D}$ and a number of periods $H \geqslant 1$. Then under KLUCB−RB strategy, for all $0 < \varepsilon < \varepsilon_\nu$, for all bandit $b \in \mathcal{B}$, for all sub-optimal arm $a \neq a_b^\star$,*

$$\mathbb{E}_\nu[N_{a,b}(H,T)] \leqslant \frac{f(\beta_b^H HT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\mu_b^\star)} + \sum_{h=1}^{H} \mathbb{I}_{\{b_\star^h = b\}}\left[\tau_\nu^h + 4\,|\mathcal{A}|\left(\frac{1}{\varepsilon^2}+1\right)\left(5+\frac{8h\,f(hT)}{T\,\mathrm{KL}(\mu_{a,b}+\varepsilon|\mu_b^\star)}\right)\right],$$

where, for all period $h \geqslant 2$, $\tau_\nu^h \coloneqq 2\varphi\big(8|\mathcal{A}|\big[\varepsilon_\nu^{-2} + 65\gamma_\nu^{-2}\log\big(128|\mathcal{A}|(4h)^{1/3}\gamma_\nu^{-2}\big)\big]\big)$, $\varphi : x \geqslant 1 \mapsto x\log(x)$, $\varepsilon_\nu = \min_{b\in\mathcal{B}} \min_{a\neq a_b^\star} \Delta_{a,b}/2$ and $\gamma_\nu = \min_{b\neq b'} \min_{a\in\mathcal{A}} \{|\mu_{a,b} - \mu_{a,b'}|, 1\}$.

This implies that the dependency on the time horizon T in these upper bounds is asymptotically optimal with regard to the lower bound on the regret given in Proposition 5. From Eq. 6.2, by considering the case when the time horizon T tends to infinity, we deduce that KLUCB−RB achieves asymptotic optimality.

Corollary 4 (Asymptotic optimality)**.** *With the same notations and under the assumptions as in Theorem 5, KLUCB−RB achieves*

$$\limsup_{T\to\infty} \frac{R(\nu, H, T)}{\log(T)} \leqslant c_\nu^\star\,,$$

where $c_\nu^\star$ is defined as in Proposition 5.

For comparison, let us remark that under the strategy that runs a separate KLUCB type strategy for each period, the regret normalized by $\log(T)$ asymptotically scales as $H\sum_{b\in\mathcal{B}} \beta_b^H \sum_{a\neq a_b^\star} \Delta_{a,b}/\mathrm{KL}(\mu_{a,b}|\mu_b^\star)$. KLUCB−RB strategy then performs better than this naive strategy by a factor of the order of $H/|\mathcal{B}|$. Also, up to our knowledge, this result is the first showing provably asymptotic optimal regret guarantee in a setting when an

agent attempts at transferring information from past to current bandits without contextual information. In the related but different settings considered in Gheshlaghi Azar et al. (2013); Gentile et al. (2014); Maillard and Mannor (2014), only logarithmic regret was shown, however asymptotic optimality was not proved for the considered strategies. Also, let us remind that $|\mathcal{B}|$ does not need to be known ahead of time by the KLUCB-RB algorithm.

6.3 Sketch of Proof

This section contains a sketch of proof for Theorem 5. We refer to Appendix C.2 for more insights and detailed derivations. The first preoccupation is to ensure that KLUCB-RB is a consistent strategy. This is achieved by showing that KLUCB-RB aggregates observations that indeed come from the same bandits with high probability. In other words, we want to control the number of previously encountered bandits falsely identified as similar to the current one.

Definition 4 (False positive)**.** *At period $h \geqslant 2$ and step $t \geqslant 1$, a previous period $k \in [\![1, h-1]\!]$ is called a false positive if the test $\mathrm{T}^{k,h}(t)$ is true while previous bandit $b_\star^k$ differs from current bandit $b_\star^h$.*

Combining the triangle inequality and time-uniform Gaussian concentration inequalities (see e.g., Abbasi-Yadkori et al. (2011)), we prove necessary condition for having $Z_a^{k,h}(t) \leqslant 0$ for some arm $a \in \mathcal{A}$ at current period h and time step t, while having $b_\star^k \neq b_\star^h$.

Lemma 18 (Condition for false positives)**.** *If there exists a* false positive *at period $h \geqslant 2$ and time step $t > |\mathcal{A}|$, then with probability $1-1/t(t+1)$, it must be that*

$$\min_{k \in [\![1,h-1]\!] : b_\star^k \neq b_\star^h} \min_{a \in \mathcal{A}} \left| \mu_{a,b_\star^h} - \mu_{a,b_\star^k} \right| \leqslant 4\mathrm{d}\left(\frac{t}{|\mathcal{A}|}, \delta^h(t) \right) .$$

The proof of this key result is provided in Appendix C.2.1. It relies on time-uniform concentration inequalities. We now introduce a few quantities.
Let us first consider at period $h \geqslant 2$ the time step

$$t_\nu^h := \max \left\{ t \geqslant |\mathcal{A}| : \ \gamma_\nu \leqslant 4\mathrm{d}\left(\frac{t}{|\mathcal{A}|}, \delta^h(t) \right) \right\} + 1 , \tag{6.8}$$

beyond which there is no false positives with high probability. We define for all $a \neq a_\star^h$, for all $0 < \varepsilon < \varepsilon_\nu := \min_{b \in \mathcal{B}} \min_{a \neq a_b^\star} \{\Delta_{a,b}, 1\}/2$ the subsets of times when there is a false positive

$$\mathcal{T}_a^h := \left\{ t \geqslant t_\nu^h : \ a_{t+1}^h = a \text{ and } \mathcal{K}_+^h(t) \neq \mathcal{K}_\star^h(t) \right\} \qquad \mathcal{T}^h := \bigcup_{a \neq a_\star^h} \mathcal{T}_a^h , \tag{6.9}$$

where we introduced for convenience the sets $\mathcal{K}_+^h := \left\{ k \in [\![1, h-1]\!] : \mathrm{T}^{k,h}(t) \text{ is true} \right\}$ and $\mathcal{K}_\star^h(t) := \left\{ k \in [\![1, h-1]\!] : b_\star^k =$
We also consider the times when the mean of the current pulled arm is poorly estimated or the best arm $a_\star^h$ is below its mean (either for the current period or by aggregation) and define

$$\mathcal{C}_{a,\varepsilon}^h := \left\{ t \geqslant 1 : a_{t+1}^h = a \text{ and } \left(\left| \widehat{\mu}_a^h(t) - \mu_a^h \right| > \varepsilon \text{ or } u_{a_\star^h}^h(t) = U_{a_\star^h}^h(t) < \mu_\star^h \right) \right\}$$

$$\mathcal{C}_\varepsilon^h := \bigcup_{a \neq a_\star^h} \mathcal{C}_{a,\varepsilon}^h \tag{6.10}$$

$$\overline{\mathcal{C}}_{a,\varepsilon}^h := \mathcal{T}_a^h \cup \left\{ t \geqslant t_\nu^h : t \notin \mathcal{T}^h, a_{t+1}^h = a \text{ and } \left(\left| \overline{\mu}_a^h(t) - \mu_a^h \right| > \varepsilon \text{ or } u_{a_\star^h}^h(t) = \overline{U}_{a_\star^h}^h(t) < \mu_\star^h \right) \right\}$$

$$\overline{\mathcal{C}}_{\varepsilon}^{h} \coloneqq \bigcup_{a \neq a_{\star}^{h}} \overline{\mathcal{C}}_{a,\varepsilon}^{h} \,. \tag{6.11}$$

The size of these (bad events) sets can be controlled by resorting to concentration arguments. The next lemma borrows elements of proof from Combes and Proutiere (2014b) for the estimation of the mean of current pulled arm and Cappé et al. (2013) for the effectiveness of the upper confidence bounds on the empirical means of optimal arms. We adapt these arguments to the routine-bandit setup, and provide additional details in the appendix.

Lemma 19 (Bounded subsets of times)**.** *For all period $h \geqslant 2$, for all arm $a \in \mathcal{A}$, for all $0 < \varepsilon < \varepsilon_{\nu}$,*

$$\mathbb{E}_{\nu}\left[\left|\mathcal{T}^{h}\right|\right] \leqslant 1 \qquad \mathbb{E}_{\nu}\left[\left|\mathcal{C}_{a,\varepsilon}^{h}\right|\right] \leqslant 4\varepsilon^{-2} + 2 \qquad \mathbb{E}_{\nu}\left[\left|\overline{\mathcal{C}}_{a,\varepsilon}^{h}\right|\right] \leqslant 4\varepsilon^{-2} + 3 \,.$$

By definition of the index (Eq. 6.7), we have

$$\begin{aligned} \forall t > |\mathcal{A}| \,, \qquad & N_{a}^{h}(t) \mathrm{KL}\big(\widehat{\mu}_{a}^{h}(t) \big| U_{a}^{h}(t)\big) = f(t) \\ & \overline{N}_{a}^{h}(t) \mathrm{KL}\big(\overline{\mu}_{a}^{h}(t) \Big| \overline{U}_{a}^{h}(t)\Big) = f\Big(\overline{K}_{t}^{h} T + t\Big) \,. \end{aligned}$$

We then provide logarithmic upper bounds on the aggregated number of pulls $\overline{N}_{a}^{h}(t)$ to deduce the consistency of `KLUCB-RB` strategy. The following non-trivial result combines standard techniques with the key mechanism of the algorithm.

Lemma 20 (Consistency)**.** *Under `KLUCB-RB` strategy for all period $h \geqslant 2$, for all $0 < \varepsilon < \varepsilon_{\nu}$, for all sub-optimal arm $a \neq a_{\star}^{h}$, for all $t > |\mathcal{A}|$ such that $a_{t+1}^{h} = a$,*

$$\text{if } t \notin \mathcal{C}_{a,\varepsilon}^{h}, \; N_{a}^{h}(t) \leqslant \frac{f(t)}{\mathrm{KL}(\mu_{a}^{h} + \varepsilon | \mu_{\star}^{h})}, \quad \text{if } t \geqslant t_{\nu}^{h} \text{ and } t \notin \overline{\mathcal{C}}_{a,\varepsilon}^{h}, \overline{N}_{a}^{h}(t) \leqslant \frac{f\big(\underline{K}_{t}^{h} T + t\big)}{\mathrm{KL}(\mu_{a}^{h} + \varepsilon | \mu_{\star}^{h})},$$

where $\underline{K}_{t}^{h} \coloneqq \min\Big\{\overline{K}_{t}^{h}, \beta_{b_{\star}^{h}}^{h-1}(h-1)\Big\}$. In particular this implies

$$\forall t \geqslant 1, \forall a \neq a_{\star}^{h}, \quad N_{a}^{h}(t) \leqslant \frac{f(t)}{\mathrm{KL}(\mu_{a}^{h} + \varepsilon | \mu_{\star}^{h})} + \left|\mathcal{C}_{a,\varepsilon}^{h}\right| + N_{a}^{h}(|\mathcal{A}|+1) \;,$$

where $N_{a}^{h}(|\mathcal{A}|+1) \leqslant 2$ and $\mathbb{E}_{\nu}\left[\left|\mathcal{C}_{a,\varepsilon}^{h}\right|\right] \leqslant 4\varepsilon^{-2} + 2$.

Thanks to Eq. 6.5 that involves the minimum of the aggregated index $\overline{U}_{a}^{h}(t)$ on past episodes and (not aggregated) indexes $U_{a}^{h}(t)$ for the current epoch, the proof proceeds by considering the appropriate sets of time, namely $t \notin \mathcal{C}_{a,\varepsilon}^{h}$ or $t \notin \overline{\mathcal{C}}_{a,\varepsilon}^{h}$ depending on the situation. In particular, we get for the considered a that the maximum index $u_{a}^{h}(t)$ is either greater than $u_{a_{\star}^{h}}^{h}(t) = U_{a_{\star}^{h}}^{h}(t)$ or $u_{a_{\star}^{h}}^{h}(t) = \overline{U}_{a_{\star}^{h}}^{h}(t)$, which in turns enable to have a control either on $N_{a}^{h}(t)$ or $\overline{N}_{a}^{h}(t)$. In order to obtain the last statement, it essentially remains to consider the maximum time $t' \in [\![|\mathcal{A}|+1; t]\!]$ such that $a_{t'+1}^{h} = a$ and $t' \notin \mathcal{C}_{a,\varepsilon}^{h}$.
In order to be asymptotically optimal (in the sense of Corollary 4), the second preoccupation is to ensure with high probability that we aggregate all of the observations coming from current bandit $b_{\star}^{h}$ when computing the indexes. From the definition of $\mathcal{T}^{h}$ (Eq. 6.9) and Lemma 19, this amounts to ensure that the current most pulled arm and the most pulled arms of previous periods are the optimal arms of the corresponding periods with high probability. By using the consistency of `KLUCB-RB`, we prove necessary conditions for the most pulled arms being different from the optimal ones.

Lemma 21 (Most pulled arms). *For all period $h \geqslant 2$, for all $0 < \varepsilon < \varepsilon_\nu$, for all $t \geqslant t_\nu^h$ such that $t \notin \mathcal{T}^h$ and $\overline{a}_t^h \neq a_\star^h$,*

$$\frac{t+|\mathcal{K}_\star^h(t)|T}{2} - \Big(f(t)+\big|\mathcal{K}_\star^h(t)\big|f(T)\Big)\sum_{a\neq a_\star^h}\frac{1}{\mathrm{KL}(\mu_a^h+\varepsilon|\mu_\star^h)} - \Big(1+\big|\mathcal{K}_\star^h(t)\big|\Big)|\mathcal{A}| \leqslant \sum_{k\in\mathcal{K}_\star^h(t)\cup\{h\}}\big|\mathcal{C}_\varepsilon^k\big|.$$

Let us remind that $\mathcal{K}_\star^h(t)$, defined after Lemma 18, counts the previous phases before h facing the same bandit as the current one, and for which the most-played arm until then agree. Then, by combining Lemma 20 and Lemma 21 we obtain randomized upper bounds on the number of pulls of sub-optimal arms.

Proposition 6 (Randomized upper bounds). *Under* `KLUCB-RB` *strategy, for all bandit $b \in \mathcal{B}$, for all sub-optimal arm $a \neq a_b^\star$, for all $0 < \varepsilon < \varepsilon_\nu$,*

$$\begin{aligned} N_{a,b}(H,T) \;\leqslant\; & \frac{f(\beta_b^H HT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\mu_b^\star)} \\ & + \sum_{h=1}^{H}\mathbb{I}_{\{b_\star^h=b\}}\left[T_{\nu,\varepsilon}^h + 4\big|\mathcal{C}_\varepsilon^h\big| + \big|\overline{\mathcal{C}}_\varepsilon^h\big| + \frac{f(hT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\mu_b^\star)}\sum_{k=1}^{h}\frac{8\big|\mathcal{C}_\varepsilon^k\big|}{T} + \mathbb{I}_{\{T\in\mathcal{T}^k\}}\right], \end{aligned}$$

where $T_{\nu,\varepsilon}^h := \max\left\{t \geqslant t_\nu^h : \frac{t}{4} - \sum_{a\neq a_\star^h}\frac{f(t)}{\mathrm{KL}(\mu_a^h+\varepsilon|\mu_\star^h)} \leqslant |\mathcal{A}|\right\} + 1$ *for $h \geqslant 2$, with t_ν^h defined in Eq. (6.8).*

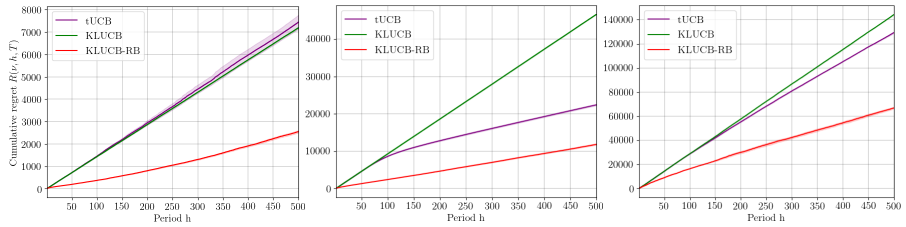
We prove Theorem 5 by averaging the randomized upper bounds from Proposition 6.

6.4 Numerical Experiments

We now perform experiments to illustrate the performance of the proposed `KLUCB-RB` under different empirical conditions. We compare `KLUCB-RB` with a baseline strategy which consists in using a `KLUCB` that restarts from scratch at every new period, that is the default strategy when no information (features) is provided to share information across periods. We also include a comparison with the sequential transfer algorithm `tUCB` Gheshlaghi Azar et al. (2013) which constitutes interesting baseline to compare with, since it transfers the knowledge of past periods to minimize the regret in a very similar context. Through the periods $h \in [\![1, H]\!]$, `tUCB` incrementally estimates the mean vectors by the Robust Tensor Power method Anandkumar et al. (2013, 2014), then yielding a deviation of rate $\mathcal{O}(1/\sqrt{h})$ over the empirical means. Thus, it needs to know in advance the total number of instances $|\mathcal{B}|$. Besides the RTP method requires the mean vectors to be linearly independent mutually, which forces the number of arms $|\mathcal{A}|$ to be larger than $|\mathcal{B}|$, while `KLUCB-RB` can tackle this kind of distributions. The next comparisons between `KLUCB-RB` and `tUCB` will mainly illustrate the ability of the former to make large profits from the very first periods, while the later needs to get a sufficiently high confidence over the models estimates before beginning to use knowledge from the previous periods.
All experiments are repeated 100 times. Sequence $(b^h)_{1\leqslant h\leqslant H}$ is chosen randomly each time. All the different strategies are compared based on their cumulative regret (Eq. 6.1). Additional experiments are provided in Appendix C.3.

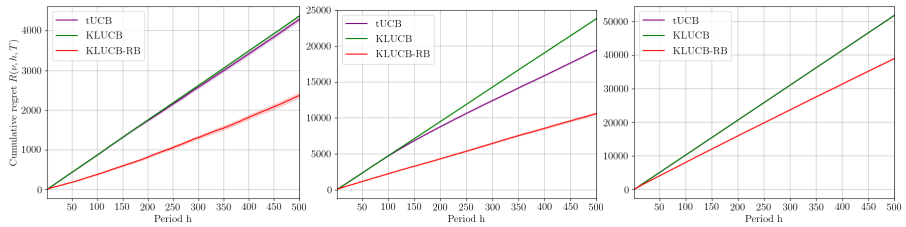
6.4.1 More Arms than Bandits: A Beneficial Case

We first investigate how Assumption 10 can be relaxed in practice. Indeed `KLUCB-RB` is designed such that only data from previous periods $k < h$ for which the most pulled arm $\bar{a}_\star^k$ is the same as the current most pulled arm $\bar{a}_t^h$ may be aggregated. Consequently, let us define $\gamma_\nu^\star := \min_{b\neq b'}\min_{a\in\mathcal{A}^\star}|\mu_{a,b} - \mu_{a,b'}|$ with $\mathcal{A}^\star$ being the set of arms optimal on at least one instance $b \in \mathcal{B}$. Assuming that `KLUCB-RB` converges to the optimal action in a



(a) $|\mathcal{A}| = 2$ (b) $|\mathcal{A}| = 10$ (c) $|\mathcal{A}| = 50$

Figure 6.1: Cumulative regret of KLUCB, KLUCB-RB and tUCB along $H = 500$ periods of $\boldsymbol{T = 10^3}$ **rounds**, for different action sets.



(a) $|\mathcal{A}| = 2$ (b) $|\mathcal{A}| = 10$ (c) $|\mathcal{A}| = 50$

Figure 6.2: Cumulative regret of KLUCB, KLUCB-RB and tUCB along $H = 500$ periods of $\boldsymbol{T = 100}$ **rounds**, for different action sets.

given period, it is natural in practice to relax Assumption 10 from $\gamma_\nu > 0$ to $\gamma_\nu^\star > 0$. Let us consider a routine two-bandit setting $\mathcal{B} = \{b_1, b_2\}$ with actions $\mathcal{A}$ such that

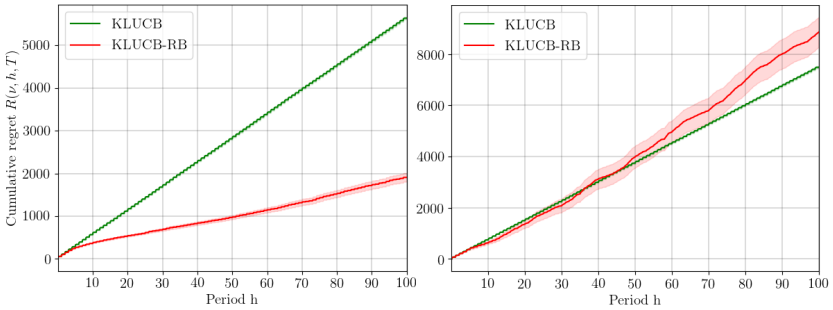
$$b_1 : (\mu_{1,b_1}, \mu_{2,b_1}) = (\frac{\Delta}{2}, -\frac{\Delta}{2}) \qquad \text{and} \qquad \forall a \geqslant 3,\ \mu_{a,b_1} = \mu \tag{6.12}$$

$$b_2 : (\mu_{1,b_2}, \mu_{2,b_2}) = (\frac{\Delta}{2} - \gamma, -\frac{\Delta}{2} + \gamma) \qquad \text{and} \qquad \forall a \geqslant 3,\ \mu_{a,b_2} = \mu, \tag{6.13}$$

with $\mu = -\frac{\Delta}{2}$, and $\gamma = 0.85\Delta$, and where $\Delta = 10\sqrt{\frac{\log(HT)}{T}}$ is set to accomodate the convergence of KLUCB in the experiment. Note that Assumption 10 is not satisfied anymore since $\gamma_\nu = 0$, but that $\gamma_\nu^\star = \gamma$. Fig. 6.1 shows the average cumulative regret with one standard deviation after $H = 500$ periods of $T = 10^3$ rounds on settings where $|\mathcal{A}^\star| = 2$ and $|\mathcal{A}| \geqslant 2$.
We observe that KLUCB-RB can largely benefit from relying on previous periods when the number of arms exceeds the number of optimal arms, which naturally happens when $|\mathcal{A}| > |\mathcal{B}|$. This can also happen for $|\mathcal{A}| \leqslant |\mathcal{B}|$ if several bandits $b \in \mathcal{B}$ share the same optimal arm. Besides, Fig. 6.2 shows a remake of the same experiment, that is $\Delta = 10\sqrt{\frac{\log(H \times 10^3)}{10^3}}$, where the number of rounds per period is decreased from 10^3 to $T = 100$. We can see that KLUCB-RB still yields good satisfying performances, although T is not large enough to enable a sure identification at each period of the current instance.

6.4.2 Increasing the Number of Bandit Instances

We now consider experiments where we switch among $|\mathcal{B}| = 5$ four-armed bandits. This highlights the kind of settings which may cause more difficulties to KLUCB-RB in distinguishing the different instances: the lesser is the number of arms $|\mathcal{A}|$ compared to the number of bandits $|\mathcal{B}|$, the harder it should be for KLUCB-RB to distinguish efficiently the different instances, in particular when the separation gaps are tight. Let us precise



(a) Setting $\nu^{(1)}$ (b) Setting $\nu^{(2)}$

Figure 6.3: Cumulative regret of KLUCB and KLUCB−RB along $H = 100$ periods of $T = 5000$ rounds over three generated settings of $|\mathcal{B}| = 5$ bandit instances with $|\mathcal{A}| = 4$ arms per instance.

that tUCB cannot be tested on such settings, where the number of models $|\mathcal{B}|$ exceeds the number of arms $|\mathcal{A}|$, since it requires that the mean vectors $(\mu_{a,b})_{a\in\mathcal{A}}$ for all b in $\mathcal{B}$ to be linearly independent.
Generating specific settings is far more complicated here than in cases where $|\mathcal{B}| = 2$ because of the intrinsic dependency between regret gaps $(\Delta_{a,b})_{a\in\mathcal{A},b\in\mathcal{B}}$ and separation gaps $(|\mu_{a,b} - \mu_{a,b'}|)_{a\in\mathcal{A},b\neq b'}$. Thus, distributions of bandits $\nu \in \mathcal{D}$ used in the next experiments are generated randomly so that some conditions are satisfied (see Eq. 6.14, 6.15). Recall that $\nu : (\nu_{b_1}, \dots, \nu_{b_{|\mathcal{B}|}})$ is the set of bandit configurations in the bandit set $\mathcal{B}$. We consider two different distributions $\nu^{(1)}$ and $\nu^{(2)}$, resulting in associated sets of bandits $\mathcal{B}_1$ and $\mathcal{B}_2$, satisfying the condition $C(\nu)$ in order to ensure the convergence of algorithms at each period:

$$C(\nu) : \forall b \in \mathcal{B}, \quad 8\sqrt{\frac{\log(HT)}{T}} \leqslant \min_{a\neq a_b^\star} \Delta_{a,b} \leqslant 12\sqrt{\frac{\log(HT)}{T}}. \tag{6.14}$$

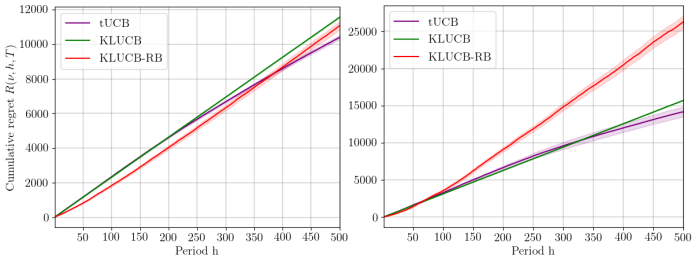
Let $\gamma(\alpha) \coloneqq \alpha\sqrt{\frac{\log(HT)}{T}}$. We generate two sets of bandits $\mathcal{B}_1$ and $\mathcal{B}_2$ such as to ensure that $\nu^{(1)}$ and $\nu^{(2)}$ satisfy

$$\gamma(12) \leqslant \gamma^\star_{\nu^{(1)}} \leqslant \gamma(16) \qquad \gamma(4) \leqslant \gamma^\star_{\nu^{(2)}} \leqslant \gamma(8). \tag{6.15}$$

Fig. C.3 (Appendix C.3.3) shows the bandit instances in the two generated bandit sets.
All experiments are conducted under the fair frequency $\beta = 1/|\mathcal{B}|$. More precisely, once a period $h \geqslant 1$ ends, $b_\star^{h+1}$ is sampled uniformly in $\mathcal{B}$ and independently of the past sequence $(b_\star^k)_{1\leqslant k\leqslant h}$. Fig. 6.3 shows the average cumulative regret with one standard deviation after $H = 100$ periods of $T = 5000$ rounds for the two settings. We observe that the performance of KLUCB−RB is tied to the smallest sub-optimal gap for all bandit instances. Fig. 6.3a highlights that KLUCB−RB outperforms KLUCB if the minimal sub-optimal gap of each bandit is less than the characteristic smaller separation gap $\gamma^\star_\nu$. This supports the observation from Sec. 6.4.1 that separation on optimal arms is sufficient. When arms are easier to separate than bandits, one might as well restart a classical KLUCB from scratch on each period (Fig. 6.3b). Note that situations where $0 < \gamma_\nu \ll \min_{b\in\mathcal{B}} \min_{a\neq a_b^\star} \Delta_{a,b}$ may not result in a catastrophic loss in learning performances if the arms in $\mathcal{A}^\star$ are *close enough* not to distort estimates computed on aggregated samples of from false positive models (see Appendix C.3).

6.4.3 Critical Settings

We saw previously that settings where bandit instances are difficult to distinguish may yield poor performance (see Section 6.4.2, Fig. 6.3b). Indeed, to determine if two estimated bandit models might result from the same bandit, both KLUCB−RB and tUCB rely on a compatibility over each arm, i.e. the intersection of confidence intervals. Therefore, it is generally harder to distinguish rollouts from many different distributions



(a) $T = 100$ (b) $T = 500$

Figure 6.4: Cumulative regret of KLUCB, KLUCB−RB and tUCB along $H = 500$ period for different numbers of rounds.

(that is the cardinal of $|\mathcal{B}|$ is high) when $|\mathcal{A}|$ is low and differences between arms are tight. To illustrate that, we consider an experiment on the setting described in Figure C.4 (Appendix C.3.3), composed of 4-armed bandits. We recall that tUCB requires in particular $|\mathcal{A}| \geqslant |\mathcal{B}|$. Thus we choose a set $|\mathcal{B}|$ of cardinal 4 in order to include a comparison of our algorithm with tUCB.
Here we have $|\mathcal{A}^\star| = \{0, 1, 3\}$ and $\gamma_\nu^\star := \min_{b \neq b'} \min_{a \in \mathcal{A}^\star} |\mu_{a,b} - \mu_{a,b'}| = 0.15$, while the minimal regret gaps of each instances are $(\min_{a \neq a_b^\star} \Delta_{a,b})_{b \in \mathcal{B}} = (0.74, 0.80, 0.81, 0.89)$. Consequently, finding the optimal arm at each period independently is here far less difficult than separating the different instances. Such a setting is clearly unfavorable for KLUCB−RB and we expect KLUCB to perform better.
Fig. 6.4a and Fig. 6.4b the cumulative regret for the three strategies, along $H = 500$ periods of $T = 100$ and $T = 500$ rounds respectively. As expected, KLUCB outperforms KLUCB−RB under this critical setting. On the other hand tUCB seems more robust and displays a cumulative regret trend that would be improving compared with KLUCB in the long run. One should still recall that tUCB requires knowing the cardinality of $|\mathcal{B}|$, while KLUCB−RB does not.
We may notice (Fig. 6.4a) that if the number of rounds T is sufficiently small, that is KLUCB does not have enough time to converge for each bandit, then KLUCB−RB does not perform significantly worse than KLUCB for the first periods. Then, as T rises (Fig. 6.4b), KLUCB begins to converge while KLUCB−RB still aggregate samples from confusing instances, which yields an explosion of the cumulative regret curve. We then expect for such setting that KLUCB will need far more longer periods ($T \to \infty$) to reach a regime in which it will discard all false positive rollouts and takes advantage over KLUCB. On the contrary, tUCB takes advantage of the knowledge of $|\mathcal{B}|$ and then waits to have enough confidence over the mean vectors of the 4 models to exploit them.

6.5 Conclusion

In this chapter we introduced the new routine bandits framework, for which we provided lower bounds on the regret (Proposition 5). This setting applies well to problems where, for example, customers anonymously return to interact with a system. These dynamics are known to be of interest to the community, as evidenced by the existing literature Gheshlaghi Azar et al. (2013); Gentile et al. (2014); Maillard and Mannor (2014). Routine bandits complement well these existing settings.
We then proposed the KLUCB−RB strategy (Alg. 10) to tackle the routine bandit setting by building on the seminal KLUCB algorithm for classical bandits. We proved upper bounds on the number of sub-optimal plays by KLUCB−RB (Theorem 5), which were used to prove asymptotic upper bounds on the regret (Corollary 4). This result shows the asymptotic optimality of the strategy and thanks to the proof technique that we considered, which is of independent interest, we further obtained finite-time regret guarantees with explicit quantities. We

indeed believe the proof technique may be useful to handle other structured setups beyond routine bandits.
We finally provided extensive numerical experiments to highlight the situations where KLUCB−RB can efficiently leverage information from previously encountered bandit instances to improve over a classical KLUCB. More importantly, we highlighted the cost to pay for re-using observations from previous periods, and showed that easy tasks may be better tackled independently. This is akin to an agent that would behave badly by relying on a wrong inductive bias. Fortunately, there are many situations where one can leverage knowledge from bandit instances faced in the past. This would notably be the case if the agent has to select products to recommend from a large set ($\mathcal{A}$) and it turns out that there exists a much smaller set of products ($\mathcal{A}^\star$) that is preferred by users (Sec. 6.4.1).
Our results notably show that transferring information from previously encountered bandits can be highly beneficial (e.g., see Fig. 6.1 and 6.3a). However, the lack of prior knowledge about previous instances (including the cardinality of the set of instances) introduces many challenges in transfer learning. For example, attempting to leverage knowledge from previous instances could result in negative transfer if bandits cannot be distinguished properly (e.g., see Fig. 6.4).
Therefore, reducing the cost incurred for separating bandit instances should constitute a relevant angle to tackle as future work. Another natural line of other future work could investigate extensions of KLUCB−RB to the recurring occurrence of other bandit instances, e.g., linear bandits, contextual bandits, and others.

Chapter 7

Conclusion

7.1 Summary of the presented contributions

First, we have revisited the setup of unimodal multi-armed bandits: We introduced a novel variant based on IMED algorithm. This algorithm does not separate exploration from exploitation rounds and is proven optimal for one-dimensional exponential family distributions. Remarkably, IMED-UB algorithm does not require any optimization procedure, which can be interesting for practitioners. We also provided a novel proof algorithm, in which we make explicit empirical lower and upper bounds, before tackling the handling of bad events by specific concentration tools, in particular Theorem 6 from Maillard (2018). This proof technique greatly simplifies and shortens the analysis of IMED-UB. Last, we provided numerical experiments that show the practical effectiveness of IMED-UB. Then, we extend our approach for unimodal bandits to what we called multimodal structure: We introduced IMED-MB algorithm that effectively explores local maximums by involving second order indexes (Equation 3.17). Interestingly, this second order exploration is made easy due to the consideration of IMED type indexes. Finally, we introduced the novel graph-structured bandit framework, for which we provided instance-dependent lower bounds on the regret (Proposition 3). This setting encompasses Unimodal and Lipschitz structures, which are known to be of interest to the community as evidenced by the existing literature (Saber et al. (2021a); Trinh et al. (2020); Combes and Proutiere (2014a); Magureanu et al. (2014)), and enables to have a unified treatment for such structures. We then proposed IMED-GS algorithm (Algorithm 9) to tackle the graph-structured bandit setting by building on popular IMED algorithm for unstructured bandits. We proved asymptotically optimal and fully explicit finite time guarantees on the regret, which are not very common in this framework. The analysis of the optimality of the proposed approach led us to state and prove a novel concentration inequality (Theorem 4) that can be useful for subsequent work. Finally, we show that novel IMED-GS algorithm has good performance with smaller samples (Figure 4.1).

7.2 Some personal satisfactions

The first problem I wanted to solve during my PhD was to propose an optimal and efficient algorithm for Aggregate of Bandits (Section A.1). This problem has been introduced to me by Odalric-Ambrym Maillard, my supervisor. I succeeded in solving this problem by considering the graph-structured structure, a more general setting, and by introducing IMED-GS algorithm. What I appreciated the most in solving this problem was to solving it by establishing strong links between IMED approaches and Theorem 6 from Maillard (2018), a theoretical paper of my supervisor on boundary crossing probabilities. Above all, I literally loved working with IMED algorithm from Honda and Takemura (2015). I appreciated promoting it at SCOOL, our research team at Inria Lille, and teaching it as a teacher assistant at l'X or CentraleSupélec. These teaching opportunies granted me professional and personal enrichment, and have been possible thanks to Odalric-Ambrym Maillard,

to whom I am very grateful.

7.3 Future work

First of all, the content of Chapters 3, 4 and 5 about multimodal and graph-structured bandits could be exploited to follow up on future publications. Now, we briefly provide some possible guidelines for future work.

7.3.1 Algorithms with computational efficiency

When the number of arms $|\mathcal{A}|$ is large, even for non-structured bandit problems, proposing algorithms that are computationally efficient is challenging. We illustrated in Figure 2 from Saber et al. (2020) how `d-IMED-UB`, algorithm based on `IMED` type indexes, can be efficient despite an increasing number of arms when assuming unimodal bandit structure. Further, the `IMED` indexes and its dichotomic exploration are easy to compute. That is why computation times under `d-IMED-UB` algorithm are quite reasonable. Thus, `d-IMED-UB` efficiency comes from `IMED` types indexes that allow second order exploration. We think that this mechanism of exploring the underlying structure based on second order `IMED` type indexes can be extended to structures other than the unimodal one.

7.3.2 `IMED` for Markov Decision Processes with known transition probabilities

An optimal algorithm for a given Markov Decision Process (MDP) follows optimal random cycle composed in particular of an optimal state-action pairs. Furthermore, there is two levels of sub-optimality for a sub-optimal state-action (s, a):

- s is a sub-optimal state but a is an optimal action in state s **or** state s is optimal but action a is sub-optimal
- s is a sub-optimal state and a a sub-optimal action in state s.

Thus, a mechanism allowing second order exploration may be relevant to identify and deal with these two types of sub-optimality. We think our work that consists in proposing `IMED` approaches for bandit problems can be inspiring to provide algorithms for MDPs with efficient mechanism of exploring sub-optimal state-action pairs. However, unlike the case of bandits, the learner cannot choose an arbitrary state-action pair at each time step when dealing with a Markov Decision Process. To work around this difficulty, we may assume that the transition probabilities plus some bounds on the rewards are known. These additional assumptions would make it possible to better estimate how much does it cost (in terms of regret) to reach (before being able to explore) a specific state from the current state.

7.3.3 From routine bandit to non-stationary bandit problem

The routine bandit setting is a variant of the multi-armed bandit problem in which a learner faces every day one of $\mathcal{B}$ many bandit instances. More specifically, at each period $h \in [\![1, H]\!]$, the same bandit $b_\star^h$ is considered during $T > 1$ consecutive time steps, but the identity $b_\star^h$ is unknown to the learner. Such a situation typically occurs in recommender systems when a learner may repeatedly serve the same user whose identity is unknown due to privacy issues. The numbers of time steps of episodes are supposed to be the same (equal to T) only for convenience. Episodes with different numbers of time steps can be considered. For routine bandits, the end of each period is known. If we assume that the ends of periods are now unknown, then we recover a non-stationary bandit setting where we assume a finite (but unknown) possible changes of distribution. By combining bandit-identification tests with a `KLUCB` type algorithm, we introduced `KLUCB` for Routine Bandits

(KLUCB-RB) algorithm. A natural question to explore is: How to adapt KLUCB-RB when the end (and the beginning) of each period is now unknown ?

Bibliography

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, R., Teneketzis, D., and Anantharam, V. (1989). Asymptotically efficient adaptive allocation schemes for controlled iid processes: Finite parameter space. *IEEE Transactions on Automatic Control*, 34(3).

Anandkumar, A., Ge, R., Hsu, D., and Kakade, S. (2013). A tensor spectral approach to learning mixed membership community models. In *Conference on Learning Theory*, pages 867–881. PMLR.

Anandkumar, A., Ge, R., Hsu, D. J., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832.

Baransi, A., Maillard, O.-A., and Mannor, S. (2014). Sub-sampling for multi-armed bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 115–131. Springer.

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.

Burnetas, A. N. and Katehakis, M. N. (1997). Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222–255.

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.

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

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. (2014a). Unimodal bandits: Regret lower bounds and optimal algorithms. In *International Conference on Machine Learning*.

Combes, R. and Proutiere, A. (2014b). Unimodal bandits: Regret lower bounds and optimal algorithms. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 521–529.

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

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

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.

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

Gao, X., Qi, H., Wen, X., Zheng, W., Lu, Z., and Hu, Z. (2019). Energy detection adjustment for fair co-existence of wi-fi and laa: A unimodal bandit approach. In *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, pages 1086–1091.

Garivier, A. (2013). Informational confidence bounds for self-normalized averages and applications. In *Information Theory Workshop (ITW), 2013 IEEE*, pages 1–5. IEEE.

Garivier, A. and Cappé, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, pages 359–376.

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*.

Gentile, C., Li, S., and Zappella, G. (2014). Online clustering of bandits. In *International Conference on Machine Learning*, pages 757–765.

Gheshlaghi Azar, M., Lazaric, A., and Brunskill, E. (2013). Sequential transfer in multi-armed bandit with finite set of models. In *Proc. NIPS*, pages 2220–2228. Curran Associates, Inc.

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 multi-armed 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.

Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 199–213.

Kleinberg, R. D., Niculescu-mizil, A., and Sharma, Y. (2008). Regret bounds for sleeping experts and bandits. In Servedio, R. A. and Zhang, T., editors, *Proceedings of the 21st annual Conference On Learning Theory*, volume 80 of *COLT '08*, pages 425–436, Helsinki, Finland. Omnipress.

Korda, N., Kaufmann, E., and Munos, R. (2013). Thompson Sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1448–1456.

Kunne, S., Maggi, L., Cohen, J., and Xu, X. (2020). Anytime backtrack unimodal bandits and applications to cloud computing. In *2020 IFIP Networking Conference (Networking)*, pages 82–90.

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

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.

Langford, J. and Zhang, T. (2007). The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information. In Platt, J. C., Koller, D., Singer, Y., Roweis, S. T., Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *NIPS*. MIT Press.

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.

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

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.

Lu, T., Pál, D., and Pál, M. (2010). Contextual multi-armed bandits. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the 13th international conference on Artificial Intelligence and Statistics*, volume 9, pages 485–492.

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

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.

Maillard, O.-A. and Mannor, S. (2014). Latent bandits. In *International Conference on Machine Learning (ICML)*.

Peña, V. H., Lai, T. L., and Shao, Q.-M. (2008). *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media.

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

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

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

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

Saber, H., Saci, L., Maillard, O.-A., and Durand, A. (2021b). Routine bandits: Minimizing regret on recurring problems. *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PPKD)*.

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.

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

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*.

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

Appendix A

Graph-Structured Bandits: Complements

We reintroduce some relevant notations in the following paragraph.

Notations. Let $\nu \in \mathcal{D}_\Theta$. Let $\mu^\star = \max_{a\in\mathcal{A}} \mu_a$ be the optimal mean and $\mathcal{A}^\star(\nu) = \operatorname{argmax}_{a\in\mathcal{A}} \mu_a$ be the set of optimal arms of ν. We define for an arm $a \in \mathcal{A}$ its sub-optimality gap $\Delta_a = \mu^\star - \mu_a$. Considering an horizon $T \geqslant 1$, thanks to the tower rule we can rewrite the regret as follows:

$$R(\nu, T) = \sum_{a\in\mathcal{A}} \Delta_a \, \mathbb{E}_\nu\big[N_a(T)\big] \,, \tag{A.1}$$

where $N_a(t) = \sum_{s=1}^t \mathbb{I}_{\{a_s=a\}}$ is the number of pulls of arm a at time t. For arms $a, a' \in \mathcal{A}$ and a relationship matrix $\theta \in \Theta$ we define their relative gap $\delta_{a,a'}(\theta) = \theta_{a,a'} - (\mu_a - \mu_{a'})$ and their algebraic gap $d_{a,a'}(\theta) = \Delta_a - \delta_{a,a'}(\theta)$. We note that under Assumption 4 for all $\theta \in \Theta$, $\delta_{a,a}(\theta) = 0$ and $d_{a,a}(\theta) = \Delta_a \geqslant 0$. Finally, we define $\Theta^\star(\nu) := \{\theta \in \Theta : \nu \in \mathcal{D}_\theta\}$ and $\overline{\Theta}_a(\nu) = \{\theta \in \Theta^\star(\nu) : \forall a' \in \mathcal{A}, \theta_{a,a'} \geqslant 0\}$ for $a \in \mathcal{A}$. When there is no possible confusion, $\mathcal{A}^\star(\nu)$, $\Theta^\star(\nu)$ and $\overline{\Theta}_a(\nu)$ are simply denoted $\mathcal{A}^\star$, $\Theta^\star$ and $\overline{\Theta}_a$. We note that for all $\theta \in \Theta^\star$, for all $a, a' \in \mathcal{A}$, $\delta_{a,a'}(\theta) \geqslant 0$. For a subset of arms $\mathcal{A}' \subset \mathcal{A}$, we denote $N_{\mathcal{A}'}(t) = \sum_{a'\in\mathcal{A}'} N_{a'}(t)$ the aggregated pulls from arms in $\mathcal{A}' \subset \mathcal{A}$. We remind that $\Delta_{\min} = \min_{a\notin\mathcal{A}^\star} \Delta_a$ and, with regard to Assumption 6 and 8, we introduce for convenience the quantity $\varepsilon_\nu = \dfrac{\delta_{\min}}{4} \wedge (1-\mu^\star)$, where $\delta_{\min} = \Delta_{\min}$ if Θ is a singleton, and $\delta_{\min} = \min_{a\neq a'}\{\Delta_{\min},\ \theta^\star_{a,a'} - (\mu_a - \mu_{a'})\}$ otherwise.

A.1 Structures Unimodal, Lipschitz and Aggregate of Bandits

We assume for convenience that $\mathcal{A} = \{1, \dots, |\mathcal{A}|\} \subset \mathbb{N}$. Interestingly, the graph-structure encompasses the classical structures Lipschitz, Unimodal and Aggregate of bandits. We detail each case separately in the following.

Lipschitz bandits. The graph-structure can be specified to handle Lipschitz bandits[1] by assuming that $\Theta = \{\theta\}$ is a singleton and there exists a positive constant $k > 0$ such that

$$\forall a, a' \in \mathcal{A}, \quad \theta_{a,a'} = k\,|a - a'| \,. \tag{A.2}$$

Indeed in such a case, the graph structure specializes to the following condition

$$\forall \nu \in \mathcal{D}_\Theta,\ \forall a, a' \in \mathcal{A}, \quad |\mu_a - \mu_{a'}| \leqslant k\,|a - a'| \,.$$

[1] The notion of Lipschitz bandit we refer to is the one used in Magureanu et al. (2014). We could have considered more general Lipschitz bandits.

Unimodal bandits. Likewise, a unimodal structure can be recovered by assuming

$$\Theta = \left\{ \theta \in \{0,1\}^{\mathcal{A}^2} : \exists a^\star \in \mathcal{A},\ \forall a \neq a',\quad \theta_{a,a'} = \left| \begin{array}{ll} 0 & \text{if } a' \in [\![a, a^\star]\!] \text{ or } a' \in [\![a^\star, a]\!] \\ 1 & \text{otherwise} \end{array} \right. \right\}. \tag{A.3}$$

Indeed in such a case, the graph structure specializes into a classical unimodal structure

$$\forall \nu \in \mathcal{D}_\Theta,\ \exists a^\star \in \mathcal{A},\ \forall a < a' < a^\star \text{ or } a > a' > a^\star,\quad \mu_a \leqslant \mu_{a'} \leqslant \mu_{a^\star}\,.$$

Aggregate of Bandits. We now introduce yet another specialization of the graph structure, which shows the graph structure goes beyond the previous classical examples. Here, we assume $\mathcal{A} = \mathcal{X} \times \mathcal{K}$ and view $\nu = (\nu_a)_{a\in\mathcal{A}}$ as a set of bandits $\nu = (\nu_x)_{x\in\mathcal{X}}$, where for all $x \in \mathcal{X}$, $\nu_x = (\nu_{x,k})_{k\in\mathcal{K}}$ is a $|\mathcal{K}|$-multi-armed bandit. Each bandit ν_x, for $x \in \mathcal{X}$ can be seen as a customer segment and the $|\mathcal{K}|$ arms as $|\mathcal{K}|$ new customer offers. The goal of the learner is then to exploit the best twinnings between customer segments $\mathcal{X}$ and new customer offers $\mathcal{K}$ while minimizing regret. We assume some similarities and dissimilarities between the customer segments in the appreciation of customer offers are provided. These similarities and dissimilarities are encoded in a matrix $\omega = (\omega_{x,x'})_{x,x'\in\mathcal{X}} \subset [-1\,;1]^{\mathcal{X}^2}$, assumed to be known to the learner, such that for all customer segments $x, x' \in \mathcal{X}$, for all customer offer $k \in \mathcal{K}$,

$$\mu_{x,k} - \mu_{x',k} \leqslant \omega_{x,x'}\,.$$

The matrix ω measures the separation between the customer segments and satisfies a pseudo-metric property. That is, for all customer segments $x, x', x'' \in \mathcal{X}$, $\omega_{x,x} = 0$ and $\omega_{x,x''} \leqslant \omega_{x,x'} + \omega_{x',x''}$. Now, one can express this structure as a graph structure assuming that

$$\Theta = \{\theta\}\,, \qquad \theta_{a,a'} = \omega_{x,x'}\,, \quad \forall a = (x,k), a' = (x',k') \in \mathcal{A} = \mathcal{X}\times\mathcal{K}\,. \tag{A.4}$$

Indeed, this definition yields

$$\forall \nu \in \mathcal{D}_\Theta,\ \forall (x,k),(x',k') \in \mathcal{X}\times\mathcal{K},\quad \mu_{x,k} - \mu_{x',k'} \leqslant \omega_{x,x'}\,.$$

Let us note the Aggregate of bandits is reminiscent of contextual bandits, where similarities between means are dictated by similarities between contexts.

A.2 Proof related to the regret lower bound

In this section, we provide a lower bound on the cumulative regret when assuming a graph structure. To this end, we follow the classical approach from Lai and Robbins (1985) that we apply to the case of a graph structure. In order to obtain non trivial lower bound we consider algorithms that are consistent (Definition 1).

For each sub-optimal arm $a \notin \mathcal{A}^\star(\nu)$, we introduce the following distribution-dependent subset of relationship matrices

$$\Theta_a(\nu) := \{\theta \in \Theta^\star(\nu) :\ \forall a^\star \in \mathcal{A}^\star(\nu),\ \theta_{a,a^\star} > 0 \quad \text{and} \quad \forall a' \notin \mathcal{A}^\star(\nu),\ \theta_{a,a'} \geqslant 0\}\,, \tag{A.5}$$

where $\Theta^\star(\nu) = \{\theta \in \Theta \colon \nu \in \mathcal{D}_\theta\}$. Then, we derive from the notion of consistency asymptotic regret lower bounds by considering *most confusing configurations* for each relationship matrix in $\Theta_a(\nu)$. These lower bounds involve the set of *informative* sub-optimal arms

$$\mathcal{A}_a(\theta) := \{a' \in \mathcal{A} :\ d_{a,a'}(\theta) \geqslant 0\} = \{a' \in \mathcal{A} :\ \mu_{a'} \leqslant \mu^\star - \theta_{a,a'}\}\,, \quad \theta \in \Theta_a(\nu)\,. \tag{A.6}$$

When "moving" sub-optimal arm a to make it optimal in a most confusing configuration, $\mathcal{A}_a(\theta)$ represents the set of sub-optimal arms which must also be "moved" in order to ensure the "most confusing"[2] bandit for sub-optimal a belongs to the structure $\mathcal{D}_\theta$. By definition of $\mathcal{A}_a(\theta)$, for all $a' \in \mathcal{A}$, it holds that $d_{a,a'}(\theta) < 0$ if $a' \notin \mathcal{A}_a(\theta)$. Since $\theta_{a,a} = 0$ and $\theta_{a,a^\star} > 0$ for all $a^\star \in \mathcal{A}^\star$, we further get that $a \in \mathcal{A}_a(\theta)$ and $\mathcal{A}_a(\theta) \cap \mathcal{A}^\star = \emptyset$.

Proposition 7 (Lower bounds on pulls)**.** *Let us consider a consistent bandit algorithm. Then, for all configuration $\nu \in \mathcal{D}_\Theta$, for all sub-optimal arm $a \notin \mathcal{A}^\star(\nu)$, for all relationship matrix $\theta \in \Theta_a(\nu)$, under Assumption 4 it must be that*

$$\forall 0 < \varepsilon < \varepsilon_a, \quad \liminf_{T\to\infty} \frac{1}{\log(T)} \sum_{a' \in \mathcal{A}_a(\theta)} N_{a'}(T) \mathrm{kl}(\mu_{a'} | \mu^\star - \theta_{a,a'} + \varepsilon) \geqslant 1 \,,$$

where $\varepsilon_a := \min\left\{ \left(-d_{a,a'}(\theta)\right)_{a' \notin \mathcal{A}_a(\theta)} , (1-\mu^\star)/2 \right\} > 0.$

The proof of this result uses a change-of-measure argument and follows classical proof techniques from the literature, see Lai and Robbins (1985); Agrawal et al. (1989); Graves and Lai (1997); Cappé et al. (2013). We detail it in the following sub-section for completeness.

A.2.1 Proof of Proposition 7

Let us consider a sub-optimal arm $a \notin \mathcal{A}^\star$ and $\theta \in \Theta_a(\nu)$. From the definitions of Θ_a (Eq. 4.2) and ε_a we respectively have

$$\forall a' \in \mathcal{A}, \quad \theta_{a,a'} \geqslant 0 \,,$$

and

$$\forall a' \in \mathcal{A}_a(\theta),\ \mu^\star - \theta_{a,a'} + \varepsilon_a < 1 \quad \text{and} \quad \forall a' \notin \mathcal{A}_a(\theta),\ \mu_{a'} \geqslant \mu^\star - \theta_{a,a'} + \varepsilon_a \,. \tag{A.7}$$

Let us then consider $0 < \varepsilon < \varepsilon_a$ and the maximal confusing distributions $\tilde{\nu} = (\tilde{\nu}_{a'})_{a' \in \mathcal{A}}$ for the sub-optimal arm a with means $(\tilde{\mu}_{a'})_{a' \in \mathcal{A}}$ such that

$$\begin{aligned} &\forall a' \in \mathcal{A}_a(\theta), \quad \tilde{\mu}_{a'} = \mu^\star - \theta_{a,a'} + \varepsilon \\ &\forall a' \notin \mathcal{A}_a(\theta), \quad \tilde{\mu}_{a'} = \mu_{a'} \,. \end{aligned} \tag{A.8}$$

Then, since $\theta_{a,a'} \geqslant 0$ for all $a' \in \mathcal{A}_a(\theta)$, we have for all $a' \in \mathcal{A}$,

$$\tilde{\mu}_{a'} \leqslant \mu^\star + \varepsilon \,,$$

with equality in previous inequality if and only if $a' \in \mathcal{A}_a(\theta)$ and $\theta_{a,a'} = 0$. The set of optimal arms for $\tilde{\nu}$ is then

$$\mathcal{A}^\star(\tilde{\nu}) = \{ a' \in \mathcal{A}_a(\theta) \colon\ \theta_{a,a'} = 0 \} \ni a \,. \tag{A.9}$$

Furthermore, from the definition of $\mathcal{A}_a(\theta)$ (Eq. 4.3), we get

$$\forall a' \in \mathcal{A}_a(\theta), \quad \tilde{\mu}_{a'} > \mu_{a'} \,. \tag{A.10}$$

Then, under Assumptions 4, Eq. A.7, A.8 and A.10 imply

$$\begin{aligned} &\forall a', a'' \notin \mathcal{A}_a(\theta),\ a' \neq a'', \quad \tilde{\mu}_{a'} - \tilde{\mu}_{a''} = \mu_{a'} - \mu_{a''} \leqslant \theta_{a',a''} \\ &\forall a' \notin \mathcal{A}_a(\theta), \forall a'' \in \mathcal{A}_a(\theta), \quad \tilde{\mu}_{a'} - \tilde{\mu}_{a''} < \mu_{a'} - \mu_{a''} \leqslant \theta_{a',a''} \\ &\forall a', a'' \in \mathcal{A}_a(\theta), a' \neq a'', \quad \tilde{\mu}_{a'} - \tilde{\mu}_{a''} = \theta_{a,a''} - \theta_{a,a'} \leqslant \theta_{a',a''} \\ &\forall a' \in \mathcal{A}_a(\theta), \forall a'' \notin \mathcal{A}_a(\theta), \quad \tilde{\mu}_{a'} - \tilde{\mu}_{a''} \leqslant \mu^\star - \theta_{a,a'} + \varepsilon - (\mu^\star - \theta_{a,a''} + \varepsilon) \leqslant \theta_{a',a''} \,. \end{aligned}$$

[2] These notions of "moving" and "most confusing" refer to the generic proof technique used to derive regret lower bounds. It involves a change-of-measure argument, from the initial configuration in which the arm is sub-optimal to another one chosen to make it optimal.

This implies

$$\tilde{\nu} \in \mathcal{D}_\theta \subset \mathcal{D}_\Theta \,. \tag{A.11}$$

Let $0<c<1$. We will show that almost surely

$$\liminf_{T\to\infty} \frac{1}{\log(T)} \sum_{a'\in\mathcal{A}_a(\theta)} N_{a'}(T)\,\mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'}) \geqslant c\,.$$

We start with the following inequality

$$\begin{aligned}
&\mathbb{P}_\nu\left(\liminf_{T\to\infty} \frac{1}{\log(T)} \sum_{a'\in\mathcal{A}_a(\theta)} N_{a'}(T)\,\mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'}) < c\right) \\
\leqslant\ &\liminf_{T\to\infty} \mathbb{P}_\nu\left(\frac{1}{\log(T)} \leqslant \sum_{a'\in\mathcal{A}_a(\theta)} N_{a'}(T)\mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'}) < c\right).
\end{aligned}$$

Let us consider an horizon $T\geqslant 1$ and let us introduce the event

$$\Omega_T = \left\{\sum_{a'\in\mathcal{A}_a(\theta)} N_{a'}(T)\mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'}) < c\log(T)\right\}. \tag{A.12}$$

We want to provide an upper bound on $\mathbb{P}_\nu(\Omega_T)$ to ensure $\lim_{T\to\infty}\mathbb{P}_\nu(\Omega_T)=0$. We start by taking advantage of the following lemma.

Lemma 22 (Change of measure)**.** *For every measurable event Ω with respect to ν and $\tilde{\nu}$,*

$$\forall x\in\mathbb{R},\quad \mathbb{P}_\nu(\Omega\cap\mathcal{C}_x) \leqslant \exp(x)\mathbb{P}_{\tilde{\nu}}(\Omega)\,,$$

where $\mathcal{C}_x = \left\{\log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\tilde{\nu}}(\psi)\right)\leqslant x\right\}$ and $\psi=((a_t),X_t)_{t=1..T}$ is the sequence of pulled arms and rewards.

Let $\alpha\in(0,1)$ and let us introduce the event

$$\mathcal{C}_{\alpha,T} = \left\{\log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\tilde{\nu}}(\psi)\right) \leqslant (1-\alpha)\log(T)\right\}. \tag{A.13}$$

Then we can decompose the probability $\mathbb{P}_\nu(\Omega_T)$ as follows

$$\mathbb{P}_\nu(\Omega_T) = \mathbb{P}_\nu(\Omega_T\cap\mathcal{C}_{\alpha,T}) + \mathbb{P}_\nu(\Omega_T\cap\mathcal{C}^c_{\alpha,T}) \leqslant T^{1-\alpha}\mathbb{P}_{\tilde{\nu}}(\Omega_T) + \mathbb{P}_\nu\left(\Omega_T\cap\mathcal{C}^c_{\alpha,T}\right). \tag{A.14}$$

Now, we control successively the terms $T^{1-\alpha}\mathbb{P}_{\tilde{\nu}}(\Omega_T)$ and $\mathbb{P}_\nu\left(\Omega_T\cap\mathcal{C}^c_{\alpha,T}\right)$ and show that they both tend to 0 as T tends to ∞.

We first provide an upper bound on $\mathbb{I}_{\{\Omega_T\}}$ by noting that from Eq. A.9 we have

$$\begin{aligned}
\Omega_T &\subset \left\{\sum_{a'\in\mathcal{A}_a(\theta)} N_{a'}(T) < \frac{c}{\min\limits_{a\in\mathcal{A}_a(\theta)}\mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'})}\log(T)\right\} \\
&= \left\{T < \frac{c}{\min\limits_{a\in\mathcal{A}_a(\theta)}\mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'})}\log(T) + \sum_{a'\notin\mathcal{A}_a(\theta)} N_{a'}(T)\right\} \\
&\subset \left\{T < \frac{c}{\min\limits_{a\in\mathcal{A}_a(\theta)}\mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'})}\log(T) + \sum_{a'\notin\mathcal{A}^\star(\tilde{\nu})} N_{a'}(T)\right\}.
\end{aligned}$$

This implies

$$\mathbb{I}_{\{\Omega_T\}} \leqslant \frac{c}{\min\limits_{a\in\mathcal{A}_a(\theta)} \mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'})} \frac{\log(T)}{T} + \sum_{a'\notin\mathcal{A}^\star(\tilde{\nu})} \frac{N_{a'}(T)}{T}\,. \tag{A.15}$$

Since we assume a consistent algorithm on $\mathcal{D}_\Theta$ and $\tilde{\nu}\in\mathcal{D}_\Theta$ (Eq. A.11), we know that

$$\forall a' \notin \mathcal{A}^\star(\tilde{\nu}), \quad \mathbb{E}_{\tilde{\nu}}\left[\frac{N_{a'}(T)}{T^\alpha}\right] = o(1)\,, \tag{A.16}$$

therefore from Eq. A.15 and A.16 we get

$$T^{1-\alpha}\mathbb{P}_{\tilde{\nu}}(\Omega_T) = o(1)\,. \tag{A.17}$$

Now we control the remaining term $\mathbb{P}_\nu\big(\Omega_T \cap \mathcal{C}^c_{\alpha,T}\big)$.

For each time $t\in[\![1,T]\!]$, the reward X_t is sampled independently from the past and according to ν_{a_t}. Hence the likelihood ratio rewrites

$$\frac{\mathrm{d}\nu}{\mathrm{d}\tilde{\nu}}(\psi) = \prod_{t=1}^{T} \frac{\mathrm{d}\nu_{a_t}}{\mathrm{d}\tilde{\nu}_{a_t}}(X_t)\,, \quad \text{where } \frac{\mathrm{d}\nu_a}{\mathrm{d}\tilde{\nu}_a}(x) = \frac{\mu_a^x(1-\mu_a)^{1-x}}{\tilde{\mu}_a^x(1-\tilde{\mu}_a)^{1-x}},\ \forall a\in\mathcal{A}, \forall x\in\{0,1\}\,. \tag{A.18}$$

Thus, since for all $a'\notin\mathcal{A}_a(\theta)$, $\tilde{\mu}_a = \mu_a$, the log-likelihood ratio is

$$\log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\tilde{\nu}}(\psi)\right) = \sum_{a'\in\mathcal{A}_a(\theta)} \sum_{t=1}^{T} \mathbb{I}_{\{a_t=a'\}} \log\left(\frac{\mathrm{d}\nu_{a'}}{\mathrm{d}\tilde{\nu}_{a'}}(X_t)\right). \tag{A.19}$$

Hence from Eq. A.12, A.13 and A.19 we can rewrite the set

$$\Omega\cap\mathcal{C}^c_{\alpha,T} = \left\{ \begin{array}{l} \displaystyle\sum_{a'\in\mathcal{A}_a(\theta)} \sum_{t=1}^{T} \mathbb{I}_{\{a_t=a'\}}\left[\log\left(\frac{\mathrm{d}\nu_{a'}}{\mathrm{d}\tilde{\nu}_{a'}}(X_t)\right) - \mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'})\right] > (1-\alpha-c)\log(T) \\ \displaystyle\sum_{a'\in\mathcal{A}_a(\theta)} N_{a'}(T)\mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'}) < c\log(T) \end{array} \right\}.$$

Let us introduce $X^n_a = X_{\tau^n_a}$ where $\tau^n_a = \min\{t\geqslant 1 : N_a(t) = n\}$ for all $a\in\mathcal{A}$. Note that the random variables τ^n_a are predictable stopping times, since $\{\tau^n_a = t\}$ is measurable with respect to the filtration generated by $(a_1, X_1, ..., a_{t-1}, X_{t-1})$. For $a'\in\mathcal{A}_a(\theta)$ and $n\geqslant 1$, let us consider

$$Z^n_{a'} = \frac{\mathrm{d}\nu_{a'}}{\mathrm{d}\tilde{\nu}_{a'}}(X^n_{a'})\,.$$

Then $Z^n_{a'}$ is positive and bounded by $B_{a'} = 1/\tilde{\mu}_{a'}(1-\tilde{\mu}_{a'})$, with mean $\mathbb{E}_\nu[Z^n_{a'}] = \mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'})$. Furthermore, the random variables $Z^n_{a'}$, for $a'\in\mathcal{A}_a(\theta)$ and $n\geqslant 1$, are independent. Thus, it holds

$$\Omega_T\cap\mathcal{C}^c_{\alpha,T} \subset \left\{ \max_{m\in\mathcal{M}_T} \sum_{a'\in\mathcal{A}_a(\theta)} \sum_{n=1..m_{a'}} Z^n_{a'} - \mathbb{E}_\nu[Z^n_{a'}] > \left(\frac{1-\alpha}{c} - 1\right) c\log(T) \right\}, \tag{A.20}$$

where $\mathcal{M}_T := \Big\{(m_{a'})\in[\![1,T]\!]^{\mathcal{A}_a(\theta)} : \sum_{a'\in\mathcal{A}_a(\theta)} m_{a'}\mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'}) < c\log(T)\Big\}$.

In the following, we control the asymptotic concentration of random variables $(Z^n_{a'})$ by applying Doob's maximal inequality. For $a' \in \mathcal{A}_a(\theta)$ and $\lambda > 0$, let us introduce the super-martingale

$$(M_{a',m})_{m\geqslant 0} = \left(\exp\left(\lambda \sum_{n=1}^{m} \left(Z^n_{a'} - E[Z^n_{a'}] \right) - m\lambda^2 \frac{B^2_{a'}}{8} \right) \right)_{m\geqslant 0} . \tag{A.21}$$

Then noting that

$$\forall (m_{a'}) \in \mathcal{M}_T, \ \frac{\sum\limits_{a'\in\mathcal{A}_a(\theta)} \lambda^2 m_{a'} \frac{B^2_{a'}}{8}}{c\log(T)} < \frac{\sum\limits_{a'\in\mathcal{A}_a(\theta)} \lambda^2 m_{a'} \frac{B^2_{a'}}{8}}{\sum\limits_{a'\in\mathcal{A}_a(\theta)} m_{a'} \mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'})} \leqslant \lambda^2 \frac{\max\limits_{a'\in\mathcal{A}_a(\theta)} B^2_{a'}}{8 \min\limits_{a'\in\mathcal{A}_a(\theta)} \mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'})} , \tag{A.22}$$

we obtain from Eq. A.20 and A.21 that

$$\begin{aligned} \Omega_T \cap \mathcal{C}^c_{\alpha,T} &\subset \left\{ \max_{(m_{a'})\in\mathcal{M}_T} \prod_{a'\in\mathcal{A}_a(\theta)} M_{a',m_{a'}} > T^{\left[\lambda(\frac{1-\alpha}{c}-1) - \lambda^2 \frac{\max\limits_{a'\in\mathcal{A}_a(\theta)} B^2_{a'}}{8 \min\limits_{a'\in\mathcal{A}_a(\theta)} \mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'})}\right] c} \right\} \\ &\subset \left\{ \exists a' \in \mathcal{A}_a(\theta) : \max_{m\leqslant \overline{m}} M_{a',m} > T^{\gamma} \right\} , \end{aligned} \tag{A.23}$$

where $\overline{m} = \frac{c\log(T)}{\min\limits_{a'\in\mathcal{A}_a(\theta)} \mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'})}$ and $\gamma = \left[\lambda(\frac{1-\alpha}{c}-1) - \lambda^2 \frac{\max_{a'\in\mathcal{A}_a(\theta)} B^2_{a'}}{8 \min_{a'\in\mathcal{A}_a(\theta)} \mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'})}\right] \frac{c}{|\mathcal{A}_a(\theta)|}$.

In order to have $\gamma > 0$ in (A.23), we impose:
- $0 < \alpha < 1 - c$ (this implies $\frac{1-\alpha}{c} - 1 > 0$)
- $\lambda \in \underset{\lambda'\geqslant 0}{\mathrm{argmax}} \left\{ \lambda'(\frac{1-\alpha}{c} - 1) - \lambda'^2 \frac{\max_{a'\in\mathcal{A}_a(\theta)} B^2_{a'}}{8 \min_{a'\in\mathcal{A}_a(\theta)} \mathrm{kl}(\mu_{a'}|\tilde{\mu}_{a'})} \right\} > 0.$

Then from (A.23) we have

$$\begin{aligned} \mathbb{P}_\nu(\Omega_T \cap \mathcal{C}^c_{\alpha,T}) &\leqslant \sum_{a'\in\mathcal{A}_a(\theta)} \mathbb{P}_\nu\left(\max_{m\leqslant\overline{m}} M_{a',m} > T^\gamma \right) \quad \text{(Union bound)} \\ &\leqslant \sum_{a'\in\mathcal{A}_a(\theta)} \frac{\mathbb{E}_\nu[M_{a',0}]}{T^\gamma} \quad \text{(Doob's maximal inequality)} \\ &= \frac{|\mathcal{A}_a(\theta)|}{T^\gamma} . \end{aligned}$$

This implies the following control

$$\mathbb{P}_\nu(\Omega_T \cap E^c_T) = o(1) . \tag{A.24}$$

Finally, by combining (A.8), (A.12), (A.14), (A.17) and (A.24), we show

$$\forall 0 < \varepsilon < \varepsilon_a, \ \forall 0 < c < 1, \quad \liminf_{T\to\infty} \frac{1}{\log(T)} \sum_{a'\in\mathcal{A}_a(\theta)} N_{a'}(T) \, \mathrm{kl}(\mu_{a'}|\mu^\star - \theta_{a,a'} + \varepsilon) \geqslant c . \tag{A.25}$$

The proof ends by doing $c \to 1$.

A.2.2 Proof of Proposition 3

Let $(T_k)_{k\in\mathbb{N}}$ be a sub-sequence such that

$$\liminf_{T\to\infty} \frac{R(T,\nu)}{\log(T)} = \lim_{k\to\infty} \frac{R(T_k,\nu)}{\log(T_k)}\,.$$

We assume that this limit is finite otherwise the result is straightforward. This implies in particular

$$\forall a \notin \mathcal{A}, \quad \limsup_{k\to\infty} \frac{\mathbb{E}_\nu\left[N_a(T_k)\right]}{\log(T_k)} < +\infty\,.$$

By Cantor's diagonal argument there exists an extraction of $(T_k)_{k\in\mathbb{N}}$ denoted by $(T'_k)_{k\in\mathbb{N}}$ such that for all $a\notin\mathcal{A}^\star$, there exist $N_a\in\mathbb{R}_+$ such that

$$\lim_{k'\to\infty} \frac{\mathbb{E}_\nu\left[N_a(T'_k)\right]}{\log(T'_k)} = N_a\,.$$

Hence we get

$$\liminf_{T\to\infty} \frac{R(T,\nu)}{\log(T)} = \sum_{a\notin\mathcal{A}^\star} N_a\Delta_a\,.$$

But thanks to Proposition 7, under Assumptions 4, we have for all $a\notin\mathcal{A}^\star$ such that $\Theta_a\neq\emptyset$,

$$\begin{aligned}
&\inf_{\theta\in\Theta_a} \sum_{a'\in\mathcal{A}_a(\theta)} \mathrm{kl}(\mu_{a'}|\mu^\star-\theta_{a,a'})N_a \\
=\;& \inf_{\theta\in\Theta_a} \min_{0<\varepsilon<\varepsilon_a} \sum_{a'\in\mathcal{A}_a(\theta)} \mathrm{kl}(\mu_{a'}|\mu^\star-\theta_{a,a'}+\varepsilon)N_a \\
=\;& \inf_{\theta\in\Theta_a} \min_{0<\varepsilon<\varepsilon_a} \lim_{k\to\infty} \mathbb{E}_\nu\left[\sum_{a'\in\mathcal{A}_a(\theta)} \mathrm{kl}(\mu_{a'}|\mu^\star-\theta_{a,a'}+\varepsilon)\frac{N_{a'}(T'_k)}{\log(T'_k)}\right] \\
\geqslant\;& \inf_{\theta\in\Theta_a} \min_{0<\varepsilon<\varepsilon_a} \mathbb{E}_\nu\left[\liminf_{k\to\infty} \sum_{a'\in\mathcal{A}_a(\theta)} \mathrm{kl}(\mu_{a'}|\mu^\star-\theta_{a,a'}+\varepsilon)\frac{N_{a'}(T'_k)}{\log(T'_k)}\right] \\
\geqslant\;& \inf_{\theta\in\Theta_a} \min_{0<\varepsilon<\varepsilon_a} \mathbb{E}_\nu\left[\liminf_{T\to\infty} \sum_{a'\in\mathcal{A}_a(\theta)} \mathrm{kl}(\mu_{a'}|\mu^\star-\theta_{a,a'}+\varepsilon)\frac{N_{a'}(T)}{\log(T)}\right] \geqslant 1\,.
\end{aligned}$$

Therefore we obtain the lower bound

$$\begin{aligned}
\liminf_{T\to\infty} \frac{R(\nu,T)}{\log(T)} \geqslant \mathfrak{C}_\Theta(\nu) := \inf_{n\in\mathbb{R}_+^{\mathcal{A}}} \quad & \sum_{a\notin\mathcal{A}^\star} n_a\Delta_a \\
s.t. \quad & \forall a\notin\mathcal{A}^\star,\ \Theta_a\neq\emptyset,\ \inf_{\theta\in\Theta_a} \sum_{a'\in\mathcal{A}_a(\theta)} \mathrm{kl}(\mu_{a'}|\mu^\star-\theta_{a,a'})n_{a'} \geqslant 1\,.
\end{aligned}$$

To end the proof we show that for all $n\in\mathbb{R}_+^{\mathcal{A}}$ for all sub-optimal arm $a\notin\mathcal{A}^\star$ such that $\Theta_a\neq\emptyset$,

$$\inf_{\theta\in\Theta_a} \sum_{a'\in\mathcal{A}_a(\theta)} \mathrm{kl}(\mu_{a'}|\mu^\star-\theta_{a,a'})n_{a'} = \min_{\theta\in\overline{\Theta}_a} \sum_{a'\in\mathcal{A}_a(\theta)} \mathrm{kl}(\mu_{a'}|\mu^\star-\theta_{a,a'})n_{a'}\,.$$

First, we rewrite

$$\inf_{\theta\in\Theta_a}\sum_{a'\in\mathcal{A}_a(\theta)}\mathrm{kl}(\mu_{a'}|\mu^\star-\theta_{a,a'})n_{a'}=\inf_{\theta\in\Theta_a}\sum_{a'\notin\mathcal{A}^\star}n_{a'}\,\mathrm{kl}(\mu_{a'}|\mu^\star-\theta_{a,a'})\mathbb{I}_{\left\{\mu_{a'}\leqslant\mu^\star-\theta_{a,a'}\right\}}\,.$$

Then, we simply note that the function

$$:\theta\in\overline{\Theta}_a\mapsto\sum_{a'\notin\mathcal{A}^\star}n_{a'}\,\mathrm{kl}(\mu_{a'}|\mu^\star-\theta_{a,a'})\mathbb{I}_{\left\{\mu_{a'}\leqslant\mu^\star-\theta_{a,a'}\right\}}$$

is continuous.

A.3 Technical results

In this section, some practical results, in particular about $\mathrm{kl}(\cdot|\cdot)$ and the continuity of minimization problem, are detailed.

Lemma 23 (kl inequality 1)**.** *For* $0<p<q<1$,

$$\mathrm{kl}(p|q)\leqslant\frac{(q-p)^2}{q(1-q)}\,.$$

Proof. Let us consider

$$f:x\in[p\,;q]\mapsto\mathrm{kl}(p|x)-\frac{(x-p)^2}{x(1-x)}\,.$$

Then f admits a derivative in $]p\,;q[$ and for $x\in]p\,;q[$,

$$\begin{aligned}
f'(x) &= \frac{x-p}{x(1-x)}-\frac{2(x-p)}{x(1-x)}+\frac{(x-p)^2(1-x-x)}{x^2(1-x)^2}\\
&= -\frac{x-p}{x(1-x)}+\frac{(x-p)^2(1-2x)}{x^2(1-x)^2}\\
&= \frac{x-p}{x^2(1-x)^2}\left[-x(1-x)+(x-p)(1-2x)\right]\\
&= \frac{x-p}{x^2(1-x)^2}\left[-x^2+2px-p\right]\\
&= \frac{x-p}{x^2(1-x)^2}\left[-(x-p)^2+p^2-p\right]\\
&< 0\,.
\end{aligned}$$

Thus, f is a decreasing function. Since $f(p)=0$, we have $f(q)\leqslant 0$, which ends the proof. □

Lemma 24 (kl inequality 2)**.** *For* $0<p<q<q'<1$,

$$\mathrm{kl}(p|q')-\mathrm{kl}(p|q)\geqslant 2(q-q')^2\,.$$

Proof. We note that

$$:x\in(0,q]\mapsto\mathrm{kl}(x|q')-\mathrm{kl}(x|q)=x\log\left(\frac{q}{q'}\right)+(1-x)\log\left(\frac{1-q}{1-q'}\right)$$

is an affine function with slope $\log\left(\dfrac{q(1-q')}{q'(1-q)}\right)<0$ since $q<q'$. This implies

$$\mathrm{kl}(p|q')-\mathrm{kl}(p|q)\geqslant\mathrm{kl}(q|q')-\mathrm{kl}(q|q)=\mathrm{kl}(q|q')\,.$$

Then Pinsker's inequality implies

$$\mathrm{kl}(q|q') \geqslant 2(q-q')^2\,,$$

which ends the proof. □

Lemma 25 (kl inequality 3)**.** *For* $z \geqslant 0$*, for* $0<p'<p<q-z<q'-z<1$*,*

$$\frac{\mathrm{kl}(p|q-z)}{q-p} \leqslant \frac{\mathrm{kl}(p'|q'-z)}{q'-p'}\,.$$

Proof. We show that functions

$$f:\ x\in[p'\,,p] \mapsto \frac{\mathrm{kl}(x|q-z)}{q-x} \tag{A.26}$$

$$g:\ y\in[q\,,q'] \mapsto \frac{\mathrm{kl}(p'|y-z)}{y-p'} \tag{A.27}$$

are respectively decreasing and increasing functions. For $x\in[p'\,,p]$,

$$\begin{aligned} f'(x) &= \frac{1}{q-x}\times\frac{\partial \mathrm{kl}}{\partial p}(x|q-z)+\frac{1}{(q-x)^2}\times\mathrm{kl}(x|q-z) \\ &= \frac{1}{(q-x)^2}\times\left[\mathrm{kl}(x|q-z)-\left(-(q-z)\frac{\partial \mathrm{kl}}{\partial p}(x|q-z)\right)\right]\,. \end{aligned} \tag{A.28}$$

Since $x<p<q-z$, from Equation (A.28) and Lemma 26 we get

$$f'(x) = -\frac{\mathrm{kl}(q-z|x)}{(q-x)^2} \leqslant 0\,. \tag{A.29}$$

For $y\in[q\,,q']$,

$$\begin{aligned} g'(y) &= \frac{1}{y-p'}\times\frac{\partial \mathrm{kl}}{\partial q}(p'|y-z)-\frac{1}{(y-p')^2}\times\mathrm{kl}(p'|y-z) \\ &= \frac{1}{y-p'}\times\frac{y-z-p'}{(y-z)(1-y+z)}-\frac{1}{(y-p')^2}\times\mathrm{kl}(p'|y-z) \\ &= \frac{(y-z)(y-z-p')-(y-z)(1-y+z)\mathrm{kl}(p'|y-z)}{(y-p')^2(y-z)(1-y+z)}\,. \end{aligned} \tag{A.30}$$

Since $p'<q-z<y-z$, from Equation (A.30) and Lemma 23 we get

$$\begin{aligned} g'(y) &\geqslant \frac{(y-z)(y-z-p')-(y-z-p')^2}{(y-p')^2(y-z)(1-y+z)} \\ &= \frac{p'(y-z-p')}{(y-p')^2(y-z)(1-y+z)} \\ &\geqslant 0\,. \end{aligned} \tag{A.31}$$

□

Lemma 26 (kl inequality 4)**.** *For* $0<p<q<1$*,*

$$-(q-p)\frac{\partial \mathrm{kl}}{\partial p}(p|q) = \mathrm{kl}(p|q)+\mathrm{kl}(q|p)\,.$$

Proof. We have

$$\mathrm{kl}(p|q) = p\log(p) - p\log(q) + (1-p)\log(1-p) - (1-p)\log(1-q) \tag{A.32}$$

$$\mathrm{kl}(q|p) = q\log(q) - q\log(p) + (1-q)\log(1-q) - (1-q)\log(1-p)\,, \tag{A.33}$$

which implies

$$\begin{aligned} \frac{\partial \mathrm{kl}}{\partial p}(p|q) &= \log(p) + p \times \frac{1}{p} - \log(q) - \log(1-p) + (1-p) \times \frac{(-1)}{1-p} + \log(1-q) \\ &= \log(p) + 1 - \log(q) - 1 - \log(1-p) + \log(1-q) \\ &= \log(p) - \log(q) - \log(1-p) + \log(1-q) \end{aligned} \tag{A.34}$$

and

$$\begin{aligned} \mathrm{kl}(p|q) + \mathrm{kl}(q|p) &= (p-q)\log(p) + (q-p)\log(q) + (q-p)\log(1-p) + (p-q)\log(1-q) \\ &= -(q-p)\left[\log(p) - \log(q) - \log(1-p) + \log(1-q)\right] \\ &= -(q-p)\frac{\partial \mathrm{kl}}{\partial p}(p|q)\,. \end{aligned} \tag{A.35}$$

□

Lemma 27 (Continuity of minimization problem)**.** $:\nu' \in \mathcal{D}_\Theta \mapsto c_\Theta(\nu') \in \mathbb{R}_+$ *is well defined and continuous.*

Please refer to Section B.2 from Combes et al. (2017) for a proof of Lemma 27.

Lemma 28 (Real analysis 1)**.** *For* $x \geqslant e$*,*

$$\forall y \geqslant x\,(\mathrm{f}_\xi(x))^2\,, \quad \frac{y}{\mathrm{f}_\xi(y)} \geqslant x\,.$$

Lemma 29 (Real analysis 2)**.** *For* $x > 1$*,*

$$\frac{1}{-\log(1-1/x)} \leqslant x\,.$$

A.4 Additional experiments

In this section we introduce `PO-IMED-GS` algorithm, an inspired but simplified version of `IMED-GS`, that may be appealing to the practitioner. Then, we provide additional experiments in which the regret is averaged over many random structured configurations for each of the three considered structures (Unimodal, Lipschitz and Aggregate of bandits).

A.4.1 `PO-IMED-GS` algorithm

We introduce in this subsection `PO-IMED-GS` algorithm, for Pareto-Optimal `IMED-GS` algorithm. `PO-IMED-GS` is inspired from `IMED-GS` and simply consists in pulling the arm with minimal graph-structured `IMED`-type index at each time step. Pulling this arm intuitively ensures the constrains in the optimization problem of the regret lower bound are asymptotically satisfied (Pareto optimality). In particular, `PO-IMED-GS` algorithm does not solve any optimization problem Please refer to (4.9) for the definition of these graph-structured indexes. Hence this algorithm is simpler to implement, interpret and also has lower computational complexity,

although does not a priori enjoy as refined theoretical guarantees as IMED-GS. PO-IMED-GS algorithm is summarized in Algorithm 11.

Algorithm 11 PO-IMED-GS

Input: Structure Θ, ξ.
Pull each arm once
for $t = |\mathcal{A}| \dots T-1$ **do**
    Pull arbitrarily $a_{t+1} \in \operatorname*{argmin}_{a \in \mathcal{A}} \overline{I}_a(t)$, see (4.9).
end for

A.4.2 Regrets Averaged on Random Structured Configurations

For each structure of Figure A.1, we considered $|\mathcal{A}| = 18$ arms, a time horizon of $T = 3000$, and results averaged over 100 randomly generated structured configurations with 10 independent runs for each obtained structured configuration (that is 1000 independent runs in total). We compare each time IMED-GS to IMED algorithm for unstructured bandits and PO-IMED-GS algorithm. We also compare IMED-GS to OSSB algorithm for generic structured bandits. For Unimodal structure we add specific comparison with OSUB and UTS that are specialized to this structure, and for Lipschitz structure we add numerical comparison with CKL-UCB. We further report the IMED-GS run with setting $d = |\mathcal{A}| - 1$. Here the parameter Γ is set to $|\mathcal{A}|^{1.5}$ and $\xi = 1$. The sequence $(\gamma_t)_{t \geqslant 1}$ is set to $(\gamma_1 \log(t)^{-0.25})_{t \geqslant 1}$. The Lipschitz constant is set to $k = 0.03$ and applies for all the random configurations sampled for the experiment regarding the Lipschitz structure. For the Aggregate of bandits, the set of arms is decomposed as $\mathcal{A} = \mathcal{X} \times \mathcal{K}$, with $\mathcal{X} = [\![1\,,2]\!]$ and $\mathcal{K} = [\![1\,,9]\!]$. This means we assume an aggregate of 2 bandits with 9 arms each. The relationship matrix is set to $\omega_{x,x'} = 0.05\,|x - x'|$ and applies for all the random configurations sampled for the experiment regarding the Aggregate of bandits.

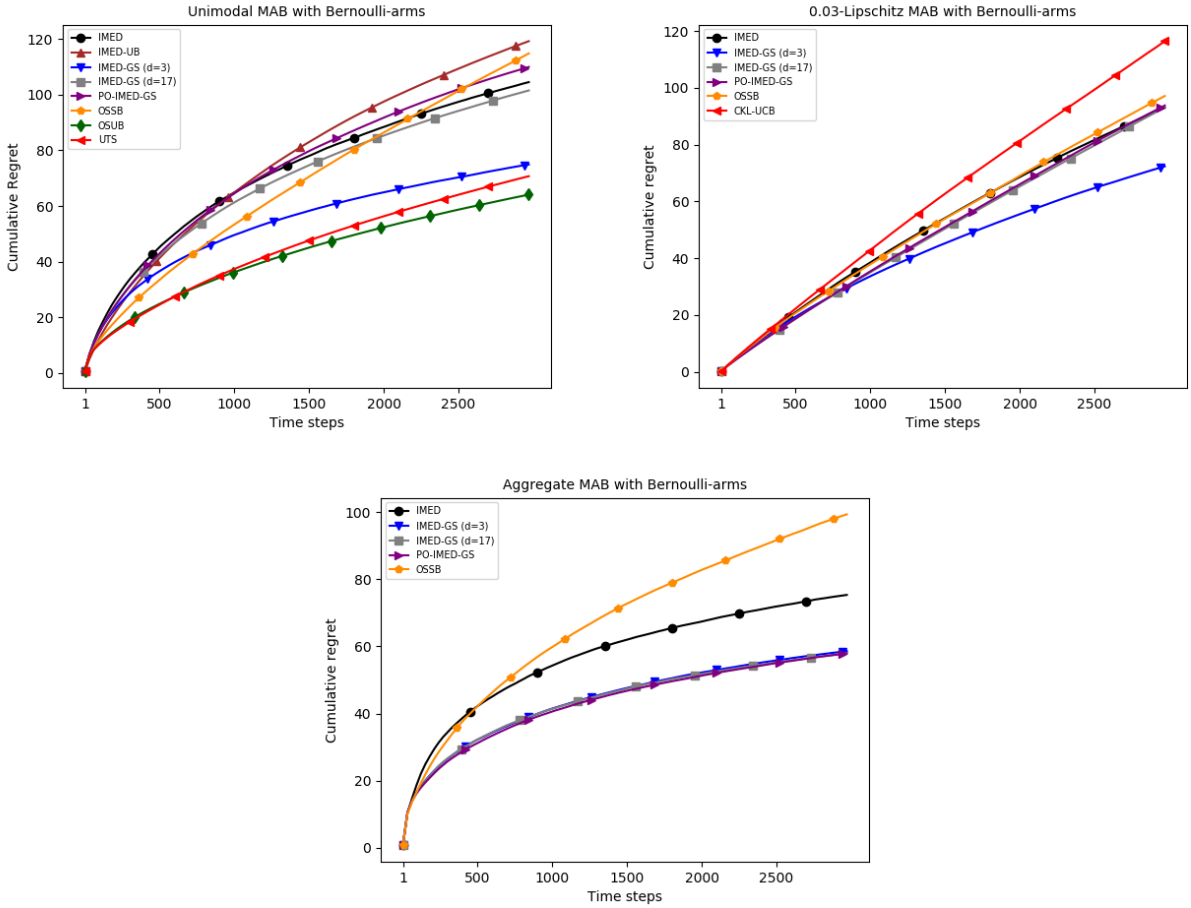


Figure A.1: Comparison of IMED−GS to other algorithms averaged over 100 randomly generated structured bandit instances.

Discussion. In the experiments, IMED−GS performs well for each considered structure and competes with specialized state-of-the-art algorithms for the Unimodal and Lipschitz structures. Furthermore, it seems that there is little benefit to use IMED−GS with no burning phase (when the parameter $d = |\mathcal{A}| - 1$) over PO−IMED−GS. This seems to confirm that reaching Pareto-Optimality provides good results in practice. Note that there is no burn-in phase in PO−IMED−GS algorithm, which makes relevant the comparison with IMED−GS with parameter d set to $|\mathcal{A}| - 1$. Finally, the theoretical benefit consisting in introducing the parameter d (less than $|\mathcal{A}| - 1$) in order to reduce the constant $C_{\xi,d,\varepsilon} = O(\varepsilon^{-2})$ in the upper bound on the regret (Theorem 3) implies practical benefit as it is highlighted in Figure A.1. Indeed, it appears that IMED−GS with parameter d set to 3 outperforms or competes with IMED−GS with parameter d set to $|\mathcal{A}| - 1$. This is a practical illustration of the benefit of having introduced the refined concentration inequality (Theorem 4). We recall that with the parameter d we replace the current informative sets of arms $\widehat{\mathcal{A}}_a(t)$ for $a \in \mathcal{A}$ with the sets $\widehat{\mathcal{A}}_a^{(d)}(t) \subset \widehat{\mathcal{A}}_a(t)$ consisting of $\{a\}$ plus the d-th most pulled arms from $\widehat{\mathcal{A}}_a(t) \setminus \{a\}$. It is justified by the fact that in the beginning, no structure can reasonably be exploited due to the poor estimates.

Appendix B

Generic Tools

B.1 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. These subsets of times appear in `IMED-UB` finite-time analysis. Similar subsets appear in finite-time analysis of `IMED-GS` (Section 4.5.3). Note that the definitions and the stated properties of these subsets of time steps are independent from the considered algorithms.

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

$$\mathcal{E}^+_{a,a'}(\varepsilon) \coloneqq \{t \in [\![1, T-1]\!] : \ a_{t+1} = a', \ N_{a'}(t) \leqslant N_a(t), \ \widehat{\mu}_a(t) \geqslant \mu_a + \varepsilon\} . \tag{B.1}$$

We similarly define

$$\mathcal{E}^-_{a,a'}(\varepsilon) \coloneqq \{t \in [\![1, T-1]\!] : \ a_{t+1} = a', \ N_{a'}(t) \leqslant N_a(t), \ \widehat{\mu}_a(t) \leqslant \mu_a - \varepsilon\} . \tag{B.2}$$

We also define

$$\mathcal{E}_{a,a'}(\varepsilon) = \mathcal{E}^+_{a,a'}(\varepsilon) \cup \mathcal{E}^-_{a,a'}(\varepsilon) . \tag{B.3}$$

Definition 5 (KL-log deviation)**.** *For $\varepsilon > 0$, the couple of arms $(a, a') \in \mathcal{A}^2$ shows ε^--KL-log deviation at time step $t \geqslant 1$ if the following conditions are satisfied*

$$\begin{aligned} &(1) \quad a_{t+1} = a' \\ &(2) \quad \widehat{\mu}_a(t) \leqslant \mu_a - \varepsilon \\ &(3) \quad \log(N_{a'}(t)) \leqslant N_a(t) \, \mathrm{KL}(\widehat{\mu}_a(t) | \mu_a - \varepsilon) + \log(N_a(t)) . \end{aligned}$$

For all couple of arms $(a, a') \in \mathcal{A}^2$ and for all accuracy $\varepsilon > 0$, let $\mathcal{K}^-_{a,a'}(\varepsilon)$ be the set of times where couple of arms (a, a') shows ε^--KL-log deviation, that is

$$\mathcal{K}^-_{a,a'}(\varepsilon) \coloneqq \left\{ t \in [\![1, T-1]\!] : \begin{array}{ll} (1) & a_{t+1} = a' \\ (2) & \widehat{\mu}_a(t) \leqslant \mu_a - \varepsilon \\ (3) & \log(N_{a'}(t)) \leqslant N_a(t) \, \mathrm{KL}(\widehat{\mu}_a(t) | \mu_a - \varepsilon) + \log(N_a(t)) \end{array} \right\} . \tag{B.4}$$

We note that

$$\mathcal{E}^-_{a,a'}(\varepsilon) \subset \mathcal{K}^-_{a,a'}(\varepsilon) .$$

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

Lemma 30 (Bounded subsets of times). *For $\varepsilon>0$, for $(a,a')\in\mathcal{A}^2$,*

$$\mathbb{E}_\nu\left[\left|\mathcal{E}^+_{a,a'}(\varepsilon)\right|\right],\ \mathbb{E}_\nu\left[\left|\mathcal{E}^-_{a,a'}(\varepsilon)\right|\right]\leqslant\frac{2\sigma_\varepsilon^2 e^{\varepsilon^2/2\sigma_\varepsilon^2}}{\varepsilon^2}$$

$$\mathbb{E}_\nu\left[\left|\mathcal{K}^-_{a,a'}(\varepsilon)\backslash\mathcal{E}^-_{a,a'}(\varepsilon)\right|\right]\leqslant 1+c_\varepsilon^{-1}+2C_\varepsilon\sqrt{\log(c_\varepsilon T)}\,,$$

where $\sigma_\varepsilon^2=\max\limits_{a\in\mathcal{A}}\left\{\mathbb{V}_{X\sim p(\mu')}(X)\colon\mu'\in[\mu_a-\varepsilon\,,\mu_a]\right\}$, $c_\varepsilon,C_\varepsilon>0$ are the constants involved in Theorem 6.

Proof. We start by proving $\mathbb{E}_\nu\left[\left|\mathcal{E}^-_{a,a'}(\varepsilon)\right|\right]\leqslant\frac{2\sigma_\varepsilon^2 e^{\varepsilon^2/2\sigma_\varepsilon^2}}{\varepsilon^2}$. The proof that $\mathbb{E}_\nu\left[\left|\mathcal{E}^+_{a,a'}(\varepsilon)\right|\right]\leqslant\frac{2\sigma_\varepsilon^2 e^{\varepsilon^2/2\sigma_\varepsilon^2}}{\varepsilon^2}$ is similar.

We write

$$\left|\mathcal{E}^-_{a,a'}(\varepsilon)\right|=\sum_{t=1}^{T-1}\mathbb{I}_{\{a_{t+1}=a',N_{a'}(t)\leqslant N_a(t),\ \mu_a-\widehat{\mu}_a(t)\geqslant\varepsilon\}}\,. \tag{B.5}$$

Considering the stopped stopping times $\tau_n=\inf\{t\geqslant 1,N_{a'}(t)=n\}$ we will rewrite the sum of indicators and use Lemma 32.

$$\begin{aligned}\left|\mathcal{E}^-_{a,a'}(\varepsilon)\right| &\leqslant \sum_{t\geqslant 1}\mathbb{I}_{\{a_{t+1}=a',\ N_{a'}(t)\leqslant N_a(t),\ \mu_a-\widehat{\mu}_a(t)\geqslant\varepsilon\}} \\ &\leqslant \sum_{n\geqslant 1}\mathbb{I}_{\{n-1\leqslant N_a(\tau_n-1),\ \mu_a-\widehat{\mu}_a(\tau_n-1)\geqslant\varepsilon\}} \\ &\leqslant 1+\sum_{n\geqslant 2}\mathbb{I}_{\{n-1\leqslant N_a(\tau_n-1),\ \mu_a-\widehat{\mu}_a(\tau_n-1)\geqslant\varepsilon\}}\,.\end{aligned} \tag{B.6}$$

Taking the expectation of Equation (B.6), by optional skipping it comes

$$\mathbb{E}_\nu\left[\left|\mathcal{E}^-_{a,a'}(\varepsilon)\right|\right]\leqslant 1+\sum_{n\geqslant 1}\mathbb{P}_\nu\left(\bigcup_{\substack{t\geqslant 1\\ N_a(t)\geqslant n}}\widehat{\mu}_a(t)\leqslant\mu_a-\varepsilon\right). \tag{B.7}$$

From Lemma 32, previous Equation (B.7) implies

$$\mathbb{E}_\nu\left[\left|\mathcal{E}^-_{a,a'}(\varepsilon)\right|\right]\leqslant 1+\sum_{n\geqslant 1}\exp(-n\,\mathrm{KL}(\mu_a-\varepsilon|\mu_a))\,. \tag{B.8}$$

From Lemma 31, previous Equation (B.8) implies

$$\mathbb{E}_\nu\left[\left|\mathcal{E}^-_{a,a'}(\varepsilon)\right|\right]\leqslant\sum_{n\geqslant 0}\exp\left(-n\varepsilon^2/2\sigma_\varepsilon^2\right)=\frac{1}{1-e^{-\varepsilon^2/2\sigma_\varepsilon^2}}\,, \tag{B.9}$$

where $\sigma_\varepsilon^2=\max\limits_{a\in\mathcal{A}}\left\{\mathbb{V}_{X\sim p(\mu')}(X)\colon\mu'\in[\mu_a-\varepsilon\,,\mu_a]\right\}$. Finally we note that

$$\frac{1}{1-e^{-\varepsilon^2/2\sigma_\varepsilon^2}}=\frac{e^{\varepsilon^2/2\sigma_\varepsilon^2}}{e^{\varepsilon^2/2\sigma_\varepsilon^2}-1}\leqslant\frac{2\sigma_\varepsilon^2 e^{\varepsilon^2/2\sigma_\varepsilon^2}}{\varepsilon^2}\,.$$

We now show that $\mathbb{E}_\nu\left[\left|\mathcal{K}^-_{a,a'}(\varepsilon)\right| \setminus \left|\mathcal{E}^-_{a,a'}(\varepsilon)\right|\right] \leqslant 1+c_\varepsilon^{-1}+2C_\varepsilon\sqrt{\log(c_\varepsilon T)}$.

We write

$$\begin{aligned}
&\left|\mathcal{K}^-_{a,a'}(\varepsilon)\setminus\mathcal{E}^-_{a,a'}(\varepsilon)\right| \\
&= \sum_{t=1}^{T-1} \mathbb{I}_{\{a_{t+1}=a',\ 1\leqslant N_a(t)<N_{a'}(t),\ \widehat{\mu}_a(t)\leqslant\mu_a-\varepsilon,\ \log(N_{a'}(t))\leqslant N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\mu_a-\varepsilon)+\log(N_a(t))\}}. \qquad \text{(B.10)}
\end{aligned}$$

Considering the stopped stopping times $\tau_n = \inf\{t\geqslant 1, N_{a'}(t)=n\}$ we shall rewrite the sum given by $\sum_{t\in[\![1,T-1]\!]} \mathbb{I}_{\{a_{t+1}=a',\ 1\leqslant N_a(t)<N_{a'}(t),\ \widehat{\mu}_a(t)\leqslant\mu_a-\varepsilon,\ \log(N_{a'}(t))\leqslant N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\mu_a-\varepsilon)+\log(N_a(t))\}}$ and use boundary crossing probabilities for one-dimensional exponential family distributions.

$$\begin{aligned}
&\left|\mathcal{K}^-_{a,a'}(\varepsilon)\setminus\mathcal{E}^-_{a,a'}(\varepsilon)\right| \\
&\leqslant \sum_{t=1}^{T-1} \mathbb{I}_{\{a_{t+1}=a',\ 1\leqslant N_a(t)<N_{a'}(t),\ \widehat{\mu}_a(t)\leqslant\mu_a-\varepsilon,\ \log(N_{a'}(t))\leqslant N_a(t)\,\mathrm{KL}(\widehat{\mu}_a(t)|\mu_a-\varepsilon)+\log(N_a(t))\}} \\
&= \sum_{t=1}^{T-1}\sum_{n=1}^{T-1} \mathbb{I}_{\{\tau_{n+1}=t+1\}}\mathbb{I}_{\{1\leqslant N_a(\tau_{n+1}-1)<n,\ \widehat{\mu}_a(\tau_{n+1}-1)\leqslant\mu_a-\varepsilon\}} \times \\
&\quad \mathbb{I}_{\{\log(n)\leqslant N_a(\tau_{n+1}-1)\,\mathrm{KL}(\widehat{\mu}_a(\tau_{n+1}-1)|\mu_a-\varepsilon)+\log(N_a(\tau_{n+1}-1))\}} \\
&= \sum_{n=1}^{T-1} \mathbb{I}_{\{1\leqslant N_a(\tau_{n+1}-1)<n,\ \widehat{\mu}_a(\tau_{n+1}-1)\leqslant\mu_a-\varepsilon\}} \times \\
&\quad \mathbb{I}_{\{\log(n)\leqslant N_a(\tau_{n+1}-1)\,\mathrm{KL}(\widehat{\mu}_a(\tau_{n+1}-1)|\mu_a-\varepsilon)+\log(N_a(\tau_{n+1}-1))\}} \sum_{t=1}^{T-1} \mathbb{I}_{\{\tau_{n+1}=t+1\}} \\
&\leqslant \sum_{n=1}^{T-1} \mathbb{I}_{\{1\leqslant N_a(\tau_{n+1}-1)<n,\ \widehat{\mu}_a(\tau_{n+1}-1)\leqslant\mu_a-\varepsilon,\ \log(n)\leqslant N_a(\tau_{n+1}-1)\,\mathrm{KL}(\widehat{\mu}_a(\tau_{n+1}-1)|\mu_a-\varepsilon)+\log(N_a(\tau_{n+1}-1))\}} \\
&= \sum_{n=2}^{T-1} \mathbb{I}_{\{1\leqslant N_a(\tau_{n+1}-1)<n,\ \widehat{\mu}_a(\tau_{n+1}-1)\leqslant\mu_a-\varepsilon,\ \log(n)\leqslant N_a(\tau_{n+1}-1)\,\mathrm{KL}(\widehat{\mu}_a(\tau_{n+1}-1)|\mu_a-\varepsilon)+\log(N_a(\tau_{n+1}-1))\}}. \qquad \text{(B.11)}
\end{aligned}$$

From Equation (B.11), we get

$$\begin{aligned}
&\left|\mathcal{K}^-_{a,a'}(\varepsilon)\setminus\mathcal{E}^-_{a,a'}(\varepsilon)\right| \qquad \text{(B.12)} \\
&\leqslant \sum_{n=2}^{T-1} \mathbb{I}_{\{1\leqslant N_a(\tau_{n+1}-1)<n,\ N_a(\tau_{n+1}-1)\,\mathrm{KL}(\widehat{\mu}_a(\tau_{n+1}-1)|\mu_a-\varepsilon)\geqslant\log(n/N_a(\tau_{n+1}-1)),\ \widehat{\mu}_a(\tau_{n+1}-1)<\mu_a-\varepsilon\}}.
\end{aligned}$$

Taking the expectation of Equation (B.12), it comes

$$\begin{aligned}
&\mathbb{E}_\nu\left[\left|\mathcal{K}^-_{a,a'}(\varepsilon)\setminus\mathcal{E}^-_{a,a'}(\varepsilon)\right|\right] \qquad \text{(B.13)} \\
&\leqslant \sum_{n=2}^{T-1} \mathbb{P}_\nu\left(\bigcup_{\substack{t\geqslant 1\\ \widehat{\mu}_a(t)<\mu_a-\varepsilon\\ 1\leqslant N_a(t)\leqslant n}} N_a(t)\mathrm{KL}(\widehat{\mu}_a(t)|\mu_a-\varepsilon)\geqslant\log(n/N_a(t))\right).
\end{aligned}$$

From Theorem 6, previous Equation (B.13) implies

$$\begin{aligned}
& \mathbb{E}_\nu\big[\big|\mathcal{K}^-_{a,a'}(\varepsilon)\setminus\mathcal{E}^-_{a,a'}(\varepsilon)\big|\big] && \text{(B.14)}\\
\leqslant\ & 1+c_\varepsilon^{-1}+C_\varepsilon\sum_{n\geqslant 1+c_\varepsilon^{-1}}^{T-1}\frac{c_\varepsilon}{c_\varepsilon n\sqrt{\log(c_\varepsilon n)}}\\
\leqslant\ & 1+c_\varepsilon^{-1}+C_\varepsilon\int_{c_\varepsilon^{-1}}^{T}\frac{c_\varepsilon\,dx}{c_\varepsilon x\sqrt{\log(c_\varepsilon x)}}\\
=\ & 1+c_\varepsilon^{-1}+2C_\varepsilon\sqrt{\log(c_\varepsilon T)}\,. && \text{(B.15)}
\end{aligned}$$

□

B.2 Concentration of measure

In this section, Pinsker's inequality for one-dimensional exponential family distributions is reminded. Please refer to Lemma 3 from Cappé et al. (2013) for more insights. We also state two concentration results from Maillard (2018).

Lemma 31 (Pinsker's inequality)**.** *For $\mu<\mu'$, it holds that*

$$\mathrm{KL}(\mu|\mu')\geqslant\frac{(\mu'-\mu)^2}{2\sigma^2}\,,$$

where $\sigma^2=\max\Big\{\mathbb{V}_{X\sim p(\mu'')}(X)\colon\mu''\in[\mu\,,\mu']\Big\}$.

Lemma 32 (Time-uniform concentration)**.** *For all arm $a\in\mathcal{A}$, for $x<\mu_a$, $m\geqslant 1$, we have*

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

Theorem 6 (Boundary crossing probabilities)**.** *For all arm $a\in\mathcal{A}$, for all $\varepsilon>0$, for all $n\geqslant 1$, we have*

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

where $c_\varepsilon, C_\varepsilon>0$ are explained in Maillard (2018).

Appendix C

Routine Bandits : Proof and Additional Experiments

C.1 Proof of Proposition 5

Let us denote by $\mathcal{S}$ the routine bandit setting and by $\mathcal{S}_0$ the setting resulting from the routine bandit setting and the *additional assumption* that now the sequence of bandits $(b_\star^h)_{h\in[\![1,H]\!]}$ is known to the learner. Then, since a consistent strategy for $\mathcal{S}$ is also consistent for $\mathcal{S}_0$ (in the sense of Definition 3), we deduce Proposition 5 from Lemma 33.

Lemma 33 (Lower bounds on the regret for $\mathcal{S}_0$)**.** *Let us consider a consistent strategy for the setting $\mathcal{S}_0$. Then, for all configuration $\nu\in\mathcal{D}$, it must be that*

$$\liminf_{T\to\infty}\frac{R(\nu,H,T)}{\log(T)}\geqslant c_\nu^\star\coloneqq\sum_{b\in\mathcal{B}}\sum_{a\neq a_b^\star}\frac{\Delta_{a,b}}{\mathrm{KL}(\mu_{a,b}|\mu_b^\star)}\,.$$

Proof. Since the sequence of bandits $(b_\star^h)_{h\in[\![1,H]\!]}$ is known to the learner and since there is no shared information between the bandits at first glance, the setting $\mathcal{S}_0$ amounts to consider each of the $|\mathcal{B}|$ bandits $(\nu_b)_{b\in\mathcal{B}}$ as a separate problem, where $\nu_b\coloneqq(\nu_{a,b})_{a\in\mathcal{A}}$ for $b\in\mathcal{B}$. Then, from the known lower bound on the regret for the classical multi-armed bandit problem Lai and Robbins (1985), we get under the assumption of consistency for all bandit $b\in\mathcal{B}$,

$$\liminf_{T\to\infty}\frac{1}{\log(N_b(H,T))}\sum_{a\neq a_b^\star}N_{a,b}(T)\Delta_{a,b}\geqslant\sum_{a\neq a_b^\star}\frac{\Delta_{a,b}}{\mathrm{KL}(\mu_{a,b}|\mu_b^\star)},\ \text{where } N_b(H,T)=\beta_b^H HT\,.$$

From previous inequalities and Eq. 6.2, we conclude that

$$\begin{aligned}\liminf_{T\to\infty}\frac{R(\nu,H,T)}{\log(T)}&\geqslant\sum_{b\in\mathcal{B}}\liminf_{T\to\infty}\frac{\log\left(\beta_b^H HT\right)}{\log(T)}\liminf_{T\to\infty}\frac{1}{\log(\beta_b^H HT)}\sum_{a\neq a_b^\star}N_{a,b}(T)\Delta_{a,b}\\&\geqslant\sum_{b\in\mathcal{B}}\sum_{a\neq a_b^\star}\frac{\Delta_{a,b}}{\mathrm{KL}(\mu_{a,b}|\mu_b^\star)}\,,\end{aligned}$$

by Fatou's Lemma and since we have $\liminf_n u_n v_n\geqslant\liminf_n u_n\liminf_n v_n$ for all positive real-valued sequences u,v. □

C.2 Proof of Theorem 5

From Proposition 6, we have the following inequality

$$N_{a,b}(H,T) \leqslant \frac{f(\beta_b^H HT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\mu_b^\star)} + \sum_{h=1}^{H} \mathbb{I}_{\{b_\star^h=b\}}\left[T_{\nu,\varepsilon}^h + 4\left|\mathcal{C}_\varepsilon^h\right| + \left|\overline{\mathcal{C}}_\varepsilon^h\right| + \frac{f(hT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\mu_b^\star)}\sum_{k=1}^{h}\frac{8\left|\mathcal{C}_\varepsilon^k\right|}{T} + \mathbb{I}_{\{T\in\mathcal{T}^k\}}\right], \tag{C.1}$$

where for all $h \geqslant 1$, $\mathbb{P}_\nu\left(T \in \mathcal{T}^h\right) \leqslant 1/T(T+1)$, $\mathbb{E}_\nu\left[\left|\mathcal{C}_\varepsilon^h\right|\right], \mathbb{E}_\nu\left[\left|\overline{\mathcal{C}}_\varepsilon^h\right|\right] \leqslant 4\left|\mathcal{A}\right|\varepsilon^{-2}+3$ according to Lemma 38 and $T_{\nu,\varepsilon}^h \leqslant \tau_\nu^h$ according to Lemma 34 stated below.
By taking the expectation in Eq. C.1 then it comes

$$\begin{aligned}
&\mathbb{E}_\nu[N_{a,b}(H,T)] \\
&\leqslant \frac{f(\beta_b^H HT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\mu_b^\star)} \\
&\quad + \sum_{h=1}^{H} \mathbb{I}_{\{b_\star^h=b\}}\left[\tau_\nu^h + 5\times\left(4\left|\mathcal{A}\right|\varepsilon^{-2}+3\right) + \frac{f(hT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\mu_b^\star)}\sum_{k=1}^{h}\frac{32\left|\mathcal{A}\right|\varepsilon^{-2}+24}{T} + 1/T(T+1)\right].
\end{aligned}$$

We conclude the proof of Theorem 5 by using the two following inequalities

$$\begin{aligned}
&4\left|\mathcal{A}\right|\varepsilon^{-2} + 3 \leqslant 4\left|\mathcal{A}\right|(\varepsilon^{-2}+1) \\
&\frac{32\left|\mathcal{A}\right|\varepsilon^{-2}+24}{T} + \frac{1}{T(T+1)} \leqslant \frac{32\left|\mathcal{A}\right|(\varepsilon^{-2}+1)}{T}.
\end{aligned}$$

Lemma 34 (Upper bound on $T_{\nu,\varepsilon}^h$)**.** *With the same notations as Proposition 6, for all* $0<\varepsilon<\varepsilon_\nu$,

$$T_{\nu,\varepsilon}^h \leqslant \tau_\nu^h := 2\varphi\left(8|\mathcal{A}|\left[\varepsilon_\nu^{-2} + 65\gamma_\nu^{-2}\log\left(128\left|\mathcal{A}\right|(4h)^{1/3}\gamma_\nu^{-2}\right)\right]\right),$$

where $\varphi\colon x \geqslant 1 \mapsto x\log(x)$.

Proof. We first show that

$$t_\nu^h < 130\left|\mathcal{A}\right|\gamma_\nu^{-2}\log\left(128(4h)^{1/3}\left|\mathcal{A}\right|\gamma_\nu^{-2}\right). \tag{C.2}$$

Let us consider $t \geqslant 3\left|\mathcal{A}\right|$. We have

$$\mathrm{d}\left(\frac{t}{|\mathcal{A}|}, \delta^h(t)\right) = \sqrt{2\left(1+\frac{|\mathcal{A}|}{t}\right)\frac{\log\left(4\left|\mathcal{A}\right|^3(h-1)\sqrt{t/\left|\mathcal{A}\right|+1}(t/\left|\mathcal{A}\right|)(t/\left|\mathcal{A}\right|+1/\left|\mathcal{A}\right|)\right)}{t/\left|\mathcal{A}\right|}}.$$

Since $1/|\mathcal{A}|<1$ and $t/\left|\mathcal{A}\right| \geqslant 3$, we have

$$\sqrt{t/\left|\mathcal{A}\right|+1}(t/\left|\mathcal{A}\right|)(t/\left|\mathcal{A}\right|+1/\left|\mathcal{A}\right|) \leqslant \sqrt{t/\left|\mathcal{A}\right|+1}(t/\left|\mathcal{A}\right|)(t/\left|\mathcal{A}\right|+1) \leqslant (t/\left|\mathcal{A}\right|)^3.$$

Then, since $1+|\mathcal{A}|/t<1+1/3$ and $h-1\leqslant h$, we get

$$\mathrm{d}\left(\frac{t}{|\mathcal{A}|}, \delta^h(t)\right) \leqslant \sqrt{\frac{8\left|\mathcal{A}\right|(4h)^{1/3}}{\Phi((4h)^{1/3}t)}}$$

where $\Phi : x \geqslant 3 \mapsto x/\log(x) \geqslant \Phi(3)$. $\Phi(\cdot)$ is a one-to-one function and $\forall y \geqslant \Phi(3), \Phi^{-1}(y) \leqslant y\log(y) + 2\log(y)$. Thus we have

$$\begin{aligned}\gamma_\nu \leqslant 4\mathrm{d}\left(\frac{t}{|\mathcal{A}|}, \delta^h(t)\right) \quad \Rightarrow \quad t &\leqslant (4h)^{-1/3}\Phi^{-1}\left(128(4h)^{1/3}\,|\mathcal{A}|\,\gamma_\nu^{-2}\right) \\ &\leqslant 128\,|\mathcal{A}|\,\gamma_\nu^{-2}\log\left(128(4h)^{1/3}\,|\mathcal{A}|\,\gamma_\nu^{-2}\right) \\ &\quad + 2(4h)^{-1/3}\log\left(128(4h)^{1/3}\,|\mathcal{A}|\,\gamma_\nu^{-2}\right) .\end{aligned}$$

In particular, we get the following implication

$$\gamma_\nu \leqslant 4\mathrm{d}\left(\frac{t}{|\mathcal{A}|}, \delta^h(t)\right) \quad \Rightarrow \quad t < 130\,|\mathcal{A}|\,\gamma_\nu^{-2}\log\left(128(4h)^{1/3}\,|\mathcal{A}|\,\gamma_\nu^{-2}\right) - 1$$

and $t_\nu^h < 130\,|\mathcal{A}|\,\gamma_\nu^{-2}\log\left(128(4h)^{1/3}\,|\mathcal{A}|\,\gamma_\nu^{-2}\right)$.

Furthermore, we have

$$\sum_{a \neq a_\star^h} \frac{f(t)}{\mathrm{KL}(\mu_a^h + \varepsilon|\mu_\star^h)} < 2\,|\mathcal{A}|\,\varepsilon_\nu^{-2} f(t) . \tag{C.3}$$

By combining Eq. C.2 and Eq. C.3, from the definition of $T_{\nu,\varepsilon}^h$ (see Proposition 6) we get

$$T_{\nu,\varepsilon}^h \leqslant \max\left\{t \geqslant 10 :\ t - 8\,|\mathcal{A}|\left(\varepsilon_\nu^{-2} f(t) - 65\gamma_\nu^{-2}\log\left(128(4h)^{1/3}\,|\mathcal{A}|\,\gamma_\nu^{-2}\right)\right) \leqslant 0\right\} . \tag{C.4}$$

We finally prove Lemma 34 by applying Lemma 35 with $c = 8\,|\mathcal{A}|\,\varepsilon_\nu^{-2}$ and $c' = 130\,|\mathcal{A}|\,\gamma_\nu^{-2}\log\left(128(4h)^{1/3}\,|\mathcal{A}|\,\gamma_\nu^{-2}\right)$. □

Lemma 35. *For all $c, c' > 10$, it holds*

$$\max\left\{t \geqslant 10 :\ t - cf(t) - c' \leqslant 0\right\} \leqslant 2\varphi(c + c') ,$$

where $\varphi : x \geqslant 1 \mapsto x\log(x)$.

Proof. It can be shown that $(t - cf(t) - c')_{t \geqslant 2\varphi(c+c')}$ is non-decreasing by standard derivative analysis and that $2\varphi(c+c') - cf(2\varphi(c+c')) - c \geqslant 0$. □

In the following we prove the results stated in Section 6.3.

C.2.1 Proof of Lemma 18

In this subsection we control the number previously encountered bandits falsely identified as different to the current one (see Definition 6) in addition to false positives and prove Lemma 36, an extension of Lemma 18.

Definition 6 (False negative)**.** *At period $h \geqslant 2$ and step $t \geqslant 1$, a previous period $k \in [\![1, h-1]\!]$ is called a false negative if the test $\mathrm{T}^{k,h}(t)$ is false while previous bandit $b_\star^k$ corresponds to current bandit $b_\star^h$.*

We prove necessary conditions for having false positives or false negatives.

Lemma 36 (Condition for false positives/negatives)**.** *At period $h \geqslant 2$ and time step $t > |\mathcal{A}|$, for all period $k \in [\![1, h-1]\!]$, , with probability $1 - 1/(h-1)t(t+1)$,*

$$\begin{aligned} k \text{ is a false positive} \quad &\Longrightarrow \quad b_\star^k \neq b_\star^h \quad \text{and} \quad \min_{a \in \mathcal{A}} \left|\mu_{a,b_\star^h} - \mu_{a,b_\star^k}\right| \leqslant 4\mathrm{d}\left(\frac{t}{|\mathcal{A}|}, \delta^h(t)\right) \\ k \text{ is a false negative} \quad &\Longleftrightarrow \quad b_\star^k = b_\star^h \quad \text{and} \quad \overline{a}_t^k \neq \overline{a}_t^h . \end{aligned}$$

Proof. From Lemma 42, with probability $1-4\,|\mathcal{A}|\,\delta^h(t)=1-1/(h-1)t(t+1)$, it holds,

$$\forall a \in \mathcal{A}, \quad \left|\widehat{\mu}_a^h(t) - \mu_a^h\right| \leqslant \mathrm{d}\big(N_a^h(t), \delta^h(t)\big) \quad \text{and} \quad \left|\widehat{\mu}_a^k(T) - \mu_a^k\right| \leqslant \mathrm{d}\big(N_a^k(T), \delta^h(t)\big)\,. \tag{C.5}$$

False negative: Here we assume that $b_\star^k = b_\star^h$. By the triangle inequality, this implies

$$\forall a \in \mathcal{A}, \quad \left|\widehat{\mu}_a^h(t) - \widehat{\mu}_a^k(T)\right| = \left|\big(\widehat{\mu}_a^h(t) - \mu_a^h\big) - \big(\widehat{\mu}_a^k(T) - \mu_a^k\big)\right| \leqslant \left|\widehat{\mu}_a^h(t) - \mu_a^h\right| + \left|\widehat{\mu}_a^k(T) - \mu_a^T\right|\,. \tag{C.6}$$

By combining Eq.C.5 and C.6, with probability $1-1/(h-1)t(t+1)$, we have

$$\forall a \in \mathcal{A}, \quad \left|\widehat{\mu}_a^h(t) - \widehat{\mu}_a^k(T)\right| - \mathrm{d}\big(N_a^h(t), \delta^h(t)\big) - \mathrm{d}\big(N_a^k(T), \delta^h(t)\big) \leqslant 0\,.$$

Then, from the definitions of the random variables $(Z_a^{k,h}(t))_{a\in\mathcal{A}}$ (Eq. 6.3) and the test $\mathrm{T}^{k,h}(t)$ (Eq. 6.4), this implies with probability $1-1/(h-1)t(t+1)$,

$$\max_{a\in\mathcal{A}} Z_a^{k,h}(t) \leqslant \infty \cdot \mathbb{I}_{\left\{\overline{a}_t^h \neq \overline{a}_\star^k\right\}}, \qquad \mathrm{T}^{k,h}(t) = \left(\overline{a}_t^h = \overline{a}_\star^k\right).$$

Thus, with probability $1-1/(h-1)t(t+1)$, period k is a false negative if, and only if, $\overline{a}_t^h \neq \overline{a}_\star^k$.

False positive: Here we assume that period k is a false positive. In particular, we have $b_\star^k \neq b_\star^h$. By the triangle inequality, this implies

$$\forall a \in \mathcal{A}, \quad \left|\widehat{\mu}_a^h(t) - \widehat{\mu}_a^k(T)\right| \geqslant \left|\mu_a^h - \mu_a^k\right| - \left|\widehat{\mu}_a^h(t) - \mu_a^h\right| - \left|\widehat{\mu}_a^k(T) - \mu_a^T\right|\,. \tag{C.7}$$

By combining Eq.C.5 and C.7, with probability $1-1/(h-1)t(t+1)$, we have

$$\forall a \in \mathcal{A}, \quad Z_a^{k,h}(t) \geqslant \infty \cdot \mathbb{I}_{\left\{\overline{a}_t^h \neq \overline{a}_\star^k\right\}} + \min_{a\in\mathcal{A}} \left|\mu_a^h - \mu_a^k\right| - 2\mathrm{d}\big(N_a^h(t), \delta^h(t)\big) - 2\mathrm{d}\big(N_a^k(T), \delta^h(t)\big)\,. \tag{C.8}$$

Since period is assumed to be a false positive, we have $\max_{a\in\mathcal{A}} Z^{k,h}(t) \leqslant 0$ and Eq. C.8 implies that, with probability $1-1/(h-1)t(t+1)$,

$$\overline{a}_t^h = \overline{a}_\star^k, \qquad \min_{a\in\mathcal{A}} \left|\mu_a^h - \mu_a^k\right| \leqslant 2\mathrm{d}\Big(N_{\overline{a}_t^h}^h(t), \delta^h(t)\Big) + 2\mathrm{d}\Big(N_{\overline{a}_\star^k}^k(T), \delta^h(t)\Big)\,. \tag{C.9}$$

Since $N_{\overline{a}_t^h}^h(t) \geqslant t/\left|\mathcal{A}\right|$, $N_{\overline{a}_\star^k}^k(T) \geqslant T/\left|\mathcal{A}\right|$ ($\overline{a}_t^h$ and $\overline{a}_\star^k$ are most pulled arms) and $\delta^h(T) \leqslant \delta^h(t)$, the monotonic properties of $\mathrm{d}(\cdot,\cdot)$ and Eq. C.9, imply that, with probability $1-1/(h-1)t(t+1)$,

$$\min_{a\in\mathcal{A}} \left|\mu_a^h - \mu_a^k\right| \leqslant 2\mathrm{d}\left(\frac{t}{|\mathcal{A}|}, \delta^h(t)\right) + 2\mathrm{d}\left(\frac{T}{|\mathcal{A}|}, \delta^h(T)\right)\,.$$

We conclude the proof of Lemma 36 by using Lemma 37 stated below. □

Lemma 37 (Monotonic properties of $\mathrm{d}(\cdot,\cdot)$)**.** *For all period $h \geqslant 2$, $\big(\mathrm{d}\big(t/\left|\mathcal{A}\right|, \delta^h(t)\big)\big)_{t>|\mathcal{A}|}$ is non-increasing.*

Proof. For all time step $t > |\mathcal{A}|$, a direct calculation gives

$$\mathrm{d}\left(\frac{t}{|\mathcal{A}|}, \delta^h(t)\right) = \sqrt{2\,|\mathcal{A}|\left(1+\frac{|\mathcal{A}|}{t}\right)\left(\frac{1}{2}\frac{\log(t/\left|\mathcal{A}\right|+1)}{t} + \frac{\log(4\,|\mathcal{A}|\,(h-1))}{t} + \frac{\log(t+1)}{t} + \frac{\log(t)}{t}\right)}.$$

Then, in order to prove Lemma 37, it is sufficient to note that $(\log(t/|\mathcal{A}|+1)/t)_{t\geqslant 1}$, $(\log(t+1)/t)_{t\geqslant 2}$ and $(\log(t)/t)_{t\geqslant 3}$ are non-increasing. □

C.2.2 Proof of Lemma 19

Let us consider the subsets of times when the mean of the current pulled arm is poorly estimated

$$\mathcal{E}^h_{a,\varepsilon} \coloneqq \left\{t > |\mathcal{A}| :\ a^h_{t+1} = a \text{ and } \left|\widehat{\mu}^h_a(t) - \mu^h_a\right| > \varepsilon\right\} \qquad \mathcal{E}^h_\varepsilon \coloneqq \bigcup_{a \neq a^h_\star} \mathcal{E}^h_{a,\varepsilon}$$

$$\overline{\mathcal{E}}^h_{a,\varepsilon} \coloneqq \left\{t \geqslant t^h_\nu :\ t \notin \mathcal{T}^h,\ a^h_{t+1} = a \text{ and } \left|\overline{\mu}^h_a(t) - \mu^h_a\right| > \varepsilon\right\} \qquad \overline{\mathcal{E}}^h_\varepsilon \coloneqq \bigcup_{a \neq a^h_\star} \overline{\mathcal{E}}^h_{a,\varepsilon}$$

and the subsets of times when the best arm $a^h_\star$ is below its mean

$$\mathcal{U}^h_a \coloneqq \left\{t > |\mathcal{A}| :\ a^h_{t+1} = a \text{ and } u^h_{a^h_\star}(t) = U^h_{a^h_\star}(t) < \mu^h_\star\right\} \qquad \mathcal{U}^h \coloneqq \bigcup_{a \neq a^h_\star} \mathcal{U}^h_a \,.$$

$$\overline{\mathcal{U}}^h_a \coloneqq \left\{t \geqslant t^h_\nu :\ t \notin \mathcal{T}^h,\ a^h_{t+1} = a \text{ and } u^h_{a^h_\star}(t) = \overline{U}^h_{a^h_\star}(t) < \mu^h_\star\right\} \qquad \overline{\mathcal{U}}^h \coloneqq \bigcup_{a \neq a^h_\star} \overline{\mathcal{U}}^h_a \,.$$

Then we have

$$\mathcal{C}^h_{a,\varepsilon} = \mathcal{T}^h_a \cup \mathcal{E}^h_{a,\varepsilon} \cup \mathcal{U}^h_a \qquad \mathcal{C}^h_\varepsilon = \mathcal{T}^h \cup \mathcal{E}^h_\varepsilon \cup \mathcal{U}^h$$
$$\overline{\mathcal{C}}^h_{a,\varepsilon} = \mathcal{T}^h_a \cup \mathcal{E}^h_{a,\varepsilon} \cup \overline{\mathcal{U}}^h_a \qquad \overline{\mathcal{C}}^h_\varepsilon = \mathcal{T}^h \cup \mathcal{E}^h_\varepsilon \cup \overline{\mathcal{U}}^h$$

and deduce Lemma 19 from the extended Lemma 38.

Lemma 38 (Bounded subsets of times)**.** *For all period $h \geqslant 2$, for all arm $a \in \mathcal{A}$, for all $0 < \varepsilon < \varepsilon_\nu$,*

$$\forall t \in [\![1, T]\!],\ \mathbb{P}_\nu\big(t \in \mathcal{T}^h\big) \leqslant \frac{1}{t(t+1)}, \quad \mathbb{E}_\nu\left[\left|\mathcal{E}^h_{a,\varepsilon}\right|\right], \mathbb{E}_\nu\left[\left|\overline{\mathcal{E}}^h_{a,\varepsilon}\right|\right] \leqslant 4\varepsilon^{-2}, \quad \mathbb{E}_\nu\left[\left|\mathcal{U}^h\right|\right], \mathbb{E}_\nu\left[\left|\overline{\mathcal{U}}^h\right|\right] \leqslant 2\,.$$

This implies

$$\mathbb{E}_\nu\left[\left|\mathcal{T}^h\right|\right] \leqslant 1, \qquad \mathbb{E}_\nu\left[\left|\mathcal{C}^h_{a,\varepsilon}\right|\right], \mathbb{E}_\nu\left[\left|\overline{\mathcal{C}}^h_{a,\varepsilon}\right|\right] \leqslant 4\varepsilon^{-2} + 3, \qquad \mathbb{E}_\nu\left[\left|\mathcal{C}^h_\varepsilon\right|\right], \mathbb{E}_\nu\left[\left|\overline{\mathcal{C}}^h_\varepsilon\right|\right] \leqslant 4\,|\mathcal{A}|\,\varepsilon^{-2} + 3\,.$$

Proof. Subset $\mathcal{T}^h$: From Lemma 38 and the definition of t^h_ν (see Eq. 6.8), for all $t \geqslant t^h_\nu$, with probability $1 - 1/t(t+1)$, there is no false positive and if a previous period $k \in [\![1, h-1]\!]$ is a false negative then $b^k_\star = b^h_\star$ and $\overline{a}^k_\star \neq \overline{a}^h_t$ (the most pulled arms are different). From the definition of $\mathcal{T}^h$ (see Eq. 6.9) this implies that for all $t \geqslant t^h_\nu$, with probability $1 - 1/t(t+1)$, $t \notin \mathcal{T}^h$. That is $\forall t \geqslant t^h_\nu$, $\mathbb{P}_\nu\big(t \in \mathcal{T}^h\big) \leqslant 1/t(t+1)$. Since on the other hand, we have

$$\left|\mathcal{T}^h\right| = \sum_{t=t^h_\nu}^{T} \mathbb{I}_{\left\{t \in \mathcal{T}^h\right\}}\,,$$

by taking expectation on both sides, it comes

$$\mathbb{E}_\nu\left[\left|\mathcal{T}^h\right|\right] = \sum_{t=t^h_\nu}^{T} \mathbb{P}_\nu\big(t \in \mathcal{T}^h\big) \leqslant \sum_{t=t^h_\nu}^{T} \frac{1}{t(t+1)} \leqslant 1\,.$$

We note that for $1 \leqslant t < t^h_\nu$, it holds that $t \notin \mathcal{T}^h$ and $\mathbb{P}_\nu\big(t \in \mathcal{T}^h\big) = 0 \leqslant 1/t(t+1)$.

Subset $\mathcal{E}^h_{a,\varepsilon}$:
Since we have

$$\left|\mathcal{E}^h_{a,\varepsilon}\right| = \sum_{t>|\mathcal{A}|}^{T} \mathbb{I}_{\left\{a^h_{t+1}=a,\ \left|\widehat{\mu}^h_a(t) - \mu^h_a\right| > \varepsilon\right\}}\,,$$

by taking the expectation on both sides, it comes

$$\mathbb{E}_\nu\left[\left|\mathcal{E}^h_{a,\varepsilon}\right|\right] \leqslant \sum_{t=1}^{T} \mathbb{P}_\nu\left(a^h_{t+1}=a,\ \left|\widehat{\mu}^h_a(t)-\mu^h_a\right|>\varepsilon\right). \tag{C.10}$$

Then, by combining Eq. C.10 and Lemma 41, we prove $\mathbb{E}_\nu\left[\left|\mathcal{E}^h_{a,\varepsilon}\right|\right] \leqslant 4\varepsilon^{-2}$.

Subset $\overline{\mathcal{E}}^h_{a,\varepsilon}$: From the definitions of t^h_ν and $\mathcal{T}^h$ (Eq. 6.8 and 6.9), we get the following inclusion

$$\left\{t \geqslant t^h_\nu :\ t\notin\mathcal{T}^h,\ a^h_{t+1}=a,\ \left|\overline{\mu}^h_a(t)-\mu^h_a\right|>\varepsilon\right\} \subset \left\{t\geqslant t^h_\nu :\ a^h_{t+1}=a,\ \left|\widehat{\mu}^{\mathcal{K}^h_\star(t),h}_a(t)-\mu^h_a\right|>\varepsilon\right\}, \tag{C.11}$$

where $\mathcal{K}^h_\star(t) \coloneqq \left\{k\in[\![1,h-1]\!] : b^k_\star = b^h_\star \text{ and } \overline{a}^k_\star = \overline{a}^h_t\right\}$ and $N^{\mathcal{K},h}_a(t) = \sum_{k\in\mathcal{K}} N^k_a(T) + N^h_a(t)$,
$S^{\mathcal{K},h}_a(t) = \sum_{k\in\mathcal{K}} S^k_a(T) + S^h_a(t)$, $\widehat{\mu}^{\mathcal{K},h}_a(t) = S^{\mathcal{K},h}_a(t)/N^{\mathcal{K},h}_a(t)$, $\forall \mathcal{K}\subset\mathcal{K}^h \coloneqq \left\{k\in[\![1,h-1]\!] : b^k_\star = b^h_\star\right\}$.

Thus, by defining $\mathcal{K}_t \coloneqq \mathcal{K}^h_\star(t)$ if $t\geqslant t^h_\nu$ and $t\notin\mathcal{T}^h$, $\emptyset$ otherwise, Eq. C.11 implies

$$\forall t\geqslant t^h_\nu, \mathbb{P}_\nu\left(t\notin\mathcal{T}^h, a^h_{t+1}=a, \left|\overline{\mu}^h_a(t)-\mu^h_a\right|>\varepsilon\right) \leqslant \mathbb{P}_\nu\left(a^h_{t+1}=a, \left|\widehat{\mu}^{\mathcal{K}_t,h}_a(t)-\mu^h_a\right|>\varepsilon\right). \tag{C.12}$$

Since we have

$$\left|\overline{\mathcal{E}}^h_{a,\varepsilon}\right| = \sum_{t=t^h_\nu}^{T} \mathbb{I}_{\left\{t\notin\mathcal{T}^h,\ a^h_{t+1}=a,\ \left|\overline{\mu}^h_a(t)-\mu^h_a\right|>\varepsilon\right\}},$$

by taking the expectation on both sides and using inequalities from Eq.C.12, it comes

$$\begin{aligned}\mathbb{E}_\nu\left[\left|\overline{\mathcal{E}}^h_{a,\varepsilon}\right|\right] &= \sum_{t=t^h_\nu}^{T} \mathbb{P}_\nu\left(t\notin\mathcal{T}^h,\ a^h_{t+1}=a,\ \left|\overline{\mu}^h_a(t)-\mu^h_a\right|>\varepsilon\right)\\ &\leqslant \sum_{t=t^h_\nu}^{T} \mathbb{P}_\nu\left(a^h_{t+1}=a, \left|\widehat{\mu}^{\mathcal{K}_t,h}_a(t)-\mu^h_a\right|>\varepsilon\right).\end{aligned} \tag{C.13}$$

Then, by combining Eq. C.13 and Lemma 41, we prove $\mathbb{E}_\nu\left[\left|\overline{\mathcal{E}}^h_{a,\varepsilon}\right|\right] \leqslant 4\varepsilon^{-2}$.

Subset $\mathcal{U}^h$:
By definition of the index (Eq. 6.6), we have

$$\forall t > |\mathcal{A}|, \quad N^h_{a^h_\star}(t)\mathrm{KL}\left(\widehat{\mu}^h_{a^h_\star}(t)\middle|U^h_{a^h_\star}(t)\right) = f(t). \tag{C.14}$$

Since $\widehat{\mu}^h_{a^h_\star}(t) \leqslant U^h_{a^h_\star}(t)$ for all $t>|\mathcal{A}|$, from the monotony of $\mathrm{KL}(x|\cdot)$ on $[x,+\infty)$, it comes

$$\forall t > |\mathcal{A}| \text{ such that } U^h_{a^h_\star}(t) \leqslant \mu^h_{a^h_\star}, \quad \mathrm{KL}\left(\widehat{\mu}^h_{a^h_\star}\middle|\mu^h_{a^h_\star}\right) \geqslant \mathrm{KL}\left(\widehat{\mu}^h_{a^h_\star}(t)\middle|U^h_{a^h_\star}(t)\right). \tag{C.15}$$

From Eq. C.14 and C.15 we deduce that

$$\mathcal{U}^h \subset \left\{t > |\mathcal{A}| :\ N^h_{a^h_\star}(t)\mathrm{KL}\left(\widehat{\mu}^h_{a^h_\star}(t)\middle|\mu_{a^h_\star,b^h_\star}\right) \geqslant f(t)\right\}. \tag{C.16}$$

From Eq. C.16 plus the union bound, it comes

$$\left|\mathcal{U}^h\right| \leqslant \sum_{t>|\mathcal{A}|}^{T} \mathbb{I}_{\left\{N^h_{a^h_\star}(t)\mathrm{KL}\left(\widehat{\mu}^h_{a^h_\star}(t)\middle|\mu_{a^h_\star,b^h_\star}\right)\geqslant f(t)\right\}}. \tag{C.17}$$

By taking the expectation on both sides in previous inequality (Eq. C.17), we have

$$\mathbb{E}_\nu\left[\left|\mathcal{U}^h\right|\right] \leqslant \sum_{t>|\mathcal{A}|}^{T} \mathbb{P}_\nu\left(N_{a_\star^h}^h(t)\mathrm{KL}\left(\widehat{\mu}_{a_\star^h}^h(t)\middle|\mu_{a_\star^h,b_\star^h}\right) \geqslant f(t)\right). \tag{C.18}$$

Combining Eq. C.18 and Lemma 42, it comes

$$\mathbb{E}_\nu\left[\left|\mathcal{U}^h\right|\right] \leqslant \sum_{t>|\mathcal{A}|} t^{-1}\log(t)^{-2}.$$

This implies $\mathbb{E}_\nu\left[\left|\mathcal{U}^h\right|\right] \leqslant 2$ since it can be shown that

$$\sum_{t>|\mathcal{A}|} t^{-1}\log(t)^{-2} \leqslant \int_{t\geqslant|\mathcal{A}|}^{\infty} t^{-1}\log(t)^{-2}\mathrm{d}t = \frac{1}{\log(|\mathcal{A}|)} \leqslant \frac{1}{\log(2)} \leqslant 2.$$

Subset $\overline{\mathcal{U}}^h$: From the definition of subset $\mathcal{T}^h$ (see Eq. 6.9), we have

$$\left\{t\geqslant t_\nu^h : t\notin\mathcal{T}^h\right\} \subset \left\{t\geqslant t_\nu^h :\ \overline{N}_{a_\star^h}^h(t) = N_{a_\star^h}^{\mathcal{K}_\star^h(t),h}(t),\ \overline{\mu}_{a_\star^h}^h(t) = \widehat{\mu}_{a_\star^h}^{\mathcal{K}_\star^h(t),h}(t),\ \overline{K}_t^h = \left|\mathcal{K}_\star^h(t)\right|\right\}, \tag{C.19}$$

where $\mathcal{K}_\star^h(t) \coloneqq \left\{k\in[\![1,h-1]\!] : b_\star^k = b_\star^h \text{ and } \overline{a}_\star^k = \overline{a}_t^h\right\} \subset \mathcal{K}^h$, $\mathcal{K}^h \coloneqq \left\{k\in[\![1,h-1]\!] : b_\star^k = b_\star^h\right\}$,
$N_a^{\mathcal{K},h}(t) = \sum_{k\in\mathcal{K}} N_a^k(T) + N_a^h(t)$, $S_a^{\mathcal{K},h}(t) = \sum_{k\in\mathcal{K}} S_a^k(T) + S_a^h(t)$ and $\widehat{\mu}_a^{\mathcal{K},h}(t) = S_a^{\mathcal{K},h}(t)/N_a^{\mathcal{K},h}(t)$ for all $\mathcal{K}\subset\mathcal{K}^h$ and $a\in\mathcal{A}$.
By definition of the index (Eq. 6.7), we have

$$\forall t\geqslant t_\nu^h,\quad \overline{N}_{a_\star^h}^h(t)\mathrm{KL}\left(\overline{\mu}_{a_\star^h}^h(t)\middle|\overline{U}_{a_\star^h}^h(t)\right) = f\left(\overline{K}_t^h T + t\right). \tag{C.20}$$

Since $\overline{\mu}_{a_\star^h}^h(t) \leqslant \overline{U}_{a_\star^h}^h(t)$ for all $t\geqslant t_\nu^h$, from the monotony of $\mathrm{KL}(x|\cdot)$ on $[x,+\infty)$, it comes

$$\forall t\geqslant t_\nu^h \text{ such that } \overline{U}_{a_\star^h}^h(t) \leqslant \mu_{a_\star^h}^h,\quad \mathrm{KL}\left(\overline{\mu}_{a_\star^h}^h(t)\middle|\mu_{a_\star^h}^h\right) \geqslant \mathrm{KL}\left(\overline{\mu}_{a_\star^h}^h(t)\middle|\overline{U}_{a_\star^h}^h(t)\right). \tag{C.21}$$

By defining $\mathcal{K}_t \coloneqq \mathcal{K}_\star^h(t)$ if $t\geqslant t_\nu^h$ and $t\notin\mathcal{T}^h$, $\emptyset$ otherwise, from Eq. C.19, C.20 and C.21 we deduce that

$$\overline{\mathcal{U}}^h \subset \left\{t\geqslant t_\nu^h :\ N_{a_\star^h}^{\mathcal{K}_t,h}(t)\mathrm{KL}\left(\widehat{\mu}_{a_\star^h}^{\mathcal{K}_t,h}(t)\middle|\mu_{a_\star^h,b_\star^h}\right) \geqslant f(|\mathcal{K}_t|\,T+t)\right\}. \tag{C.22}$$

Since we have

$$\begin{aligned}
&\left\{t\geqslant t_\nu^h :\ N_{a_\star^h}^{\mathcal{K}_t,h}(t)\mathrm{KL}\left(\widehat{\mu}_{a_\star^h}^{\mathcal{K}_t,h}(t)\middle|\mu_{a_\star^h,b_\star^h}\right) \geqslant f(|\mathcal{K}_t|\,T+t)\right\} \\
&= \bigcup_{K=0}^{h-1}\left\{t\geqslant t_\nu^h :\ |\mathcal{K}_t| = K,\ N_{a_\star^h}^{\mathcal{K}_t,h}(t)\mathrm{KL}\left(\widehat{\mu}_{a_\star^h}^{\mathcal{K}_t,h}(t)\middle|\mu_{a_\star^h,b_\star^h}\right) \geqslant f(KT+t)\right\}
\end{aligned}$$

by using the inclusion from Eq. C.22 plus the union bound, it comes

$$\left|\overline{\mathcal{U}}^h\right| \leqslant \sum_{K=0}^{h-1}\sum_{t=t_\nu^h}^{T} \mathbb{I}_{\left\{|\mathcal{K}_t|=K,\ N_{a_\star^h}^{\mathcal{K}_t,h}(t)\mathrm{KL}\left(\widehat{\mu}_{a_\star^h}^{\mathcal{K}_t,h}(t)\middle|\mu_{a_\star^h,b_\star^h}\right)\geqslant f(KT+t)\right\}}. \tag{C.23}$$

By taking the expectation on both sides in previous inequality (Eq. C.23), we have

$$\mathbb{E}_\nu\left[\left|\overline{\mathcal{U}}^h\right|\right] \leqslant \sum_{K=0}^{h-1}\sum_{t=t_\nu^h}^{T}\mathbb{P}_\nu\left(|\mathcal{K}_t|=K,\ N_{a_\star^h}^{\mathcal{K}_t,h}(t)\mathrm{KL}\left(\widehat{\mu}_{a_\star^h}^{\mathcal{K}_t,h}(t)\middle|\mu_{a_\star^h,b_\star^h}\right) \geqslant f(KT+t)\right). \tag{C.24}$$

Combining Eq. C.24 and Lemma 42, it comes

$$\begin{aligned}\mathbb{E}_\nu\Big[\big|\overline{\mathcal{U}}^h\big|\Big] &\leqslant \sum_{K=0}^{h-1}\sum_{t=t_\nu^h}^{T}(KT+t)^{-1}\log(KT+t)^{-2}\\ &\leqslant \sum_{t\geqslant t_\nu^h} t^{-1}\log(t)^{-2}\,.\end{aligned}$$

This implies $\mathbb{E}_\nu\Big[\big|\overline{\mathcal{U}}^h\big|\Big]\leqslant 2$ since it can be shown that

$$\sum_{t\geqslant t_\nu^h} t^{-1}\log(t)^{-2}\leqslant\int_{t=t_\nu^h-1}^{\infty} t^{-1}\log(t)^{-2}\mathrm{d}t=\frac{1}{\log(t_\nu^h-1)}\leqslant\frac{1}{\log(2)}\leqslant 2\,.$$

Subsets $\mathcal{C}_{a,\varepsilon}^h$, $\mathcal{C}_\varepsilon^h$, $\overline{\mathcal{C}}_{a,\varepsilon}^h$ and $\overline{\mathcal{C}}_\varepsilon^h$: We conclude the proof of Lemma 38 by taking the expectation on both sides in the following inequalities and by using the bounds on subsets $\mathcal{T}^h, \mathcal{U}^h, \overline{\mathcal{U}}^h, \mathcal{E}_{a,\varepsilon}^h$ and $\overline{\mathcal{E}}_{a,\varepsilon}^h$.

$$\begin{aligned}\big|\mathcal{C}_{a,\varepsilon}^h\big| &\leqslant \big|\mathcal{U}^h\big|+\big|\mathcal{E}_{a,\varepsilon}^h\big|\\ \big|\mathcal{C}_\varepsilon^h\big| &\leqslant \big|\mathcal{U}^h\big|+\sum_{a\neq a_\star^h}\big|\mathcal{E}_{a,\varepsilon}^h\big|\\ \Big|\overline{\mathcal{C}}_{a,\varepsilon}^h\Big| &\leqslant \big|\mathcal{T}^h\big|+\Big|\overline{\mathcal{U}}^h\Big|+\Big|\overline{\mathcal{E}}_{a,\varepsilon}^h\Big|\\ \Big|\overline{\mathcal{C}}_\varepsilon^h\Big| &\leqslant \big|\mathcal{T}^h\big|+\Big|\overline{\mathcal{U}}^h\Big|+\sum_{a\neq a_\star^h}\Big|\overline{\mathcal{E}}_{a,\varepsilon}^h\Big|\,.\end{aligned}$$

□

C.2.3 Proof of Lemma 20

From Lemma 38, we have the bound $\mathbb{E}_\nu\big[\big|\mathcal{C}_{a,\varepsilon}^h\big|\big]\leqslant 4\varepsilon^{-2}+2$ and $\mathbb{E}_\nu\Big[\Big|\overline{\mathcal{C}}_{a,\varepsilon}^h\Big|\Big]\leqslant 4\varepsilon^{-2}+3$.

Let us consider $t>|\mathcal{A}|$ such that $t\notin\mathcal{C}_{a,\varepsilon}^h=\mathcal{E}_{a,\varepsilon}^h\cup\mathcal{U}_a^h$ and $a_{t+1}^h=a$. By definition of the index (Eq. 6.6), we have

$$N_a^h(t)\mathrm{KL}\big(\widehat{\mu}_a^h(t)\big|U_a^h(t)\big)=f(t)\,. \tag{C.25}$$

Since $a_{t+1}^h=a$, it follows from the `KLUCB-RB` strategy that

$$u_{a_\star^h}^h(t)\leqslant u_a^h(t)\leqslant U_a^h(t)\,. \tag{C.26}$$

Since $a_{t+1}^h=a$, we have $t\notin\mathcal{U}^h$ and

$$\mu_\star^h\leqslant u_{a_\star^h}^h(t)=U_{a_\star^h}^h(t)\,. \tag{C.27}$$

Since $\varepsilon<\varepsilon_\nu$ and since a is a sub-optimal arm, we have

$$\mu_a^h+\varepsilon<\mu_\star^h\,. \tag{C.28}$$

Since $a_{t+1}^h=a$, we have $t\notin\mathcal{E}_{a,\varepsilon}^h$ and

$$\widehat{\mu}_a^h(t)\leqslant\mu_a^h+\varepsilon\,. \tag{C.29}$$

Then Eq. C.26, C.27, C.28 and C.29 imply

$$\widehat{\mu}_a^h(t)\leqslant\mu_a^h+\varepsilon<\mu_\star^h\leqslant U_a^h(t)\,. \tag{C.30}$$

Combining Eq. C.25 and Eq. C.30, it holds

$$\mathrm{KL}\big(\mu_a^h+\varepsilon\big|\mu_\star^h\big)\leqslant\mathrm{KL}\big(\widehat{\mu}_a^h(t)\big|U_a^h(t)\big)\quad\text{and}\quad N_a^h(t)\mathrm{KL}\big(\mu_a^h+\varepsilon\big|\mu_\star^h\big)\leqslant f(t)\,.$$

Let us consider $t\geqslant t_\nu^h$ such that $t\notin\overline{\mathcal{C}}_{a,\varepsilon}^h=\mathcal{T}\cup\overline{\mathcal{E}}_{a,\varepsilon}^h\cup\overline{\mathcal{U}}_a^h$ and $a_{t+1}^h=a$. By definition of the index (Eq. 6.7), we have

$$\overline{N}_a^h(t)\mathrm{KL}\Big(\overline{\mu}_a^h(t)\Big|\overline{U}_a^h(t)\Big)=f\Big(\overline{K}_t^hT+t\Big)\,. \tag{C.31}$$

Since $a_{t+1}^h=a$, it follows from the `KLUCB-RB` strategy that

$$u_{a_\star^h}^h(t)\leqslant u_a^h(t)\leqslant\overline{U}_a^h(t)\,. \tag{C.32}$$

Since $a_{t+1}^h=a$, we have $t\notin\mathcal{T}^h\cup\overline{\mathcal{U}}^h$ and

$$\mu_\star^h\leqslant u_{a_\star^h}^h(t)=\overline{U}_{a_\star^h}^h(t)\,. \tag{C.33}$$

Since $\varepsilon<\varepsilon_\nu$ and since a is a sub-optimal arm, we have

$$\mu_a^h+\varepsilon<\mu_\star^h\,. \tag{C.34}$$

Since $a_{t+1}^h=a$, we have $t\notin\mathcal{T}^h\cup\overline{\mathcal{E}}_{a,\varepsilon}^h$ and

$$\overline{\mu}_a^h(t)\leqslant\mu_a^h+\varepsilon\,. \tag{C.35}$$

Then Eq. C.32, C.33, C.34 and C.35 imply

$$\overline{\mu}_a^h(t)\leqslant\mu_a^h+\varepsilon<\mu_\star^h\leqslant\overline{U}_a^h(t)\,. \tag{C.36}$$

Combining Eq. C.31 and Eq. C.36, it holds

$$\mathrm{KL}\big(\mu_a^h+\varepsilon\big|\mu_\star^h\big)\leqslant\mathrm{KL}\Big(\overline{\mu}_a^h(t)\Big|\overline{U}_a^h(t)\Big)\quad\text{and}\quad\overline{N}_a^h(t)\mathrm{KL}\big(\mu_a^h+\varepsilon\big|\mu_\star^h\big)\leqslant f\Big(\overline{K}_t^hT+t\Big)\,.$$

In order to conclude the proof it remains to show that $\overline{K}_t^h\leqslant\beta_{b_\star^h}^{h-1}(h-1)$. Since $a_{t+1}^h=a$, we have $t\notin\mathcal{T}^h$ and we deduce from the definition of $\mathcal{T}^h$ (see Eq. 6.9) that

$$\overline{K}_t^h=\big|\mathcal{K}_\star^h(t)\big|\leqslant\big|\big\{k\in[\![1,h-1]\!]:\ b_\star^k=b_\star^h\big\}\big|=\ \beta_{b_\star^h}^{h-1}(h-1)\,,$$

where $\mathcal{K}_\star^h(t)\coloneqq\big\{k\in[\![1,h-1]\!]\colon b_\star^k=b_\star^h\text{ and }\overline{a}_\star^k=\overline{a}_t^h\big\}$.

Finally, we prove the last statement of Lemma 20. For all sub-optimal arm $a\in\mathcal{A}$, for all period $h\geqslant1$, for all time step $t>|\mathcal{A}|$, we denote by

$$\tau_a^h(t)=\max\big\{t'\in[\![|\mathcal{A}|+1\,;t]\!]:\ a_{t'+1}^h=a\quad\text{and}\quad t'\notin\mathcal{C}_{a,\varepsilon}^h\big\} \tag{C.37}$$

the last time step before time step t that does not belong to $\mathcal{C}_{a,\varepsilon}^h$ such that we pull arm a in period h. In particular, we have

$$N_a^h\big(\tau_a^h(t)\big)\leqslant\frac{f\big(\tau_a^h(t)\big)}{\mathrm{KL}(\mu_a^h+\varepsilon|\mu_\star^h)}\leqslant\frac{f(t)}{\mathrm{KL}(\mu_a^h+\varepsilon|\mu_\star^h)}\,. \tag{C.38}$$

Then, from Eq. C.37 and Eq. C.38 we have

$$\begin{aligned} N_a^h(t) &= N_a^h(|\mathcal{A}|+1) + \sum_{t'>|\mathcal{A}|}^{t-1} \mathbb{I}_{\{a_{t'+1}^h=a\}} \\ &= N_a^h(|\mathcal{A}|+1) + \sum_{t'>|\mathcal{A}|}^{t-1} \mathbb{I}_{\{a_{t'+1}^h=a,\, t'\in\mathcal{C}_{a,\varepsilon}^h\}} + \sum_{t'>|\mathcal{A}|}^{t-1} \mathbb{I}_{\{a_{t'+1}^h=a,\, t'\notin\mathcal{C}_{a,\varepsilon}^h\}} \\ &\leqslant N_a^h(|\mathcal{A}|+1) + \left|\mathcal{C}_{a,\varepsilon}^h\right| + N_a^h\left(\tau_a^h(t)\right) \\ &\leqslant N_a^h(|\mathcal{A}|+1) + \left|\mathcal{C}_{a,\varepsilon}^h\right| + \frac{f(t)}{\mathrm{KL}(\mu_a^h+\varepsilon|\mu_\star^h)}\,. \end{aligned}$$

C.2.4 Proof of Lemma 21

Let us consider $t \geqslant t_\nu^h$ such that $t \notin \mathcal{T}^h$ and $\overline{a}_t^h \neq a_\star^h$. Since $t \notin \mathcal{T}^h$ (see Eq. 6.9),

$$\forall a \in \mathcal{A}, \quad \overline{N}_a^h(t) = N_a^h(t) + \sum_{k\in\mathcal{K}_\star^h(t)} N_a^k(T)\,, \tag{C.39}$$

where $\mathcal{K}_\star^h(t) \coloneqq \left\{k \in [\![1, h-1]\!] : \overline{a}_\star^k = \overline{a}_t^h\right\}$. Since for all $k \in \mathcal{K}_\star^h$, $\overline{a}_\star^k \in \operatorname{argmax}_{a\in\mathcal{A}} N_a^k(T)$, from Eq. C.39 we deduce that $\overline{a}_t^h \in \operatorname{argmax}_{a\in\mathcal{A}} \overline{N}_a^h(t)$. Since $\overline{a}_t^h \neq a_\star^h$, this implies

$$\overline{N}_{a_\star^h}^h(t) \leqslant \overline{N}_{\overline{a}_t^h}^h(t) \quad \text{and} \quad \overline{N}_{\overline{a}_t^h}^h(t) \leqslant \sum_{a\neq a_\star^h} \overline{N}_a^h(t)\,. \tag{C.40}$$

Furthermore, since $t \notin \mathcal{T}^h$ (see Eq. 6.9), we have

$$\overline{K}_t^h \coloneqq \left|\mathcal{K}_+^h(t)\right| = \left|\mathcal{K}_\star^h(t)\right|\,. \tag{C.41}$$

Then it comes

$$\overline{N}_{a_\star^h}^h(t) = \left|\mathcal{K}_\star^h(t)\right| T + t - \sum_{a\neq a_\star^h} \overline{N}_a^h(t)\,. \tag{C.42}$$

Then Eq. C.39, C.40 and C.42 imply

$$\frac{\left|\mathcal{K}_\star^h(t)\right| T}{2} + \frac{t}{2} \leqslant \sum_{a\neq a_\star^h} N_a^h(t) + \sum_{k\in\mathcal{K}_\star^h(t)} N_a^k(T)\,. \tag{C.43}$$

For $a \neq a_\star^h$ and for $k \in \mathcal{K}_\star^h(t)$, the arm a is sub-optimal for bandit $b_\star^k = b_\star^h$. Thus, from Lemma 20, we have

$$\begin{aligned} \forall a \neq a_\star^h, \forall k \in \mathcal{K}_\star^h(t), \qquad N_a^h(t) &\leqslant \frac{f(t)}{\mathrm{KL}(\mu_a^h+\varepsilon|\mu_\star^h)} + \left|\mathcal{C}_{a,\varepsilon}^h\right| + N_a^h(|\mathcal{A}|+1) \\ N_a^k(T) &\leqslant \frac{f(T)}{\mathrm{KL}(\mu_a^k+\varepsilon|\mu_\star^k)} + \left|\mathcal{C}_{a,\varepsilon}^k\right| + N_a^k(|\mathcal{A}|+1)\,. \end{aligned} \tag{C.44}$$

Then, by combining Eq. C.43 and Eq. C.44, we get

$$\begin{aligned} &\frac{t}{2} - \left(\sum_{a\neq a_\star^h} \frac{f(t)}{\mathrm{KL}(\mu_a^h+\varepsilon|\mu_\star^h)} + N_a^h(|\mathcal{A}|+1)\right) \\ + \;& \frac{\left|\mathcal{K}_\star^h(t)\right| T}{2} - \left(\sum_{k\in\mathcal{K}_\star^h(t)} \sum_{a\neq a_\star^h} \frac{f(T)}{\mathrm{KL}(\mu_a^k+\varepsilon|\mu_\star^k)} + N_a^k(|\mathcal{A}|+1)\right) \\ \leqslant \;& \sum_{k\in\mathcal{K}_\star^h(t)\cup\{h\}} \sum_{a\neq a_\star^h} \left|\mathcal{C}_{a,\varepsilon}^k\right| \end{aligned} \tag{C.45}$$

We finally prove Lemma 21 from Eq. C.45 and the following inequalities

$$\begin{aligned}
&\forall k \in \mathcal{K}_\star^h(t) \cup \{h\}\,, \quad a_\star^k = a_\star^h\,,\\
&\forall k \in \mathcal{K}_\star^h(t) \cup \{h\}\,, \quad \sum_{a\neq a_\star^h} N_a^k(|\mathcal{A}|+1) = \sum_{a\neq a_\star^k} N_a^k(|\mathcal{A}|+1) \leqslant |\mathcal{A}|\,,\\
&\forall k \in \mathcal{K}_\star^h(t) \cup \{h\}\,, \quad \sum_{a\neq a_\star^h} \left|\mathcal{C}_{a,\varepsilon}^k\right| = \sum_{a\neq a_\star^k} \left|\mathcal{C}_{a,\varepsilon}^k\right| = \left|\mathcal{C}_\varepsilon^k\right|\,.
\end{aligned}$$

C.2.5 Proof of Proposition 6

We first deduce Lemma 39 from Lemma 21.

Lemma 39 (Conditions for misidentifying the best arms)**.** *For all period $h \geqslant 1$, for all $0 < \varepsilon < \varepsilon_\nu$, for all $t \geqslant T_{\nu,\varepsilon}^h$,*

$$\left(t \notin \mathcal{T}^h \text{ and } \overline{a}_t^h \neq a_\star^h \right) \iff \left(t < 4\left|\mathcal{C}_\varepsilon^h\right| \text{ or } \exists k \in [\![1,h]\!],\, T < 8\left|\mathcal{C}_\varepsilon^k\right| \right).$$

This implies

$$\left(T \notin \mathcal{T}^h \text{ and } \overline{a}_\star^h \neq a_\star^h \right) \iff \exists k \in [\![1,h]\!],\, T < 8\left|\mathcal{C}_\varepsilon^k\right|\,.$$

We respectively refer to Proposition 6, Eq. 6.9 and Eq. 6.11 for the definitions of $T_{\nu,\varepsilon}^h$, $\mathcal{T}^h$ and $\mathcal{C}_\varepsilon^h$.

The proof of Lemma 39 is deferred to the Section C.2.5. We prove Proposition 6 in the following.

Let us introduce the subset $\mathcal{P}$ of pairs period-time when there is false positives or false negatives, or when the mean of the current pulled arm is underestimated, or when the index of the best arm is below its mean, or when the most pulled arms are different from the best arms. More formally,

$$\mathcal{P} := \left\{ (h,t) \in [\![1,h]\!] \times [\![1,T]\!] :\ \left(t \geqslant T_{\nu,\varepsilon}^h, t \in \overline{\mathcal{C}}_\varepsilon^h \cup \mathcal{M}_\varepsilon^h \right) \text{ or } \left(\exists k \in [\![1,h-1]\!], T \in \mathcal{T}^k \cup \mathcal{M}_\varepsilon^k \right) \right\}, \tag{C.46}$$

where $\mathcal{M}_\varepsilon^h := \left\{ t \geqslant T_{\nu,\varepsilon}^h : t \notin \mathcal{T}^h \text{ and } \overline{a}_t^h \neq a_\star^h \right\}$, for all period $h \geqslant 1$.

Then, for a bandit $b \in \mathcal{B}$ and a sub-optimal arm $a \in \mathcal{A}$, from Lemma 20 we have

$$N_{a,b}(H,T) \leqslant \frac{f\left(\beta_b^H HT\right)}{\mathrm{KL}(\mu_{a,b} + \varepsilon | \mu_b^\star)} + \sum_{h=1}^{H}\sum_{t=0}^{T-1} \mathbb{I}_{\left\{ b_\star^h = b,\ a_{t+1}^h = a,\ t < T_{\nu,\varepsilon}^h \text{ or } (h,t) \in \mathcal{P} \right\}}\,. \tag{C.47}$$

From the definitions of $\mathcal{P}$ (Eq. C.46), $T_{\nu,\varepsilon}^h$ (Proposition 6) and $\overline{\mathcal{C}}_\varepsilon^h$ (Eq. 6.11) for $h \geqslant 1$, this implies

$$\begin{aligned}
N_{a,b}(H,T) \ \leqslant\ & \frac{f\left(\beta_b^H HT\right)}{\mathrm{KL}(\mu_{a,b} + \varepsilon | \mu_b^\star)} \\
&+ \sum_{h=1}^{H}\sum_{t=0}^{T-1} \mathbb{I}_{\left\{ b_\star^h = b,\ a_{t+1}^h = a,\ t < T_{\nu,\varepsilon}^h \right\}} \\
&+ \sum_{h=1}^{H}\sum_{t=0}^{T-1} \mathbb{I}_{\left\{ b_\star^h = b,\ a_{t+1}^h = a,\ t \in \overline{\mathcal{C}}_\varepsilon^h \right\}} \\
&+ \sum_{h=1}^{H}\sum_{t=0}^{T-1} \mathbb{I}_{\left\{ b_\star^h = b,\ a_{t+1}^h = a,\ t \notin \overline{\mathcal{C}}_{a,\varepsilon}^h \right\}} \mathbb{I}_{\left\{ \exists k \in [\![1,h-1]\!], T \in \mathcal{T}^k \right\}} \\
&+ \sum_{h=1}^{H}\sum_{t=0}^{T-1} \mathbb{I}_{\left\{ b_\star^h = b,\ a_{t+1}^h = a,\ t \notin \overline{\mathcal{C}}_{a,\varepsilon}^h \right\}} \mathbb{I}_{\left\{ t \in \mathcal{M}_\varepsilon^h \text{ or } \exists k \in [\![1,h-1]\!], T \in \mathcal{M}_\varepsilon^k \right\}}\,.
\end{aligned} \tag{C.48}$$

Furthermore, from Lemma 39 we have for all period $h \geqslant 1$,

$$\mathbb{I}_{\left\{t\in\mathcal{M}_\varepsilon^h \text{ or } \exists k\in[\![1,h-1]\!],T\in\mathcal{M}_\varepsilon^k\right\}} \leqslant \mathbb{I}_{\left\{t<4\left|\mathcal{C}_\varepsilon^h\right|\right\}} + \sum_{k=1}^{h} \mathbb{I}_{\left\{T<8\left|\mathcal{C}_\varepsilon^k\right|\right\}} \,. \tag{C.49}$$

By combining Eq.C.48 and Eq.C.49, we get

$$\begin{aligned} N_{a,b}(H,T) \;\leqslant\; & \frac{f\left(\beta_b^H HT\right)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\mu_b^\star)} \\ & + \sum_{h=1}^{H} \mathbb{I}_{\left\{b_\star^h=b\right\}} \left[T_{\nu,\varepsilon}^h + 4\left|\mathcal{C}_\varepsilon^h\right| + \left|\overline{\mathcal{C}}_\varepsilon^h\right| + \left(\sum_{t=0}^{T-1} \mathbb{I}_{\left\{b_\star^h=b,\; a_{t+1}^h=a,\; t\notin\overline{\mathcal{C}}_{a,\varepsilon}^h\right\}} \right) \left(\sum_{k=1}^{h} \mathbb{I}_{\left\{T<8\left|\mathcal{C}_\varepsilon^k\right|\right\}} + \mathbb{I}_{\left\{T\in\mathcal{T}^k\right\}} \right) \right]. \end{aligned} \tag{C.50}$$

Since the arm a is sub-optimal for the bandit b, the consistency (Lemma 20) implies

$$\forall h \geqslant 1, \quad \sum_{t=0}^{T-1} \mathbb{I}_{\left\{b_\star^h=b,\; a_{t+1}^h=a,\; t\notin\overline{\mathcal{C}}_{a,\varepsilon}^h\right\}} \leqslant \frac{f(hT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\mu_b^\star)} \,. \tag{C.51}$$

In addition, the following Markov's type inequalities are satisfied

$$\forall k \geqslant 1, \quad \mathbb{I}_{\left\{T<8\left|\mathcal{C}_\varepsilon^k\right|\right\}} \leqslant \frac{8\left|\mathcal{C}_\varepsilon^k\right|}{T} \,. \tag{C.52}$$

By combining Eq. C.50, C.51 and C.52, we prove Proposition 6, that is

$$\begin{aligned} & N_{a,b}(H,T) \\ \leqslant\; & \frac{f(\beta_b^H HT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\mu_b^\star)} + \sum_{h=1}^{H} \mathbb{I}_{\left\{b_\star^h=b\right\}} \left[T_{\nu,\varepsilon}^h + 4\left|\mathcal{C}_\varepsilon^h\right| + \left|\overline{\mathcal{C}}_\varepsilon^h\right| + \frac{f(hT)}{\mathrm{KL}(\mu_{a,b}+\varepsilon|\mu_b^\star)} \sum_{k=1}^{h} \frac{8\left|\mathcal{C}_\varepsilon^k\right|}{T} + \mathbb{I}_{\left\{T\in\mathcal{T}^k\right\}} \right]. \end{aligned}$$

Proof of Lemma 39

Let us consider a period $h \geqslant 1$, $0 < \varepsilon < \varepsilon_\nu$, and a time step all $t \geqslant T_{\nu,\varepsilon}^h$ such that $t \notin \mathcal{T}^h$ and $\overline{a}_t^h \neq a_\star^h$. Then, since $T_{\nu,\varepsilon}^h \geqslant t_\nu^h$, from Lemma 21 we have

$$\frac{t+\left|\mathcal{K}_\star^h(t)\right|T}{2} - \left(1+\left|\mathcal{K}_\star^h(t)\right|\right)|\mathcal{A}| - \left(f(t)+\left|\mathcal{K}_\star^h(t)\right|f(T)\right) \sum_{a\neq a_\star^h} \frac{1}{\mathrm{KL}(\mu_a^h+\varepsilon|\mu_\star^h)} \leqslant \sum_{k\in\mathcal{K}_\star^h(t)\cup\{h\}} \left|\mathcal{C}_\varepsilon^k\right| . \tag{C.53}$$

Furthermore, by definition of $T_{\nu,\varepsilon}^h$, since $t \geqslant T_{\nu,\varepsilon}^h$, we have

$$\begin{aligned} \frac{t}{2} - \textstyle\sum_{a\neq a_\star^h} \dfrac{f(t)}{\mathrm{KL}(\mu_a^h+\varepsilon|\mu_\star^h)} - |\mathcal{A}| \;&>\; \frac{t}{4} \\ \frac{T}{2} - \textstyle\sum_{a\neq a_\star^h} \dfrac{f(T)}{\mathrm{KL}(\mu_a^h+\varepsilon|\mu_\star^h)} - |\mathcal{A}| \;&>\; \frac{T}{4} \,. \end{aligned} \tag{C.54}$$

By respectively combining Eq. C.53 and Eq. C.54, we thus deduce

$$\begin{aligned} \left|\mathcal{K}_\star^h(t)\right| = 0 \quad &\Rightarrow \quad t \leqslant 4\left|\mathcal{C}_\varepsilon^h\right| \\ \left|\mathcal{K}_\star^h(t)\right| \geqslant 1 \quad &\Rightarrow \quad \exists k \in [\![1,h]\!], T \leqslant 8\left|\mathcal{C}_\varepsilon^k\right| \end{aligned}$$

which implies Lemma 39.

C.2.6 Tools from Concentration of Measure

This subsection gathers useful concentration lemmas that do not depend on the considered strategy.

Notations For all period $h \geqslant 2$, for all time step $t > |\mathcal{A}|$, for each (possible random) subset of past periods $\mathcal{K} \subset \mathcal{K}^h := \big\{k \in [\![1, h-1]\!] : b^k_\star = b^h_\star\big\}$, for all arm $a \in \mathcal{A}$, we define $N_a^{\mathcal{K},h}(t) := \sum_{k\in\mathcal{K}} N_a^k(T) + N_a^h(t)$, $S_a^{\mathcal{K},h}(t) := \sum_{k\in\mathcal{K}} S_a^k(T) + S_a^h(t)$ and $\widehat{\mu}_a^{\mathcal{K},h}(t) := S_a^{\mathcal{K},h}(t)/N_a^{\mathcal{K},h}(t)$.
In particular, for $\mathcal{K} = \mathcal{K}^h_\star(t) := \big\{k \in [\![1, h-1]\!] : b^k_\star = b^h_\star \text{ and } \overline{a}^k_\star = \overline{a}^h_t\big\}$, we have $N_a^{\mathcal{K}^h_\star(t),h}(t) = \overline{N}^h_a(t)$ and $\widehat{\mu}_a^{\mathcal{K}^h_\star(t),h}(t) = \overline{\mu}^h_a(t)$ when $t \geqslant t^h_\nu$ and $t \notin \mathcal{T}^h$ (see Eq. 6.8 and 6.9).

Uniform bounds based on the Laplace method (method of mixtures for sub-Gaussian random variables, see Peña et al. (2008)) are given in Lemma 40.

Lemma 40 (Uniform sub-Gaussian concentration)**.** *For all period $h \geqslant 2$, for all time step $t > |\mathcal{A}|$, for all arm $a \in \mathcal{A}$, for all $\delta \in (0,1)$, it holds*

$$\mathbb{P}_\nu\big(\widehat{\mu}^h_a(t) - \mu_{a,b^h_\star} \geqslant \mathrm{d}\big(N^h_a(t), \delta\big)\big) \leqslant \delta$$
$$\mathbb{P}_\nu\big(\mu_{a,b^h_\star} - \widehat{\mu}^h_a(t) \geqslant \mathrm{d}\big(N^h_a(t), \delta\big)\big) \leqslant \delta\,,$$

where $\mathrm{d}(n,\delta) = \sqrt{2(1+1/n)\log\big(\sqrt{n+1}/\delta\big)/n}$*, for all* $n \geqslant 1$.

Lemma 41 reformulates Lemma B.1 from Combes and Proutiere (2014b).

Lemma 41 (Concentration inequalities)**.** *For all period $h \geqslant 2$, for all arm $a \in \mathcal{A}$, for all $\varepsilon \in (0, 1/2)$, and all possibly random subset of periods $\mathcal{K}_t$ such that the random variable $N_a^{\mathcal{K}_t,h}(t)$ is a random stopping time, it holds*

$$\sum_{t\geqslant 1} \mathbb{P}_\nu\big(a^h_{t+1} = a,\ \big|\widehat{\mu}_a^{\mathcal{K}_t,h}(t) - \mu^h_a\big| \geqslant \varepsilon\big) \leqslant 4\varepsilon^{-2}\,.$$

Lemma 42 reformulates Theorem 1 from Garivier (2013).

Lemma 42 (Self-normalized inequalities)**.** *For all period $h \geqslant 2$, for all time step $t > |\mathcal{A}|$, for all arm $a \in \mathcal{A}$, for all $K \in [\![0, h-1]\!]$, for all $\delta > 0$ and all possibly random subset of periods $\mathcal{K}_t$ such that the random variable $N_a^{\mathcal{K}_t,h}(t)$ is a random stopping time, it holds*

$$\mathbb{P}_\nu\big(|\mathcal{K}_t| = K,\ N_a^{\mathcal{K}_t,h}(t)\mathrm{KL}\big(\widehat{\mu}_a^{\mathcal{K}_t,h}(t)\big|\mu_{a,b^h_\star}\big) \geqslant \delta\big) \leqslant 2e\lceil \delta\log(KT+t)\rceil \exp(-\delta)\,.$$

In particular, this implies for $\delta = f(KT+t)$,

$$\mathbb{P}_\nu\big(|\mathcal{K}_t| = K,\ N_a^{\mathcal{K}_t,h}(t)\mathrm{KL}\big(\widehat{\mu}_a^{\mathcal{K}_t,h}(t)\big|\mu_{a,b^h_\star}\big) \geqslant f(KT+t)\big) \leqslant (KT+t)^{-1}\log(KT+t)^{-2}\,.$$

C.3 Additional Experiments: Ideal Cases for which Bandits are Close Enough on the Subset of Optimal Arms

This section provides additional experiments where we investigate some favorable distributions ν where it is hard to separate the different bandits from each other. All experiments are repeated 100 times.

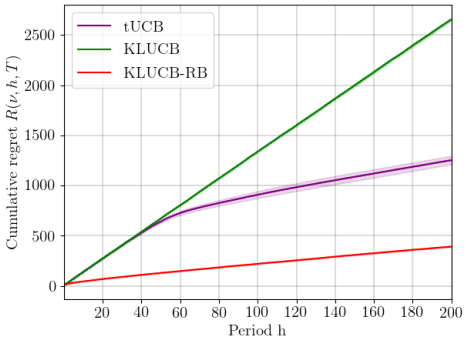


Figure C.1: Cumulative regret of KLUCB, KLUCB−RB and tUCB along $H = 200$ periods of $T = 10^3$ rounds for the bandit set $\mathcal{B} = \{b\}$.

C.3.1 A Single Instance

Let us first make a remark in the trivial limit case of a unique bandit, that is to say $\mathcal{B} = \{b\}$. In such cases, playing KLUCB−RB is obviously equivalent to playing for $T_{total} := HT$ rounds a KLUCB strategy on the bandit instance b, with an additional term $(H-1)\sum_{a\in\mathcal{A}} \Delta_{a,b}$ in the final cumulative regret due to the initialization at each period. Figure C.1 highlights this fact for the two-armed bandit b defined in Eq. C.55, over $H = 200$ periods of $T = 10^3$ rounds.

$$b : (\mu_{1,b}, \mu_{2,b}) = (-\frac{\Delta}{2}, \frac{\Delta}{2}) \qquad \text{where} \qquad \Delta = 10\sqrt{\frac{\log(HT)}{T}}. \tag{C.55}$$

Although the case $|\mathcal{B}| = 1$ is not an interesting one since there is no switches between different bandits instances, it enables to understand what happens when $|\mathcal{B}| > 1$ and bandits are similar, that is $\max_{a\in\mathcal{A}^\star} \max_{b\neq b'} |\mu_{a,b} - \mu_{a,b'}|$ approaches 0. Besides, it highlights the need for tUCB to see a sufficient number of periods before exploiting the estimated models of the bandits.

C.3.2 Similarity of Different Instances on the Optimal Subset $\mathcal{A}^\star$

Let us consider routines over two bandits b_1 and b_2 composed of two arms such that $(\mu_{1,b_2}, \mu_{2,b_2}) = (\mu_{1,b_1} + \gamma, \mu_{2,b_1} - \gamma)$ and $a^\star_{b_1} = 2$. If $\gamma > \Delta_{2,b_1}/2$, arms arrangements are different in both instances and these cases are studied in subsection 6.4.1. Otherwise we have $a^\star_{b_1} = a^\star_{b_2} = 2$ if ever $0 < \gamma < \Delta_{2,b_1}/2$. Although separation of instances is particularly hard in such cases, samples aggregation from false positive periods does not perturb the empirical means arrangement, and thus yields great performances for KLUCB−RB. To explain how this kind of distribution generalizes to settings composed of arbitrary numbers of bandits and arms, we present in Fig. C.2b a distribution ν such that $|\mathcal{B}| = 5$ and $|\mathcal{A}| = 4$. In this setting, we have $\mathcal{A}^\star = \{1, 4\}$. Considering distributions of bandits restricted to $\mathcal{A}^\star$, $\mathcal{B}$ naturally decomposes into 3 clusters $\mathcal{C}^{(1)} := \{b_1, b_4\}$, $\mathcal{C}^{(2)} := \{b_2, b_3\}$ and $\mathcal{C}^{(3)} := \{b_5\}$ so that

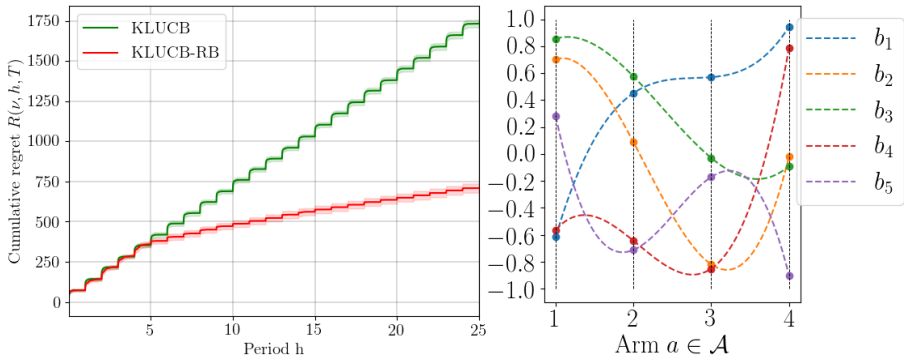
$$\forall i \in \{1,2,3\}\,, \forall b, b' \in \mathcal{C}^{(i)}, \forall a \in \mathcal{A}^\star, |\mu_{a,b} - \mu_{a,b'}| < \frac{1}{2} \min_{y\in\mathcal{C}^{(i)}} \min_{x\in\mathcal{A}, x\neq a^\star_y} \Delta_{x,y} \tag{C.56}$$

which entails in particular

$$\forall i \in \{1,2,3\}\,, \exists a^{(i)} \in \mathcal{A}^\star, \forall b \in \mathcal{C}^{(i)}, a^\star_b = a^{(i)}\ ,$$

and

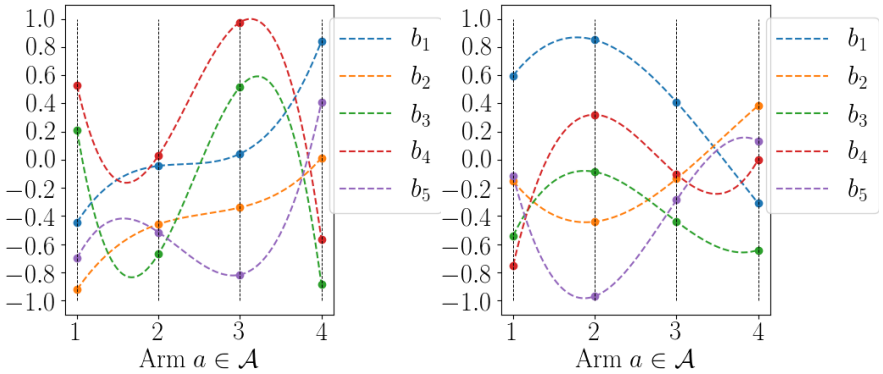
$$\forall i \in \{1,2,3\}, \forall b \in \mathcal{C}^{(i)}, \forall b' \notin \mathcal{C}^{(i)}, |\mu_{a^{(i)},b} - \mu_{a^{(i)},b'}| > \min_{x\in\mathcal{C}^{(i)}} \min_{a\neq a^\star_x} \Delta_{a,x}. \tag{C.57}$$



(a) Cumulative regret rounds.

(b) $\nu : \{\nu_{b_1}, \ldots, \nu_{b_5}\}$

Figure C.2: `KLUCB-RB` and `KLUCB` performances on a clusterized distribution according to $\mathcal{A}^\star$, along $H = 25$ periods of 2×10^4



(a) $\nu^{(1)} : \{\nu^{(1)}_{b_1}, \ldots, \nu^{(1)}_{b_5}\}$

(b) $\nu^{(2)} : \{\nu^{(2)}_{b_1}, \ldots, \nu^{(2)}_{b_5}\}$

Figure C.3: Distribution ν for each bandit in sets $\mathcal{B}_1$ and $\mathcal{B}_2$.

On the one hand, Eq. C.56 sums up that different bandits from a same cluster are hard to distinguish, in comparison with the difficulty of learning each instance independently. On the other hand, Eq. C.57 implies that clusters are easy to separate from each other. Besides Eq. C.56 also implies that the permutation of $\mathcal{A}$ sorting arms according to an increasing order is the same for all instances from a same cluster. Thus `KLUCB-RB` is expected to perform well for this kind of arms distributions. Fig. C.2a shows the cumulative regret curves with one standard deviation obtained on this setting, along $H = 25$ periods of $T = 2 \times 10^4$ rounds. As expected, it highlights that a positive cluster effect causes an improvement in regret minimization. In practice, `KLUCB-RB` naturally clusterizes the previously seen periods while the current period index h increases. More specifically, noting $\mathcal{C}(h)$ the cluster containing $b^h_\star$, `KLUCB-RB` makes all the different bandits from $\mathcal{C}(h)$ share their samples with $b^h_\star$ for a large amount of rounds, which enables to boost the minimization of regret across period h.

C.3.3 Complement of Sections 6.4.2 and 6.4.3

Figure C.3 shows the generated settings used in experiments of Section 6.4.2, and Figure C.4 the setting used in Section 6.4.3. More specifically, each sub-figure displays the expected reward for each of the four arms, in each of the bandits, for the three considered bandit sets.

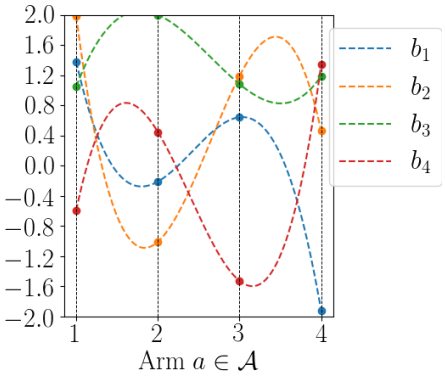


Figure C.4: Distribution ν for each bandit in the critical setting.

Appendix D

Bandits with Groups of Similar Arms

This work is the subject of Pesquerel et al. (2021) and has been initiated and mainly driven by Fabien Pesquerel. The proofs are not reproduced in this chapter. We refer to Pesquerel et al. (2021) and its Appendix for more details.

D.1 Introduction

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

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

Assuming the set of arms $\mathcal{A}$ and base distributions $\mathcal{D}$ is known to the learner, we denote by $\mathcal{D}_{\text{sim}}(q)$ the set of bandit configurations having the q-equivalence property.

Definition 7 (Arm equivalence and equivalence class)**.** *Given a bandit configuration ν, two arms $a, a' \in \mathcal{A}$ are said to be equivalent if their associated distributions have the same expected values:*

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

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

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

Goal. For the structure $\mathcal{D}_{\text{sim}}(q)$, as we show in Theorem 7 below, the term $\mathfrak{C}_{\mathcal{D}_{\text{sim}}(q)}(\mu)$ unfortunately makes appear in general a combinatorial optimization problem. This makes resorting to OSSB or any strategy targeting exact asymptotic optimality a daunting task for the practitioner. In this chapter, our goal is to provide a computationally efficient strategy adapted to the structure $\mathcal{D}_{\text{sim}}(q)$, that is able to reach optimality up to controlled error term.

Outline and contributions The rest of this chapter is organized as follows. In section D.2, we derive a lower bound on the regret for the structured set of bandit configurations $\mathcal{D}_{\text{sim}}(q)$. This bound makes appear two components, one that we call *non-combinatorial* as optimizing it can be done efficiently, and a second term that we term *combinatorial* as it involves solving a combinatorial problem. Interestingly, using in Lemma 43 and Theorem 9 that the contribution of the combinatorial part of the lower bound can be controlled. Owing to this key insight, we introduce in section D.3, `IMED-EC`, an adaptation of the `IMED` strategy from Honda and Takemura (2015) to the structured set $\mathcal{D}_{\text{sim}}(q)$. One advantage of `IMED` over a `KLUCB` alternative is its reduced complexity, which translates to the equivalence class setup. At each time step, the complexity of computing the next arm to be pulled by `IMED-EC` is no more than the one of sorting a list of $|\mathcal{A}|$ elements once the `IMED` indexes have been computed is at most the one of sorting a list of $|\mathcal{A}|$ elements, which is only $\log|\mathcal{A}|$ times larger than looking for the minimal `IMED` index. In Section D.4, we prove that `IMED-EC` achieves a controlled asymptotic regret that matches the non-combinatorial part of the lower bound and is at most (less than) 2 times from the optimal regret bound. Last, we illustrate the benefit of the `IMED-EC` over its unstructured version in section D.5, where it shows a substantial improvement. Our experiments also highlights the robustness of the algorithm to a misspecified parameter q, which is a desirable feature for the practicioner.

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

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

Definition 8 (Confusing instance)**.** *Given a bandit configuration $\nu \in \mathcal{D}_{sim}(q)$, a real number λ and a subset $c_q \subseteq \mathcal{A}$ of q equivalent arms in ν, we denote by $\mathcal{D}_{sim}(q,\nu,c_q,\lambda)$ the set of all bandit configurations having the same set of arms as ν and such that for all $\nu' \in \mathcal{D}_{sim}(q,\nu,c_q,\lambda)$, $\nu' \in \mathcal{D}_{sim}(q)$ and for every arm a in c_q, $\mu'_a \geqslant \lambda$. When $\lambda > \mu^\star$, and c_q is a subset of a suboptimal class, a bandit configuration in $\mathcal{D}_{sim}(q,\nu,c_q,\lambda)$ is called a **confusing instance** of ν. Similarly to the notation introduced above, we will use the notation $\mathcal{D}_{sim}(q,\mu,c_q,\lambda)$ to specify the set of means of bandit configurations in $\mathcal{D}_{sim}(q,\nu,c_q,\lambda)$.*

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

Theorem 7 (Asymptotic lower bound)**.** *Let $q \in \mathbb{N} - \{0\}$ be a positive integer and $\nu \in \mathcal{D}_{sim}(q)$ be a bandit configuration having the q-equivalence property. Let $c \subset \mathcal{A}$ be a suboptimal equivalence class in ν. Assuming uniform consistency, for all suboptimal arm a,*

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

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

$$\liminf_{T\to\infty} \frac{\min\limits_{c_q\subseteq c} \sum\limits_{a\in c_q} \mathbb{E}_\nu[N_a(T)]\,\mathrm{KL}(\mu_a|\mu^\star) + \inf\limits_{\mu'\in\mathcal{D}_{sim}(q,\mu,c_q,\lambda)} \sum\limits_{a\notin c_q} \mathbb{E}_\nu[N_a(T)]\,\mathrm{KL}(\mu_a|\mu'_a)}{\log T} \geqslant 1 \tag{D.1}$$

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

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

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

$$\liminf_{T\to\infty} \frac{\min\limits_{c_q \subseteq c} \sum\limits_{a\in c_q} \mathbb{E}_\nu[N_a(T)]\,\mathrm{KL}(\mu_a|\mu^\star)}{\log T} \geqslant 1.$$

We call this the *non-combinatorial* regime because the minimum over all q-partitions of c is in fact the sum of the q smallest elements of $\{\mathbb{E}_\nu[N_a(T)]\,\mathrm{KL}(\mu_a|\lambda)\}_{a\in c}$. The search amongst all the q-partitions of c amount to a research of the q smallest elements which is not more complex than sorting a list of $|c|$ elements.

Lemma 43. *Let $\nu \in \mathcal{D}_{sim}(q)$ be a bandit configuration having the q-equivalence property. Let c be a suboptimal class in the non-combinatorial regime, then, under assumtion 1 and 2,*

$$\liminf_{T\to\infty} \frac{\sum\limits_{a\in c} \mathbb{E}_\nu[N_a(T)]}{\log T} \geqslant \frac{|c|}{q} \frac{1}{\mathrm{KL}(\mu_a|\lambda)} \tag{D.2}$$

While we do not have information about individual number of time an arm in a class has been sampled, lemma 43 roughly tells us than on average, the lower bound on the minimal amount of exploration of an arm in a suboptimal class has been divided by q.

Lemma 44. *If all suboptimal classes are in the non-combinatorial regime, the regret may be asymptotically lower bounded by*

$$\liminf_{T\to\infty} \frac{R(\nu,T)}{\log T} \geqslant \frac{1}{q} \sum_{a\in\mathcal{A}\setminus\mathcal{A}^\star} \frac{\Delta_a}{\mathrm{KL}(\mu_a|\lambda)}. \tag{D.3}$$

Lemma 44 informs us that in the non-combinatorial regime, the classical lower bound on the regret given by equation (1.3) has been divided by q.

Combinatorial regime For a suboptimal class c to be in the combinatorial regime, we need $q < |c| < 2q$. In that case, the lower bound (D.1) involves a combinatorial optimization problem. The difficulty arising from the term

$$\inf_{\mu'\in\mathcal{D}_{\mathrm{sim}}(q,\mu,c_q,\lambda)} \sum_{a\notin c_q} \mathbb{E}_\nu[N_a(T)]\,\mathrm{KL}(\mu_a|\mu'_a)\,,$$

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

Lemma 45. *Let $\nu \in \mathcal{D}_{sim}(q)$ be a bandit configuration having the q-equivalence property and c be a suboptimal class in the combinatorial regime. Then, we have*

$$\liminf_{T\to\infty} \frac{\sum\limits_{a\in c} \mathbb{E}_\nu[N_a(T)]}{\log T} \geqslant \frac{\log(T)}{2q} \sum_{a\in c} \frac{1}{\mathrm{KL}(\mu_a|\lambda)}. \tag{D.4}$$

Those equations can be compared to the equation (D.2) from the non-combinatorial regime. We emphasize the fact that the lower bounds given by equation (D.4) are not the *largest* possible lower bound and hence do not provide as much information about the algorithmically achievable regret as the largest one given by equation (D.1). However, together with a regret upper bound on the algorithm IMED−EC, those quantities will help us control the asymptotic discrepancy between IMED−EC's regret and the asymptotic lower bound given by Theorem 7.

D.3 Information Minimization for bandits with equivalence class

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

$$I_a(t) = N_a(t)\mathrm{KL}(\widehat{\mu}_a(t)|\widehat{\mu}^\star(t)) + \log N_a(t),$$

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

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

$I(t)$ can be computed efficiently by summing the q smallest elements of the list of IMED indexes. The complexity of computing $I(t)$ once the IMED indexes have been computed is at most the one of sorting a list of $|\mathcal{A}|$ elements, $\mathcal{O}(|\mathcal{A}| \log |\mathcal{A}|)$, which is only $\log |\mathcal{A}|$ times larger than looking for the minimal IMED index. Using selection algorithms, we may even achieve a better mean time complexity. The IMED−EC algorithm is presented in Algorithm 12.

Algorithm 12 IMED for Equivalent classes

```
Pull each arm once
for t = |A| ... T − 1 do
  if I*(t) ⩽ I(t) then
    Pull a_{t+1} ∈ arg min_{a∈A*(t)} N_a(t) (chosen arbitrarily)
  else
    Pull a_{t+1} ∈ arg min_{a∉A*(t)} I_a(t) (chosen arbitrarily)
  end if
end for
```

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

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

D.4 Regret analysis

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

Theorem 8 (Upper bound on the number of pulls)**.** *Under the `IMED-EC` algorithms, the number of pulls of a suboptimal arm a is upper bounded by:*

$$\mathbb{E}_\nu[N_a(T)] \leqslant \frac{\log T}{q\,\mathrm{KL}(\mu_a|\mu^\star)}\,(1+\alpha(\varepsilon)) + f(\varepsilon) \tag{D.5}$$

where $0 < \varepsilon < \frac{1}{3}\min_{a\in\mathcal{A}\setminus\mathcal{A}^\star}(\mu^\star - \mu_a)$ α and α and f tends to 0 as ε tends to 0.

Corollary 5. *Under the `IMED-EC` algorithms, the number of pulls of a suboptimal arm a is upper bounded by:*

$$\min_{c_q\subseteq c}\sum_{a\in c_q}\mathbb{E}_\nu[N_a(T)]\,\mathrm{KL}(\mu_a|\mu^\star) \leqslant (1+\alpha(\varepsilon))\log T + g(\varepsilon) \tag{D.6}$$

where $0 < \varepsilon < \frac{1}{3}\min_{a\in\mathcal{A}\setminus\mathcal{A}^\star}(\mu^\star - \mu_a)$ α and α and g tends to 0 as ε tends to 0.

Theorem 9 (Asymptotic upper bound on the number of pulls)**.** *Under the `IMED-EC` algorithms, the number of pulls of a suboptimal arm a is asymptotically upper bounded by:*

$$\liminf_{t\to+\infty}\frac{\mathbb{E}_\nu[N_a(T)]}{\log T} \leqslant \frac{1}{q\,\mathrm{KL}(\mu_a|\mu^\star)} \tag{D.7}$$

Discussion This upper bound shows that in particular, the number of pulls of a suboptimal class, $\sum_{a\in c}\mathbb{E}_\nu[N_a(T)]$ is asymptotically no more than $\frac{|c|}{q\,\mathrm{KL}(\mu_a|\mu^\star)}\log T$. This hence matches the lower bound in the *non-combinatorial regime*. In the *combinatorial regime*, along with equation (D.4), this regret upper bound shows that

$$\frac{|c|}{q\,\mathrm{KL}(\mu_a|\mu^\star)} \geqslant \liminf_{T\to\infty}\sum_{a\in c}\frac{\mathbb{E}_\nu[N_a(T)]}{\log T} \geqslant \frac{1}{2}\cdot\frac{|c|}{q\,\mathrm{KL}(\mu_a|\mu^\star)}$$

proving that the regret of the proposed `IMED-EC` does not differ from the optimal lower bound by a factor more than 2. This is striking result.

Full proof of Theorem 9 and Theorem 8 are provided in Appendix C from Pesquerel et al. (2021).

D.5 Experiments

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

Balanced class, perfect knowledge In this set of experiments, see Figure D.1, we focus on the bandit configurations in which all equivalence classes have the same cardinality and assume that we know the number of elements per class. This setting is interesting for two reasons. First, one can compute the theoretical lower-bound without solving a combinatorial optimization problem. Second, the theoretical analysis shows that IMED−EC is asymptotically optimal in this case. This setting will thus allow us to numerically grasp what happens in the most structured case. We compare compare IMED−EC to unspecialized bandit algorithm, UCB, IMED and KLUCB. To make the comparison fairer we also compare IMED−EC to OSSB, an algorithm specialized in structured bandit. Since OSSB has to solve a combinatorial optimization problem at each time step, we can cannot carry experiments on large set of arms while comparing IMED−EC to it. In this particular setting,

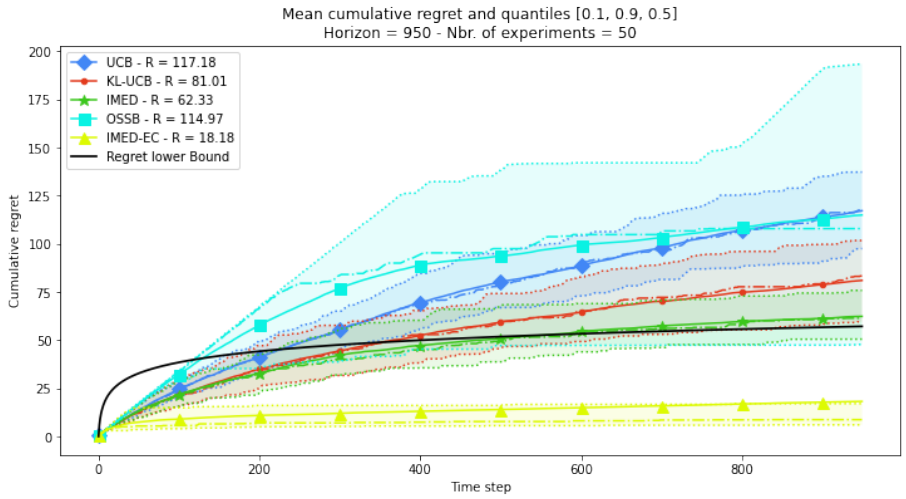


Figure D.1: 3 classes, 3 distributions per class - set of means = $(0.1, 0.3, 0.6, 0.8)$

we see that while OSSB and IMED−EC are provably asymptotically optimal, IMED−EC numerically performs better in finite time horizon. We recall that it is furthermore numerically more efficient since it does not involve any combinatorial optimization. Without too much surprises, IMED−EC also outperforms unspecialized algorithm.

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

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

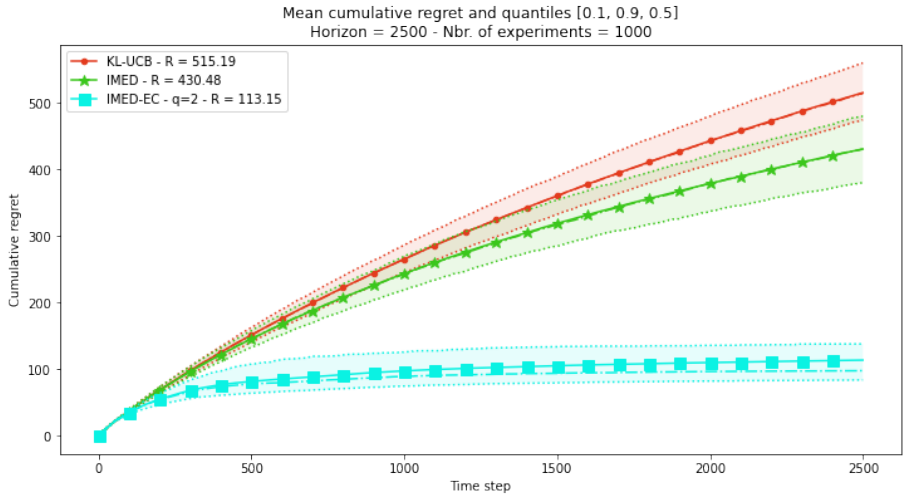


Figure D.2: 7 classes, 8 distributions per class - set of means = $(0.1, 0.3, 0.4, 0.5, 0.6, 0.75, 0.9)$

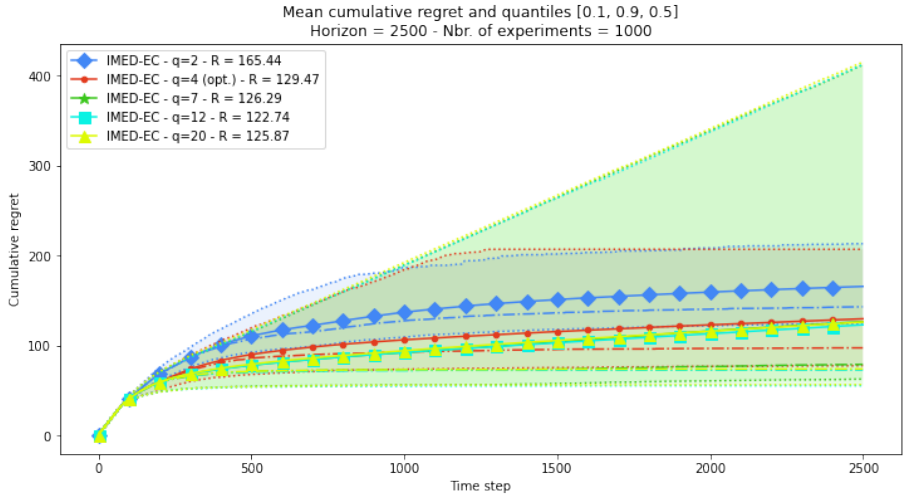


Figure D.3: 7 classes, unbalanced - set of means = $(0.1, 0.3, 0.4, 0.5, 0.6, 0.75, 0.9)$

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

D.6 Conclusion

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