Université de Lille
Ecole Doctorale MADIS

Thèse de Doctorat

Spécialité **Informatique**

présentée par
Marc Jourdan

Solving Pure Exploration Problems with the Top Two Approach

Résoudre les Problèmes d'Exploration Pure avec l'Approche Top Two

---

sous la direction d' **Emilie Kaufmann** et de **Rémy Degenne**.

Soutenue publiquement à **Villeneuve d'Ascq**, le **14/06/2024** devant le jury composé de

M. Alexandre **Proutière**	Professeur	KTH	Rapporteur
M. Sandeep **Juneja**	Professeur	TIFR	Rapporteur
M. Aurélien **Garivier**	Professeur	ENS Lyon	Examinateur
M. Wouter M. **Koolen**	Professeur	CWI	Examinateur
M. Junya **Honda**	Professeur	Kyoto University	Invité
M^{me} Emilie **Kaufmann**	Chargée de recherche	CNRS	Directrice de thèse
M. Rémy **Degenne**	Chercheur	Inria	Co-encadrant de thèse

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

Résumé

Dans les problèmes d'exploration pure pour les bandits stochastiques à bras multiples, l'objectif est de répondre à des questions concernant un ensemble de distributions inconnues (modélisant par exemple l'efficacité d'un traitement) à partir desquelles nous pouvons collecter des échantillons (mesurer son effet), et de fournir ensuite des garanties sur la réponse proposée. L'exemple archétypal est le problème de l'identification du meilleur bras, dans lequel l'agent cherche à identifier le bras étant le plus efficace en moyenne.

Cette thèse s'intéresse à la classe des algorithmes Top Two, dans lesquels un leader est opposé à un challenger, ce qui oriente les efforts d'échantillonnage ultérieurs pour valider la supériorité du leader. Nous avons introduit une définition unifiée de l'approche Top Two, mettant en avant quatre composants importants. Compte tenu de leur simplicité, de leur interprétabilité, de leur généralisation et de leur polyvalence, les algorithmes Top Two sont prometteurs pour être adoptés pour différentes applications. Cette thèse s'efforce d'établir l'approche Top Two comme une méthodologie fondée sur des principes statistiques, offrant des garanties théoriques quasiment optimales ainsi que des performances empirique excellentes.

Nous abordons différentes formulations de bandits stochastiques à plusieurs bras, avec des classes de distributions variées ou des hypothèses structurelles sur les moyennes. Nous avons aussi étudié différents problèmes d'exploration pure, notamment l'identification du meilleur bras ou d'un bras de qualité acceptable. La principale contribution de cette thèse réside dans l'obtention de garanties théoriques pour l'approche Top Two avec plusieurs mesures de performance. Dans le cas où un niveau de confiance est donné, les algorithmes Top Two collectent un nombre moyen d'échantillons qui est asymptotiquement optimal (lorsque le niveau de confiance tend vers un). Par ailleurs, nous proposons un algorithme Top Two qui offre à tout moment des garanties sur la probabilité de se tromper dans l'identification d'un bras de qualité acceptable.

Mots-clefs : prise de décision séquentielle, problème de bandit à plusieurs bras, exploration pure, identification du meilleur bras.

Abstract

In pure exploration problems for stochastic multi-armed bandits, the objective is to answer inquiries regarding a set of unknown distributions (modeling for example the efficacy of a treatment) from which we can collect samples (measure its effect), and subsequently provide guarantees on the candidate answer. The archetypal example is the best arm identification problem, in which the agent aims at identifying the arm with the highest mean.

This thesis delves into the class of Top Two algorithms, wherein a leader is pitted against a challenger, directing subsequent sampling efforts to validate the superiority of the leader. We introduce a unified definition of the Top Two approach, putting forward four key components. Given their simplicity, interpretability, generalizability, and versatility, Top Two algorithms are promising for widespread adoption among practitioners. This thesis endeavors to establish the Top Two approach as a principled methodology offering nearly optimal theoretical guarantees alongside state-of-the-art empirical performance.

We address several stochastic multi-armed bandits settings, such as various classes of distributions or structural assumptions on the means. We also study different pure exploration problems, including the identification of the best arm or one of acceptable quality. The principal contribution of this thesis lies in establishing theoretical guarantees for the Top Two approach across several performance metrics. In the fixed-confidence setting, we prove that many Top Two algorithms have an asymptotically optimal expected sample complexity (number of collected samples when the confidence level goes to one). In the anytime setting, we propose a Top Two algorithm which has guarantees on the probability of misidentifying a good enough arm at any time.

Keywords: sequential decision making, multi-armed bandits, pure exploration, best-arm identification.

Table of Contents

Chapter 1

Introduction

This manuscript concludes my doctoral thesis which started in October 2021 and took place in the Scool team, hosted by Inria and the CRIStAL computer science lab at the University of Lille. I had the honor of being supervised by Dr. Emilie Kaufmann and Dr. Rémy Degenne.

After introducing the context of this thesis, we present formally the setting of pure exploration problems in stochastic multi-armed bandits. Then, we give an overview of the main contributions of this thesis, and how they fit in the existing (and vast) literature. While the contributions presented in this thesis are theoretical, the considered questions have been motivated by practical considerations.

Contents

1.1 Context and Scope

The stochastic multi-armed bandit problem has a rich history, initially conceived as a simple model for sequential clinical trials [Thompson, 1933, Robbins, 1952]. Imagine a clinical trial where a physician seeks to evaluate the efficacy of K potential treatments for a novel disease. Each treatment i carries a certain probability μ_i of curing a patient upon administration.

μ_1 μ_2 μ_3 μ_4

In the phase III of a clinical trial, a central question emerges: which treatment distribution yields the highest efficiency, meaning has the largest mean μ_i? Various methodologies have been explored to address this query, depending on the intricacies of data collection. Sequential hypothesis testing deals with situations where samples are collected without explicit control [Chernoff, 1959, Robbins, 1952]. Experimental design endeavors to predetermine the data collection scheme [Chaloner and Verdinelli, 1995, Pukelsheim, 2006]. In the multi-armed bandit [Audibert et al., 2010, Jamieson and Nowak, 2014] and ranking and selection [Hong et al., 2021] literature, an algorithm sequentially select its sampling strategy based on past data.

As each patient enters the trial, the physician selects a treatment $I_n \in [K]$ for administration, as well as a treatment $\hat{\imath}_n \in [K]$ for recommendation since it is believed to be the best treatment. The patient response to the treatment I_n , modeled as a binary random variable X_{n,I_n} drawn from a Bernoulli distribution with mean μ_{I_n} , guides subsequent decisions. The crux of the matter lies in leveraging this feedback to craft a "good" allocation policy for future patients. The policy's nature depends on the physician's objectives; in phase III trials, the goal is to identify a viable treatment for large-scale production. Given the costs associated with clinical trials, timely certainty regarding treatment efficacy is crucial. An "optimal" policy may thus prioritize swift identification of promising treatments, even at the expense of fewer patient cured during the trial phase, hoping to accelerate the mass production of an efficacious treatment.

While clinical trials serve as a motivating example, the quest for identifying the most efficient item permeates numerous domains. In the context of crop management for agriculture, different fertilization policies (or planting dates) are applied on the fields to assess the yield of the crop. In A/B testing for online marketing, several versions of the same webpage are deployed to evaluate their conversion probability (*e.g.* buying products, spending time, etc). In hyperparameter optimization, different hyperparameters are tested to measure the performance of a model. In those domains, the random response observed when selecting an item (or *arm*)

might not be binary, hence other classes of distributions will also be considered (*e.g.* bounded distributions for crop management).

When identifying the most efficient item is too costly, the practitioner is often willing to settle for a good enough item whose performance is close to the best one. For other applications, the practitioner considers that an item is good enough when its average efficiency is above a certain level. In outcome scoring from gene activity (transcriptomic) data (*e.g.* treatment of encephalopathy of prematurity in infants), several protocols for the administration of stem cells are tested and the goal is only to identify one protocol that yields a strong enough positive effect on patients. Before testing the potency of a drug against other drugs, one should first evaluate its toxicity. In a toxicity study, several criteria are evaluated, and the physician should ensure that none of those toxicity levels is large enough.

Rather than focusing on individual applications, this thesis considers the mathematical framework of pure exploration problems for stochastic multi-armed bandits. As statisticians, our aim is to design policies (or *strategies*) which have simultaneously good theoretical guarantees and good empirical performance. To discern between strategies with comparable theoretical and empirical merits, we must consider the end-user (*i.e.* the practitioner). Thus, we endeavor to craft policies that are simple, interpretable, generalizable, and versatile. We believe that those characteristics are key for a policy to become widely accepted and utilized.

Simple There is often a difference between the policy and the agent: the former suggests an action and the later can decide to perform it. Since rational agents will only follow a strategy in which they believe in, the policy should be simple in order for the practitioner to understand it and implement it (preferably efficiently). In this thesis, we assume implicitly that both notions are the same, but in a real-world scenario this is often not true.

Interpretable Transparency and accountability are paramount, especially in domains fraught with ethical implications. Interpretability not only aids understanding but also justifies decisions, crucial for gaining stakeholder trust and regulatory approval. For example, in phase III of clinical trials, the treatment is administered to a patient in a pool of volunteers. Interpretability is a requirement of health authorities and it is necessary to secure volunteers since extrinsic motivation (*e.g.* sickness or money) might not outweigh the defiance towards a black-box procedure. To address the ethical dilemma of curing less patient during the trials to obtain a cure faster, the status (or purpose) of the allocated treatment should be made clear. Either it is believed to be a good treatment or it is only administered to reduce uncertainty on treatments that are believed to be sub-optimal.

Generalizable Theoretical guarantees hinge on well-defined goals (*e.g.* identify the best item, or a good enough one) and environmental assumptions (*e.g.* the class of distributions, or the underlying structure). Yet, practitioners seek methodologies adaptable to varied scenarios, removing the need to frequently learn a new methodology. Therefore, the design approach which led to the policy should be generalizable to cope for changes of the goals and assumptions with limited modifications.

Versatile Constraints in decision-making scenarios can fluctuate unpredictably. Policies must withstand such dynamism, offering guarantees that remain robust across changing contexts. For example, a physician might decide to stop earlier (resp. continue) the ongoing experiments due to insufficient (resp. additional) funding. Therefore, a policy should be versatile enough to obtain guarantees that hold at any time.

The Top Two approach In essence, the Top Two approach is pitting a leader against a challenger, and guiding subsequent sampling to verify the leader's superiority. This thesis endeavors to demonstrate the efficacy of the Top Two approach. It is a principled methodology that meets the four aforementioned criteria, offering near-optimal theoretical guarantees and good empirical performance.

1.2 Stochastic Multi-Armed Bandits

We study the stochastic multi-armed *bandit* problem [Bubeck and Cesa-Bianchi, 2012, Lattimore and Szepesvari, 2019], which allows us to reflect on fundamental information-utility trade-offs involved in interactive sequential learning. Specifically, in a bandit model, an *agent* is interacting with an environment composed of $K \in \mathbb{N}$ *arms*. Each arm $i \in [K]$[1] is associated with an unknown probability distribution over $\mathbb{R}$ denoted by $\nu_i \in \mathcal{P}(\mathbb{R})$ and having a finite mean $\mu_i = \mathbb{E}_{X\sim\nu_i}[X]$. A *bandit instance* is uniquely characterized by its vector of distributions $\nu = (\nu_i)_{i\in[K]}$, which admits $\mu = (\mu_i)_{i\in[K]}$ as vector of means.

At each stage $n \in \mathbb{N}$, the agent chooses an arm $I_n \in [K]$ based on the samples previously observed and receives a sample X_{n,I_n} , random variable with conditional distribution ν_{I_n} given I_n . It then proceeds to the next stage. An *algorithm* (or *strategy*) for the agent in this interaction is specified by a *sampling rule*, a procedure that determines I_n based on previously observed samples and some exogenous randomness. Formally, I_n is measurable with respect to the σ-algebra $\mathcal{F}_n := \sigma\left(U_1, I_1, X_{1,I_1}, \cdots, I_{n-1}, X_{n-1,I_{n-1}}, U_n\right)$, where $U_n \sim \mathcal{U}([0,1])$ materializes the possible independent randomness used by the algorithm at time n . We call that σ-algebra *history* before time n . The empirical allocation over arms is denoted by $N_n := (N_{n,i})_{i\in[K]}$ where

[1] $[K]$ is a shorthand for $\{1, \cdots K\}$.

$N_{n,i} := \sum_{t\in[n-1]} \mathbb{1}(I_t = i)$, and we have $N_n/(n-1) \in \Sigma_K := \left\{w \in \mathbb{R}_+^K \mid w \geq 0,\ \sum_{i\in[K]} w_i = 1\right\}$ where Σ_K denotes the probability simplex of dimension $K-1$.

Depending on their objectives, agents should have different sampling strategies. In the regret minimization problems [Auer et al., 2002], the agent aims at maximizing the reward accumulated over time, or equivalently minimizing the regret. In the pure exploration problems [Even-Dar et al., 2002, Bubeck et al., 2009], the agent solely wants to answer a question about the underlying distributions. The most well studied pure exploration problem is best-arm identification (BAI), in which the goal is to identify an arm with largest mean Audibert et al. [2010]. In this thesis, we focus solely on the pure exploration problems, for which we provide more details in Section 1.3.

Class of distributions When modeling a bandit problem, one first needs to make an assumption as regards the set of possible distributions $\mathcal{D}$ for the arms. The set of possible means is denoted by $\mathcal{I} := \{m(\kappa) \mid \kappa \in \mathcal{D}\}$ where $m(\kappa) := \mathbb{E}_{X\sim\kappa}[X]$. From the viewpoint of a practitioner, $\mathcal{D}$ should be chosen as a simple set of distributions which provides a good approximation of the real-world application. For example, parametric distributions are reasonable for applications such as A/B testing [Kaufmann et al., 2014], but they are unrealistic in other fields such as agriculture. From the perspective of a theoretician, the choice of $\mathcal{D}$ is guided by the type of research directions. For example, Gaussian distributions are a natural first step to study a new phenomenon or provide new guarantees for an algorithm. Hopefully the insights gained will then be generalized to wider classes of distributions.

In this thesis, we mostly use the set $\mathcal{D}_{\mathcal{N}_\sigma}$ of Gaussian distributions with known variance σ, and the set $\mathcal{D}_\sigma$ of σ-sub-Gaussian distributions. A distribution κ is σ-sub-Gaussian if it satisfies $\mathbb{E}_{X\sim\kappa}[e^{\lambda(X-m(\kappa))}] \leq e^{\sigma^2\lambda^2/2}$ for all $\lambda \in \mathbb{R}$. Most of the methods used to study $\mathcal{D}_{\mathcal{N}_\sigma}$ can be transferred for the analysis of the set $\mathcal{D}_{\exp}$ of one-parameter exponential family. This thesis will present more challenging extensions with a two-parameters exponential family, *e.g.* the set $\mathcal{D}_{\mathcal{N}}$ of Gaussian distributions with unknown variance (see Chapter 3), and a set of non-parametric distributions, *e.g.* the set $\mathcal{D}_{[0,B]}$ of bounded distributions, whose support lies in $[0, B]$ where $B > 0$ (see Chapter 4).

Other classes of non-parametric distributions could have been considered, *e.g.* heavy-tailed distributions with an upper bound on a non-centered moment [Agrawal et al., 2020]. Classes of shape constrained distributions are also promising for future work, *e.g.* logarithmically concave density functions.

Underlying structure To define a bandit problem, one also needs to make an assumption as regards its underlying structure. By structure, we mean an encoding of some prior knowledge on the vector of means μ which should lies in a known subset $\mathcal{S}$ of the set $\mathcal{I}^K$ of possible vectors

of means. The choice of the structure should be done with respect to the same trade-off as for the choice of the class of distributions, both for the practitioner and the theoretician.

In this thesis, we mostly study the *vanilla* (or *unstructured*) bandit problem in which the means are independent, *i.e.* $\mathcal{S} = \mathcal{I}^K$. Despite the "simplicity" of the vanilla bandits, there are still many open problems which remain to be answered. Since the structure can create new information-utility trade-offs, we also explore one structured bandit problem in Part III. In the *linear* bandit problem, each arm is associated with a known context vector $a_i \in \mathbb{R}^d$ and has a mean which is a linear function of unknown vector $\theta \in \mathcal{M}$ where $\mathcal{M} \subseteq \mathbb{R}^d$ is a bounded set, *i.e.* $\mu_i = \langle a_i, \theta \rangle$. In the linear setting, μ is fully characterized by the set of arms vector $\mathcal{A} = \{a_i\}_{i \in [K]}$ and the regression parameter θ , *i.e.* $\mathcal{S} = \{\mu \in \mathcal{I}^K \mid \exists \theta \in \mathcal{M}, \forall i \in [K], \ \langle a_i, \theta \rangle\}$.

Other structural assumption could have been studied in the literature: generalized linear bandits [Filippi et al., 2010] such as logistic bandits [Jun et al., 2021], combinatorial bandits [Chen et al., 2013], sparse bandits [Jamieson et al., 2015], spectral bandits [Kocák and Garivier, 2021], unimodal bandits [Combes and Proutière, 2014, Trinh et al., 2020], Lipschitz [Magureanu et al., 2014], partial monitoring [Audibert and Bubeck, 2010], etc.

1.3 Pure Exploration Problems

In pure exploration problems, the goal is to answer a question about the unknown environment by interacting with the set of K arms. In this thesis, we focus on questions whose answers are a function of the unknown vector of means. However, similar questions could be formulated on other functionals of the distribution [Wang et al., 2022], *e.g.* the conditional value at risk for heavy-tailed distributions [Agrawal et al., 2021b].

A wide range of questions Two types of questions about μ have been studied in the literature, either there is a unique correct answer or there are multiple correct answer. The set $\mathcal{Z}$ of possible answers to a pure exploration problems can coincide with the set of arms, *i.e.* $\mathcal{Z} = [K]$, or have a more elaborate structure, *e.g.* answers can be subsets of arms (*i.e.* $\mathcal{Z} \subseteq 2^{[K]}$). Its cardinality is also assume to be finite and we denote it by $Z = |\mathcal{Z}|$.

The most studied topic in pure exploration is the best arm identification (BAI) problem, which we tackle in Part I. In BAI, the agent aims at identifying an arm with highest mean, *i.e.* $i^\star \in i^\star(\mu) = \arg\max_{i \in [K]} \mu_i$. In (exact) BAI, we consider $\mathcal{S} = \{\mu \in \mathcal{I}^K \mid |i^\star(\mu)| = 1\}$, hence there is a unique correct answer $i^\star$ and $\mathcal{Z} = [K]$. In some applications such as investigating treatment protocols, BAI requires too many samples for it to be useful in practice. To avoid wasteful queries, practitioners might be interested in easier tasks that identify one "good enough" option. We consider two relaxed identification problems which admit multiple correct answers: epsilon best arm identification (ε-BAI) and good arm identification (GAI).

In ε-BAI [Mannor and Tsitsiklis, 2004, Even-Dar et al., 2006, Garivier and Kaufmann, 2021], which is the focus of Chapter 5, the agent is interested in an arm whose mean is ε-close to the highest one $\mu_\star := \max_{i\in[K]} \mu_i$, *i.e.* $i \in \mathcal{I}_\varepsilon(\mu) := \{i \in [K] \mid \mu_i \geq \mu_\star - \varepsilon\}$ where $\varepsilon \geq 0$ and $\mathcal{S} = \mathcal{I}^K$. The larger ε is, the easier the task. In the multiplicative ε-BAI problem, the means are non-negative, *i.e.* $\mathcal{S} = \{\mu \in \mathcal{I}^K \mid \min_{i\in[K]} \mu_i \geq 0\}$, and one aims at returning an arm $i \in \mathcal{I}^{\mathrm{mul}}_\varepsilon(\mu) := \{i \in [K] \mid \mu_i \geq (1-\varepsilon)\mu_\star\}$ where $\varepsilon \in [0, 1)$. The BAI setting is recovered by taking $\varepsilon = 0$ and considering $\mathcal{S} = \{\mu \in \mathcal{I}^K \mid |i^\star(\mu)| = 1\}$. In GAI, which we tackle in Chapter 6, the agent wants to return an arm whose mean exceeds a given threshold γ if it exists, *i.e.* $i \in \mathcal{I}^{\mathrm{thr}}_\gamma(\mu) := \{i \in [K] \mid \mu_i \geq \gamma\}$ where $\gamma \in \mathbb{R}$, else return $\emptyset$. In GAI, we have $\mathcal{Z} = [K] \cup \emptyset$, and we consider $\mathcal{S} = \{\mu \in \mathcal{I}^K \mid \min_{i\in[K]} |\mu_i - \gamma| > 0\}$.

Other pure exploration problems could have been studied in the literature: thresholding bandits [Locatelli et al., 2016, Cheshire et al., 2021, Ouhamma et al., 2021], Top-k identification [Katz-Samuels and Scott, 2019, Réda et al., 2021, Tirinzoni and Degenne, 2022], Pareto set identification [Auer et al., 2016, Kone et al., 2023, 2024], identifying the whole set $\mathcal{I}_\varepsilon(\mu)$ [Mason et al., 2020, Marjani et al., 2022], best partition identification [Chen et al., 2017a, Juneja and Krishnasamy, 2019, Kaufmann and Koolen, 2021], structured BAI [Huang et al., 2017], etc.

Recommendation rule Since the goal of the agent is to answer a question, its strategy should include a *recommendation* rule in addition to a sampling rule. At time n , the agent recommends a *candidate* answer $\hat{\imath}_n$ based on the samples previously observed. The recommendation $\hat{\imath}_n$ is done before pulling arm I_n , hence it is measurable with respect to the filtration $\mathcal{F}_n$. Depending on the metric of performance, the recommendation rule can only be specified at a fixed time T or when a data-dependent stopping condition is met.

Performance metrics There are several ways to evaluate the performance of an algorithm on a pure exploration problem. While those metrics are function of the candidate answer, they inherently depend on the sampling rule due to the data dependency. The two major theoretical frameworks are the *fixed-confidence* setting [Even-Dar et al., 2006, Jamieson and Nowak, 2014, Garivier and Kaufmann, 2016], which is the main focus of this thesis, and the *fixed-budget* setting [Audibert et al., 2010, Gabillon et al., 2012]. In the fixed-confidence setting, the agent aims at minimizing the number of samples used to identify a correct answer with confidence $1 - \delta \in (0, 1)$. In the fixed-budget setting, the objective is to minimize the probability of misidentifying a correct answer with a fixed number of samples T .

While the constraint on δ or T is supposed to be given, properly choosing it is challenging for the practitioner since a "good" choice typically depends on unknown quantities. Moreover, in medical applications such as clinical trials, the maximal budget is limited but might not be fixed beforehand. When the collected data shows sufficient evidence in favor of one answer, an

experiment is often stopped before the initial budget is reached, referred to as *early stopping*. When additional sampling budget have been obtained due to new funding, an experiment can continue after the initial budget has been consumed, referred to as *continuation*. While early stopping and continuation are common practices, both fixed-confidence and fixed-budget settings fail to provide useful guarantees for them. Recently, the *anytime* setting has received increased scrutiny as it fills this gap between theory and practice. In the anytime setting, the agent aims at achieving a low probability of error at any deterministic time [Jun and Nowak, 2016, Zhao et al., 2023]. In this thesis, an anytime strategy should also have good guarantees in the fixed-confidence setting. As such, an anytime strategy is versatile and can be used both in the fixed-confidence and fixed-budget settings without modification. When the candidate answer has anytime guarantees, the practitioners can use continuation or early stopping (when combined with a stopping rule). The anytime setting is presented in more details in Section 1.5, as it will be the focus of Part II.

Instead of considering minimax guarantees (*e.g.* minimax optimiality of uniform sampling for the probability of error [Bubeck et al., 2011]), this thesis focuses on *problem-dependent* guarantees on the considered metric of performance. The goal is to show that the strategy will (optimally) adapt to the bandit instance ν . In a nutshell, the contributions of this thesis are to derive upper bounds on the performance of algorithms, which are themselves designed by studying the theoretical lower bound.

1.4 Fixed-confidence Setting

In the fixed-confidence setting, the agent is given a parameter $\delta \in (0,1)$. In addition to its sampling and recommendation rule, the agent should define a *stopping* rule which is a stopping time with respect to the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$. Note that the recommendation needs only to be defined at the stopping time in the fixed-confidence setting. The stopping time τ_δ is also called the sample complexity of the algorithm. The main requirement that we impose on a fixed-confidence identification strategy is to be δ-correct.[2]

Definition 1.1 (δ-correct)**.** *Given $\delta \in (0,1)$, we say that an algorithm is* δ-correct *on the problem class $\mathcal{D}^K$ with means $\mathcal{S}$ if its probability of stopping and not recommending a correct answer is upper bounded by δ for all instances $\nu \in \mathcal{D}^K$ having mean $\mu \in \mathcal{S}$,* i.e.

$$\forall \nu \in \mathcal{D}^K \text{ s.t. } \mu \in \mathcal{S}, \quad \mathbb{P}_\nu(\{\tau_\delta < +\infty\} \cap \mathcal{E}_\mu^{\mathrm{err}}(\tau_\delta)) \leq \delta \,, \tag{1.1}$$

where $\mathcal{E}_\mu^{\mathrm{err}}(n)$ denotes the error *event at time n , meaning that $\hat{\imath}_n$ is not a correct answer for μ .*

[2] A stronger definition of δ-correctness has also been studied by requiring the algorithm to stop almost surely.

Example 1.2. *In ε-BAI, we have $\mathcal{E}_\mu^{\mathrm{err}}(n) = \{\hat{\imath}_n \notin \mathcal{I}_\varepsilon(\mu)\}$ where $\mathcal{I}_\varepsilon(\mu) = \{i \in [K] \mid \mu_i \geq \mu_\star - \varepsilon\}$. When $\varepsilon > 0$, we will denote the stopping time by $\tau_{\varepsilon,\delta}$, and use the term (ε, δ)-PAC (Probably Approximately Correct) instead of δ-correct. In GAI, we have $\mathcal{E}_\mu^{\mathrm{err}}(n) = \{\hat{\imath}_n \in \{\emptyset\} \cup ([K] \setminus \mathcal{I}_\gamma^{\mathrm{thr}}(\mu))\}$ when $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) = \{i \in [K] \mid \mu_i \geq \gamma\} \neq \emptyset$, otherwise $\mathcal{E}_\mu^{\mathrm{err}}(n) = \{\hat{\imath}_n \neq \emptyset\}$ when $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) = \emptyset$.*

A fixed-confidence identification strategy is judged based on its expected sample complexity $\mathbb{E}_\nu[\tau_\delta]$, *i.e.* the expected number of samples it needs to collect before it can stop and return a correct answer with the required confidence. A "good" fixed-confidence algorithm should minimize $\mathbb{E}_\nu[\tau_\delta]$. Historically, the sample complexity of the first fixed-confidence algorithms was upper bounded with high probability. While there is no high probability lower bound on the sample complexity, the expected sample complexity admits one (see Section 1.4.1). Upper and lower bounds on the expected sample complexity are more informative than high probability bounds. However, they only provide partial information as regards the behavior of the algorithm. For example, we are still lacking a good understanding of the right tail of the distribution of the stopping time τ_δ . It is not clear how fast it can decay, both in terms of a theoretical lower bound and an upper bound for an algorithm.

First, we detail the lower bound on the expected sample complexity of any δ-correct algorithm (Section 1.4.1), which is known to be tight in the asymptotic regime of $\delta \to 0$. Second, we present the GLR (generalized likelihood ratio) stopping rule (Section 1.4.2), which ensures δ-correctness regardless of the sampling rule when used with a well-chosen threshold. Finally, we review several approaches to define a sampling rule reaching the lower bound asymptotically (Section 1.4.3).

1.4.1 Lower Bound on the Expected Sample Complexity

To be δ-correct on $\mathcal{D}^K$, an algorithm has to be able to distinguish problems in $\mathcal{D}^K$ which are disagreeing on one of their correct answers. Garivier and Kaufmann [2016] show that this requirement leads to a problem-dependent lower bound on the expected sample complexity incurred by a δ-correct algorithm on any instance, which is tight in the asymptotic regime of $\delta \to 0$. The lower bound is obtained by using properties of the Kullback-Leibler (KL) divergence and change-of-distribution arguments with respect to alternative bandit instances. Change-of-distribution arguments were also used by previous (less tight) lower bounds on the expected sample complexity [Mannor and Tsitsiklis, 2004, Kaufmann et al., 2016].

Kullback-Leibler divergence For two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ on the same measurable space $\mathcal{X}$, the KL divergence (or relative entropy) is $\mathrm{KL}(\mathbb{P}, \mathbb{Q}) := \mathbb{E}_{X \sim \mathbb{P}}\left[\log \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(X)\right]$,

when $\mathbb{P} \ll \mathbb{Q}$, and $+\infty$ otherwise. $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}$ is the Radon–Nikodym derivative of $\mathbb{P}$ with respect to $\mathbb{Q}$. Given a distribution $\mathbb{P}^{(0)}$ with cumulant generating function ϕ , defined on an interval $\mathcal{I}_\phi$, the one-parameter exponential families defined by $\mathbb{P}^{(0)}$ is the set of distributions $\mathbb{P}^{(\lambda)}$ with density with respect to $\mathbb{P}^{(0)}$ given by $\frac{d\mathbb{P}^{(\lambda)}}{d\mathbb{P}^{(0)}}(x) = e^{\lambda x - \phi(\lambda)}$. For one-parameter exponential families, $d_{\mathrm{KL}}(x, y)$ denotes the KL divergence between the distributions having means (x, y) . The KL divergence between two Bernoulli distributions (also known as binary relative entropy) is denoted by kl , and satisfies $\mathrm{kl}(x, y) = x \log(x/y) + (1 - x) \log(1 - x)/(1 - y))$. The *data-processing* inequality (or *contraction of entropy*) is a useful tool to derive lower bounds. It states that $\mathrm{KL}(\mathbb{P}, \mathbb{Q}) \geq \mathrm{KL}(\mathbb{P}^f, \mathbb{Q}^f)$ where $f : \mathcal{X} \to \mathcal{Y}$ is a measurable function and $\mathbb{P}^f$ is the push-forward measure of $\mathbb{P}$ by f .

Change of distribution A change of distribution relates the probability of an event $\mathcal{E}$ under two different probability distributions ν and κ with their log-likelihood ratio. Let

$$\mathcal{L}_n(\nu, \kappa) = \log \frac{\ell(X_{1,I_1}, \cdots, X_{n-1,I_{n-1}}; \nu)}{\ell(X_{1,I_1}, \cdots, X_{n-1,I_{n-1}}; \kappa)} \tag{1.2}$$

denote the log-likelihood ratio of the observations collected before time n (*i.e.* in the history $\mathcal{F}_n$). The data-processing inequality yields that, for any stopping time τ and any event $\mathcal{E} \in \mathcal{F}_\tau$,

$$\mathbb{E}_\nu[\mathcal{L}_\tau(\nu, \kappa)] = \mathrm{KL}(\nu^{F_\tau}, \kappa^{F_\tau}) \geq \mathrm{kl}(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_\kappa(\mathcal{E})) \,, \tag{1.3}$$

where $F_n = (U_1, I_1, X_{1,I_1}, \cdots, I_{n-1}, X_{n,I_{n-1}}, U_n)$ is such that $\mathcal{F}_n = \sigma(F_n)$. When there is a unique correct answer, (1.3) is enough to obtain a lower bound on the expected sample complexity. However, when there are multiple correct answers, a "lower level" change of distribution is required [Garivier and Kaufmann, 2021, Degenne and Koolen, 2019]. It features the right tail of the log-likelihood ratio instead of its expectation. Direct manipulations yield that, for all $x \in \mathbb{R}$, all time $n \in \mathbb{N}$ and all event $\mathcal{E} \in \mathcal{F}_n$,

$$\mathbb{P}_\kappa(\mathcal{E}) \geq e^{-x} \left(\mathbb{P}_\nu(\mathcal{E}) - \mathbb{P}_\nu(\mathcal{L}_n(\nu, \kappa) \geq x)\right) \,. \tag{1.4}$$

Both in (1.3) and (1.4), the event $\mathcal{E}$ is chosen such that it is controlled under both distributions, likely for one and unlikely for the other. Therefore, we want to find an alternative bandit model $\kappa \in \mathcal{D}^K$ whose vector of mean $\lambda \in \mathcal{S}$ is close enough to μ but under which the algorithm should behave differently since ν and κ are disagreeing on their correct answers. Empirically, different behavior are reflected by a large log-likelihood ratio. As the log-likelihood ratio is linked to the empirical allocation N_n , those change of distribution arguments will translate into constraints on the sample complexity.

Alternative set Given an answer $i \in \mathcal{Z}$, we define the *alternative to* i as the set $\neg i$ of vectors of means $\lambda \in \mathcal{S}$ such that i is not a correct answer for λ. When there is a unique correct answer, the set $\mathrm{Alt}(\nu)$ of alternative bandit instances to ν is defined as the set of bandit instances $\kappa \in \mathcal{D}^K$ such that the mean vector $m(\kappa)$ lies in the alternative to the correct answer associated with μ.

Example 1.3. *For BAI, we have* $\neg i = \overline{\{\lambda \in \mathcal{S} \mid i \notin i^\star(\lambda)\}}$ *where* $\overline{X}$ *denotes the closure of* X, *hence* $\mathit{Alt}(\nu) := \{\kappa \in \mathcal{D}^K \mid m(\kappa) \in \neg i^\star\}$ *where* $i^\star$ *is the unique element of* $i^\star(\mu)$. *In* ε*-BAI, we have* $\neg_\varepsilon i = \overline{\{\lambda \in \mathcal{S} \mid i \notin \mathcal{I}_\varepsilon(\lambda)\}}$. *In GAI, we have* $\neg i = \overline{\{\lambda \in \mathcal{S} \mid i \notin \mathcal{I}_\gamma^{\mathrm{thr}}(\lambda)\}}$ *when* $i \in [K]$, *otherwise* $\neg i = \overline{\{\lambda \in \mathcal{S} \mid i \neq \mathcal{I}_\gamma^{\mathrm{thr}}(\lambda)\}}$ *when* $i = \emptyset$.

Lower bound For simplicity of exposure and since it covers BAI, we first consider the case where there is a unique correct answer. The following elegant information-theoretic proof of Theorem 1.4 was given by Garivier et al. [2019].

Theorem 1.4 ([Garivier and Kaufmann, 2016, Agrawal et al., 2020])**.** *An algorithm which is* δ*-correct on all problems in* $\mathcal{D}^K$ *satisfies that, for all* $\nu \in \mathcal{D}^K$ *with mean* $\mu \in \mathcal{S}$,

$$\mathbb{E}_\nu[\tau_\delta] \geq T^\star(\nu) \log\left(\frac{1}{2.4\delta}\right), \tag{1.5}$$

where $T^\star(\nu) = \min_{\beta \in (0,1)} T^\star_\beta(\nu)$ *is a* characteristic time *whose inverse is defined as*

$$T^\star_\beta(\nu)^{-1} := \sup_{w \in \Sigma_K,\, w_{i^\star} = \beta} \inf_{\kappa \in \mathit{Alt}(\nu)} \sum_{i \in [K]} w_i \mathrm{KL}(\nu_i, \kappa_i). \tag{1.6}$$

Proof. When $\mathbb{P}_\nu(\tau_\delta < +\infty) < 1$, we have $\mathbb{E}_\nu[\tau_\delta] = +\infty$, hence any lower bound will hold true. In the following, we suppose that $\tau_\delta < +\infty$ almost surely. Let $\delta \in (0,1)$, $\kappa \in \mathrm{Alt}(\nu)$ and $\mathcal{E}_\delta = \mathcal{E}^{\mathrm{err}}_\mu(\tau_\delta)$. Since answers are unique, being correct on instance κ implies that the algorithm is not correct on the instance ν, *i.e.* $\mathcal{E}^{\mathrm{err}}_{m(\kappa)}(\tau_\delta)^{\complement} \subseteq \mathcal{E}_\delta$ where $X^{\complement}$ denotes the complement of X. This key argument does not hold when there are multiple correct answers. The δ-correctness property yields that $\mathbb{P}_\nu(\mathcal{E}_\delta) \leq \delta$ and $\mathbb{P}_\kappa(\mathcal{E}_\delta) \geq 1 - \delta$. Using (1.3) and Wald's lemma, we obtain

$$\sum_{i \in [K]} \mathbb{E}_\nu[N_{n,i}] \mathrm{KL}(\nu_i, \kappa_i) = \mathbb{E}_\nu[\mathcal{L}_{\tau_\delta}(\nu, \kappa)] \geq \mathrm{kl}(\mathbb{P}_\nu(\mathcal{E}_\delta), \mathbb{P}_\kappa(\mathcal{E}_\delta)) \geq \mathrm{kl}(\delta, 1-\delta) \geq \log\left(\frac{1}{2.4\delta}\right),$$

where the last two inequalities use the monotonicity properties of kl and a known lower bound. Then, we can normalize to factor $\mathbb{E}_\nu[\tau_\delta]$ in the *r.h.s.* of the above equation. Taking the infimum over $\kappa \in \mathrm{Alt}(\nu)$ and the supremum over $w \in \Sigma_K$ concludes the proof. ■

The set $w^\star(\nu)$ (resp. $w^\star_\beta(\nu)$) of (resp. β-)*optimal allocations* is defined as the maximizer of the outer supremum on Σ_K which defines $T^\star(\nu)^{-1}$ (resp. $T^\star_\beta(\nu)^{-1}$). The inverse of the characteristic time quantifies the dissimilarity between ν and the *most confusing* (or *closest*) alternative bandit κ . The notion of dissimilarity is a reweighted summation of KL divergence, and the reweighting is chosen to maximize the dissimilarity.

In the *asymptotic* regime where $\delta \to 0$, the lower bound (1.5) is known to be tight since it is achieved by several algorithms. An algorithm is said to be *asymptotically optimal* (resp. *β-optimal*) on $\mathcal{D}^K$ if this asymptotic lower bound is matched on all instances ν , $\limsup_{\delta\to 0} \mathbb{E}_\nu[\tau_\delta]/\log(1/\delta) \le T^\star(\nu)$ (resp. $T^\star_\beta(\nu)$) for all $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathcal{S}$. In the non-asymptotic (or *moderate* confidence) regime where $\delta \in (0,1)$ is not necessarily close to 0, the lower bound (1.5) only provides part of the picture and additional lower bounds have been derived to account for different phenomenon. For example, lower bounds of order $\sum_{i=1}^K \Delta_i^{-2} \log\log \Delta_i^{-2}$ (independent of δ , but with a stronger dependence in the gaps) were derived [Jamieson et al., 2014, Chen et al., 2017b, Simchowitz et al., 2017, Chen et al., 2017c]. Deriving a tight lower bound in the non-asymptotic regime is one of the main open problem in the field of fixed-confidence pure exploration.

When there are multiple correct answer, the analysis relies on (1.4) and is purely asymptotic. The asymptotic lower bound features a characteristic time which is similar to $T^\star(\nu)$, and recovers it when there is a unique correct answer. While both characteristic times have a similar interpretation, the most confusing alternative bandit κ is now with respect to one of the multiple correct answers. The set $i_F(\nu)$ of *furthest* (or *easiest-to-verify*) answers is defined as the answers that maximize the corresponding dissimilarity. For example, in ε-BAI, we have

$$\liminf_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau_{\varepsilon,\delta}]}{\log(1/\delta)} \ge T_\varepsilon(\nu) \quad \text{with} \quad T_\varepsilon(\nu)^{-1} := \max_{i\in\mathcal{I}_\varepsilon(\mu)} \sup_{w\in\Sigma_K} \inf_{\kappa\in\mathcal{D}^K, m(\kappa)\in\neg i} \sum_{i\in[K]} w_i \mathrm{KL}(\nu_i,\kappa_i)\,, \tag{1.7}$$

and $i_F(\nu)$ is the maximizer of the outer maximization defining $T_\varepsilon(\nu)$. We refer the reader to Appendix G.1 for an asymptotic proof in the multiple correct answer setting, which is based on techniques developed by Degenne and Koolen [2019].

Illustrative example: BAI Characteristic times are ubiquitous in fixed-confidence pure exploration since they are giving the full picture in the asymptotic regime. Therefore, they have been extensively used as and inspiration for the design of an identification strategy. While the ideas can often be transferred irrespective of the question of interest and the considered structure, we will consider vanilla BAI as an illustrative example. In order to fully cover Part I, the set of distributions could be either parametric (*e.g.* Gaussian with known or unknown variance) or non-parametric (*e.g.* bounded distribution). Agrawal et al. [2020] were the first to study BAI for a class of non-parametric distribution.

Let $w \in \Sigma_K$ and $i^\star$ be the unique correct answer of ν , hence $\mathrm{Alt}(\nu) = \bigcup_{i \neq i^\star} \{\kappa \in \mathcal{D}^K \mid m(\kappa_i) \geq m(\kappa_{i^\star})\}$. Leveraging the independence between the means (*i.e.* $\mathcal{S} = \mathcal{I}^K$), we have

$$\begin{aligned}\inf_{\kappa \in \mathrm{Alt}(\nu)} \sum_{i \in [K]} w_i \mathrm{KL}(\nu_i, \kappa_i) &= \min_{i \neq i^\star} \inf_{\kappa \in \mathcal{D}^2, m(\kappa_i) \geq m(\kappa_{i^\star})} \{w_{i^\star} \mathrm{KL}(\nu_{i^\star}, \kappa_{i^\star}) + w_i \mathrm{KL}(\nu_i, \kappa_i)\} \\ &= \min_{i \neq i^\star} \inf_{u \geq v} \left\{ w_{i^\star} \mathcal{K}^-_{\inf}(\nu_{i^\star}, v) + w_i \mathcal{K}^+_{\inf}(\nu_i, u) \right\} ,\end{aligned}$$

The last equality is obtained by re-parametrizing with means and taking the infimum over distributions satisfying the mean constraint, hence it involves the function $\mathcal{K}_{\inf}$ which is defined as an infimum over KL divergence

$$\begin{aligned} \mathcal{K}^+_{\inf}(\nu, u) &:= \inf\{\mathrm{KL}(\nu, \kappa) \mid \kappa \in \mathcal{D},\ m(\kappa) > u\} \\ \text{and} \quad \mathcal{K}^-_{\inf}(\nu, u) &:= \inf\{\mathrm{KL}(\nu, \kappa) \mid \kappa \in \mathcal{D},\ m(\kappa) < u\} . \end{aligned} \tag{1.8}$$

For one-parameter exponential families, we have $\mathcal{K}^+_{\inf}(\nu, u) = d_{\mathrm{KL}}(m(\nu), \max\{m(\nu), u\})$ and $\mathcal{K}^-_{\inf}(\nu, u) = d_{\mathrm{KL}}(m(\nu), \min\{m(\nu), u\})$. Depending on the class of distribution considered and potentially additional constraints, $\mathcal{K}_{\inf}$ enjoys useful properties such as strong duality, monotonicity and strict convexity with respect to the second argument. A detailed description of the assumptions under which those properties hold is beyond the scope of this manuscript. Nonetheless, the reader can be reassured that they hold for all the distributions considered in Part I. Using monotonicity properties of $\mathcal{K}_{\inf}$, we have

$$\inf_{\kappa \in \mathrm{Alt}(\nu)} \sum_{i \in [K]} w_i \mathrm{KL}(\nu_i, \kappa_i) = \min_{i \neq i^\star} \inf_{u \in \mathcal{I}} \left\{ w_{i^\star} \mathcal{K}^-_{\inf}(\nu_{i^\star}, u) + w_i \mathcal{K}^+_{\inf}(\nu_i, u) \right\} ,$$

where the infimum over $\mathcal{I}$ could be restricted to $[\mu_i, \mu_{i^\star}]$. Let

$$C(i, j; \nu, w) = \mathbb{1}\left(\mu_i > \mu_j\right) \inf_{u \in \mathcal{I}} \left\{ w_i \mathcal{K}^-_{\inf}(\nu_i, u) + w_j \mathcal{K}^+_{\inf}(\nu_j, u) \right\} \tag{1.9}$$

be the *transportation cost* between answer i and answer j with respect to the allocation w for the bandit instance ν . Intuitively, $C(i, j; \nu, w)$ represents how far ν is from a distribution where arm i has a higher mean than arm j given allocation w . Putting things together, we have that

$$T^\star(\nu)^{-1} = \sup_{w \in \Sigma_K} \min_{i \neq i^\star} C(i^\star, i; \nu, w) .$$

Depending on the class of distribution considered, the characteristic times satisfy strong regularity properties that are useful to study algorithms inspired by this lower bound, *e.g.* see Lemma 2.11. Interestingly, the optimization problem defining $T^\star(\nu)^{-1}$ can be rewritten as a simpler optimization problem which can be approximately solved with nested binary searches.

When there is additional structure (*e.g.* linear bandit), $C(i, j; \nu, w)$ will depend on (ν, w) as a whole instead of simply depending on (μ_i, μ_j, w_i, w_j) . Moreover, it will not be possible to use the quantities $\mathcal{K}_{\inf}^{\pm}$ since the structural dependence between the distributions prevents to optimize the distributions at the level of the arms.

When there are multiple correct answers, the transportation cost should be also adapted to the question of interest. For example, in vanilla ε-BAI, we have

$$T_\varepsilon(\nu)^{-1} = \max_{i \in \mathcal{I}_\varepsilon(\mu)} \sup_{w \in \Sigma_K} \min_{j \neq i} C_\varepsilon(i, j; \nu, w) \, ,$$

where $C_\varepsilon(i, j; \nu, w) = \mathbb{1}\left(\mu_i > \mu_j - \varepsilon\right) \inf_{u \in \mathcal{I}} \left\{ w_i \mathcal{K}_{\inf}^{-}(\nu_i, u) + w_j \mathcal{K}_{\inf}^{+}(\nu_j, u - \varepsilon) \right\}$.

1.4.2 The GLR Stopping Rule

Before designing a sampling rule that matches the asymptotic lower bound presented in Section 1.4.1, one should specify a stopping rule. Given observations collected before time n , an algorithm can stop as soon as it has collected enough statistical evidence that its recommendation is a correct answer with probability at least $1 - \delta$. This problem can be seen as an adaptive hypothesis testing problem regardless of the sampling rule considered. The sampling rule collects data sequentially by adapting to past observations. Sequential hypothesis testing refers to the setting where the sampling rule follows a sequential allocation fixed beforehand. It has been studied for several decades, *e.g.* Wald [1945] for two simple hypothesis, Robbins and Siegmund [1974] and Lai [1988] for composite hypothesis.

Estimator Let ν_n be the empirical distribution associated with the empirical estimator of μ_n which maximizes the likelihood, *i.e.* $\nu_n \in \arg\max_{\kappa \in \mathcal{D}^K,\, m(\kappa) \in \mathcal{S}} \ell(X_{1,I_1}, \cdots, X_{n-1,I_{n-1}}; \kappa)$. We have $\mu_{n,i} = N_{n,i}^{-1} \sum_{t \in [n-1]} \mathbb{1}\left(I_t = i\right) X_{t,i}$ for all $i \in [K]$ when there is no structure. For a parametric distribution, ν_n is the parametric distribution associated with the maximum likelihood estimator (MLE) of the parameters uniquely characterizing the distribution. For non-parametric distribution (and without structure), ν_n is the vector of empirical distributions of the arms, *i.e.* $\nu_{n,i} = N_{n,i}^{-1} \sum_{t \in [n-1]} \mathbb{1}\left(I_t = i\right) \delta_{X_{t,i}}$ where δ_x is the Dirac mass at x . While we have $\nu_n \in \mathcal{D}^K$ when $\mathcal{D} = \mathcal{D}_{[0,B]}$, there are classes of non-parametric distributions such that ν_n does not even belong to $\mathcal{D}^K$.

Parallel tests We use the generalized likelihood ratio (GLR) stopping rule, as done extensively in the literature. The GLR stopping rule was first proposed in Garivier and Kaufmann [2016], then popularized by subsequent works [Kaufmann, 2020]. The idea is to run Z sequential tests in parallel. Given an answer $i \in \mathcal{Z}$, we consider the GLR test for the following two non-overlapping hypotheses $\mathcal{H}_{0,i} : (\mu \in \neg i)$ against $\mathcal{H}_{1,i} : (\mu \in \mathcal{S} \setminus \neg i)$, and

denote the GLR statistic of answer i at time n by $\mathrm{GLR}_n(\neg i) = \inf_{\kappa \in \mathcal{D}^K,\, m(\kappa) \in \neg i} \mathcal{L}_n(\nu_n, \kappa)$ where $\mathcal{L}_n(\nu, \kappa)$ as in (1.2). Importantly, a high value of $\mathrm{GLR}_n(\neg i)$ indicates that we should reject $\mathcal{H}_{0,i}$. Let $i_F(\nu_n, N_n)$ be the set of *instantaneous furthest* answers which maximizes the GLR, *i.e.* $i_F(\nu_n, N_n) = \arg\max_{i \in \mathcal{Z}} \mathrm{GLR}_n(\neg i)$. When we need to recommend answer at any time n , it is natural to recommend $\hat{\imath}_n \in i_F(\nu_n, N_n)$ since this is the answer for which we have collected the most evidence that it is correct. The GLR stopping rule stops as soon as one of these tests can reject the null hypothesis. In other words, it stops as soon as the GLR statistic exceeds a given stopping threshold $c : \mathbb{N} \times (0, 1) \to \mathbb{R}_+$,

$$\tau_\delta = \inf \left\{ n \mid \max_{i \in \mathcal{Z}} \mathrm{GLR}_n(\neg i) > c(n-1, \delta) \right\} , \tag{1.10}$$

and recommend $\hat{\imath}_{\tau_\delta} \in \arg\max_{i \in \mathcal{Z}} \mathrm{GLR}_{\tau_\delta}(\neg i)$ since μ is believed to admit $\hat{\imath}_{\tau_\delta}$ as a correct answer. When there are multiple correct answers, other tests could have rejected the null hypothesis if provided with more samples, *i.e.* several answers could satisfy that $\mathrm{GLR}_n(\neg i) > c(n-1, \delta)$ if we don't stop.

Regardless of the sampling rule, the stopping threshold $c(n, \delta)$ is chosen to ensure δ-correct by using time-uniform concentration results. Namely, it is such that the probability that there exists a time n such that $\mathrm{GLR}_n(\neg \hat{\imath}_n) \geq c(n-1, \delta)$ and $\hat{\imath}_n$ is not correct is upper bounded by δ . While its exact formula depends on the considered class of distribution, the δ-dependency should be such that $\lim_{\delta \to 0} c(n, \delta)/\log(1/\delta) = 1$, otherwise the algorithm is prohibited from being asymptotically optimal. The asymptotic behavior as regards the n-dependency depends on the considered class of distributions. While $c(n, \delta) =_{n \to +\infty} \mathcal{O}(\log \log n)$ for the parametric distributions considered in this thesis, the known thresholds for non-parametric distributions only satisfy $c(n, \delta) =_{n \to +\infty} \mathcal{O}(\log n)$. It is still an open problem to know whether it is possible to achieve a scaling in $\mathcal{O}(\log \log n)$ for non-parametric distributions. Empirically, the known stopping thresholds have been observed to be conservative empirically since the empirical error rates are orders of magnitude below the tolerated error of δ . Avoiding this bottleneck on the expected sample complexity is also an open problem, *e.g.* one could use a tighter stopping threshold or a different stopping rule.

Contribution 1.5. *In Section 3.3 of Chapter 3, we present several ways to define a threshold for Gaussian with unknown variance, and derive one which scales as $\mathcal{O}(\log \log n)$ when $n \to +\infty$. Lemma 4.3 in Chapter 4, we propose a threshold for bounded distributions which scales as $\mathcal{O}(\log n)$ when $n \to +\infty$. In both cases, our thresholds ensure δ-correctness for the corresponding class of distributions and satisfies that $\lim_{\delta \to 0} c(n, \delta)/\log(1/\delta) = 1$, hence reaching asymptotic optimality.*

When the stopping condition is not met, the sampling rule will return an arm I_n to be pulled next and we will update our belief based on this new observation. While examples

of sampling rule will be detailed in Section 1.4.3, we understand what a good sampling rule should do. Since we want to stop as soon as possible, we should gather more evidence to verify that $\hat{\imath}_n \in i_F(\nu_n, N_n)$ is a correct answer. Since this is the answer for which we have the most evidence that it is correct, it is likely to be the easiest to verify with additional observations.

Contribution 1.6. *In Chapter 7, the easiest-to-verify candidate answer is fundamental to reach asymptotic optimality. For ε-BAI in transductive linear bandits, we show how to use this concept in the GLR stopping rule (Section 7.3), as well as in the sampling rule (Section 7.4).*

Illustrative example: BAI The GLR statistic of answer i at time n can be written as

$$\mathrm{GLR}_n(\neg i) = \min_{j \neq i} W_n(i,j) \quad \text{with} \quad W_n(i,j) = C(i,j;\nu_n, N_n)\,,$$

where $W_n(i,j)$ denotes the empirical transportation cost between answer i and answer j . It represents the amount of evidence we have collected so far in order to reject the hypothesis that arm j has a higher mean than arm i . Moreover, we have $i_F(\nu_n, N_n) = i^\star(\mu_n)$, hence we are recommending the empirical best (EB) arm $\hat{\imath}_n \in i^\star(\mu_n)$.

To better understand how to choose $c(n,\delta)$, one can rewrite the error event as

$$\begin{aligned}\{\tau_\delta < +\infty\} \cap \mathcal{E}_\mu^{\mathrm{err}}(\tau_\delta) &= \bigcup_{n\in\mathbb{N}} \bigcup_{i \neq i^\star} \{i \in i^\star(\mu_n),\ \min_{j\neq i} C(i,j;\nu_n,N_n) > c(n-1,\delta)\} \\ &\subseteq \bigcup_{n\in\mathbb{N}} \bigcup_{i \neq i^\star} \{N_{n,i}\mathcal{K}_{\inf}^-(\nu_{n,i},\mu_i) + N_{n,i^\star}\mathcal{K}_{\inf}^+(\nu_{n,i^\star},\mu_{i^\star}) > c(n-1,\delta)\}\,.\end{aligned}$$

Therefore, the stopping threshold is chosen to control a reweighted random deviation between the empirical distributions ν_n of the arms and their true mean μ .

Both for additional structure (*e.g.* linear bandit) and multiple correct answers, the story is similar as in Section 1.4.1 since $W_n(i,j)$ is an empirical version of the transportation cost $C(i,j;\nu,w)$. For example, in vanilla ε-BAI, we have

$$\mathrm{GLR}_n(\neg_\varepsilon i) = \min_{j\neq i} W_{\varepsilon,n}(i,j) \quad \text{with} \quad W_{\varepsilon,n}(i,j) = C_\varepsilon(i,j;\nu_n,N_n)\,,$$

hence $i_F(\nu,w) = \arg\max_{i\in\mathcal{I}_\varepsilon(\mu)} C_\varepsilon(i,j;\nu,w)$.

Influence of the sampling Rule Importantly, one can notice that the GLR statistic is an empirical version of the inverse of the characteristic time $T^\star(\nu)$, where ν is replaced by ν_n and

the outer optimization over Σ_K is replaced by an evaluation at N_n , *i.e.*

$$T^\star(\nu)^{-1} = \max_{w\in\Sigma_K} \min_{j\neq i^\star} C(i^\star, j; \nu, w) \quad \text{and} \quad \mathrm{GLR}_n(\neg i) = \min_{j\neq i} C(i, j; \nu_n, N_n) \, .$$

Let $w \in \mathring{\Sigma}_K$ where $\mathring{X}$ denotes the interior of the set X . This allows to understand what will be the asymptotic behavior of an algorithm whose empirical allocation is close to w asymptotically, *i.e.* $\|N_n/n - w\|_\infty \approx_{n\to+\infty} 0$ for n large enough. For a minute, let's assume we have access to such a sampling rule. Then, for n large enough such that $n < \tau_\delta$, we will have $\hat{\imath}_n = i^\star$ and

$$n \min_{i\neq i^\star} C(i^\star, i; \nu, w) + o(n) \underset{n\to+\infty}{=} \min_{j\neq i} W_n(i, j) \leq c(n, \delta) \underset{\delta\to 0, n\to+\infty}{=} \log(1/\delta) + o(n + \log(1/\delta)) \, .$$

Therefore, we now have a sufficient property for a sampling rule to be asymptotically optimal: the empirical allocation should be close to $w^\star(\nu)$, *i.e.* $\inf_{w\in w^\star(\nu)} \|N_n/n - w\|_\infty \approx 0$ for n large enough. For such a sampling rule, we obtain $\tau_\delta \lesssim T^\star(\nu)\log(1/\delta)$ which will yield asymptotic optimality. This argument is formalized in Section 2.3.2.

Other stopping rules While (1.10) is the most common form of GLR stopping rule, one can also leverage the fact that solving ε-BAI can be done by comparing pairs of arms, *i.e.*

$$\tau_{\varepsilon,\delta} = \inf\left\{n \mid \exists i \in [K], \forall j \in [K] \setminus \{i\},\ W_{\varepsilon,n}(i, j) > c_{i,j}(N_n, \delta)\right\} \, , \tag{1.11}$$

where $c_{i,j} : \mathbb{N}^K \times (0, 1) \to \mathbb{R}_+$ is a stopping threshold for the pair of arms (i, j) . In Chapter 3, we use this expression of the GLR stopping rule. While this form of stopping time is often provably smaller than (1.10), our experiments suggest that the expected sample complexities are similar when paired with a good sampling rule.

While the GLR stopping rule has been extensively used for fixed-confidence pure exploration problems in recent years, it is not yet the most used stopping rule. The two most studied alternatives stopping rule are the elimination-based one and the confidence-based one. The elimination-based stopping rule stops when there is only one active answer left [Even-Dar et al., 2006, Karnin et al., 2013]. The set of active answers is shrinking at the end of each phase (whose length is increasing), and answers which have poor empirical mean compared to the empirical best one (up to) are eliminated. The confidence-based stopping rule used stops when the LCB (Lower Confidence Bound) of the candidate answer is not ε worse the UCB (Upper Confidence Bound) of all the alternative answers. For Gaussian distributions with known variance, the main difference between the confidence-based stopping rule and the GLR stopping rule lies in the fact the concentration results which are used. While later considers mean gaps directly, the later relies on per-arm concentration, hence will be sub-optimal in the δ-dependency (*e.g.* see Chapter 3). It is actually easy to show that the GLR stopping rule is

equivalent to a confidence-based stopping rule when the same concentration results on the mean gap is done.

1.4.3 Lower Bound Based Algorithms

The sampling rule is the last component to be defined in order to fully specify a fixed-confidence pure exploration strategy. Since the GLR stopping rule ensures δ-correctness of the algorithm, the sampling rule should be designed to stop as soon as possible, *i.e.* minimize the expected sample complexity. For the sake of space, we will only present three types of sampling rules which are all inspired by the lower bound: Track-and-Stop, an online optimization based approach and the Top Two approach. While those approaches have been generalized to other settings, they have been introduced to tackle vanilla BAI for parametric distributions, as we will do below.

While all those algorithms aims at obtaining an empirical allocation which is close to $w^\star(\nu)$, they differ in the considered strategies. The Track-and-Stop algorithm has a high-level approach which computes $w^\star(\nu_n)$ at each time n , then forces the empirical counts to be close to $\sum_{t\in[n]} w^\star(\nu_t)$ by using tracking. The online optimization based algorithms view $T^\star(\nu)^{-1}$ as an optimization problem, and they learn $w^\star(\nu)$ sequentially with online optimization algorithms. A Top Two sampling rule for bandit identification is a method which selects the next arm to sample from among two candidate arms, a leader and a challenger. In light of our recent works, the Top Two approach can be seen as a low-level approach which aims at increasing $\mathrm{GLR}_n(\neg\hat{\imath}_n)$ the most. Historically, it is interesting to notice that the first papers on Track-and-Stop and the Top Two approach were both published at the 29^{th} Annual Conference on Learning Theory (COLT 2016).

Track-and-Stop

The Track-and-Stop approach [Garivier and Kaufmann, 2016] appeared as an algorithmic by-product of the theoretical lower bound on the expected sample complexity of any δ-correct algorithm.

At each time n , Track-and-Stop solves the optimization problem defining the characteristic time of the empirical estimator ν_n of the distributions, *i.e.* it computes $w_n = w^\star(\nu_n)$. Given the vector $w_n \in \Sigma_K$, it uses a so-called *tracking* procedure to obtain an arm I_n to sample. We describe and use the one called C-Tracking by Garivier and Kaufmann [2016]. On top of this tracking a forced exploration is used to enforce that all arms are sampled. This is done here by projecting on $\Sigma_K^\varepsilon = \{w \in [\varepsilon, 1]^K \mid \sum_{i\in[K]} w_i = 1\}$ for a well chosen $\varepsilon \in (0, 1/K]$. Defining $w_n^{\varepsilon_n}$ the ℓ_∞ projection of w_n on $\Sigma_K^{\varepsilon_n}$ with $\varepsilon_n = (K^2+n)^{-1/2}/2$, C-Tracking pulls $I_n \in \arg\max_{i\in[K]}\{\sum_{t\in[n]} w_{t,i}^{\varepsilon_t} - N_{n,i}\}$.

The forced exploration ensures that each arm is sampled enough for ν_n to converge towards ν . Using arguments of continuity, we obtain that w_n converge towards $w^\star(\nu)$, hence N_n/n does too by tracking properties. Therefore, as detailed in Section 1.4.2, the Track-and-Stop algorithm can be shown to be asymptotically optimal when combined with the GLR stopping rule.

Since it requires to solve a difficult optimization problem, the computational cost of Track-and-Stop algorithm is several orders of magnitude larger than previous algorithms, *e.g.* LUCB which is based on confidence intervals [Kalyanakrishnan et al., 2012]. Finding alternative approaches to Track-and-Stop that are asymptotically optimal without the need to compute the optimal proportion in every round was an active line of research which has been addressed in recent years with the Top Two approach and the game based approach (among others).

Contribution 1.7. *In Section 3.4 of Chapter 3, we study the Track-and-Stop approach in vanilla bandits for Gaussian distributions with unknown variance, and show its asymptotic optimality.*

Variants of Track-and-Stop have also been analyzed. The EBS (Exploration-Biased Weights) algorithm [Barrier et al., 2022] computes the optimal allocation for a modified mean parameter. The intuition is to wrap the optimal weight vector from above, by ensuring that its minimal value is never underestimated.

Online Optimization Based Approach

The online optimization approach aims at solving the lower bound with generic optimization techniques, *i.e.*

$$\sup_{w\in\Sigma_K} F(w,\nu) \quad \text{with} \quad F(w,\nu) = \inf_{\kappa\in \text{Alt}(\nu)} \sum_{i\in[K]} w_i \text{KL}(\nu_i,\kappa_i)\,,$$

where $w \to F(w,\nu)$ is concave (but not necessarily smooth) and admits sub-gradients. Ménard [2019] replaced the oracle call from Track-and-Stop by a one step of lazy mirror descent in each round of the algorithm. Then, Ménard [2019] uses a tracking procedure which mixes the target allocation with the uniform distribution to ensure forced exploration. Instead of using lazy mirror descent, the FWS (Frank-Wolfe Sampling) algorithm [Wang et al., 2021] uses an adapted Frank-Wolfe step at each round to update the target to be tracked, and adds forced exploration.

The game based approach [Degenne et al., 2019] came from the interpretation of the lower bound as the solution of a two-players game. The quantity

$$\sup_{w \in \Sigma_K} \inf_{\kappa \in \text{Alt}(\nu)} \sum_{i \in [K]} w_i \text{KL}(\nu_i, \kappa_i)$$

is viewed as the value of a min-max game between the agent that chooses arms based on an allocation $w \in \Sigma_K$ and the *nature* that plays an alternative vector of distributions $\kappa \in \text{Alt}(\nu)$. The game based approach plays two no-regret learning algorithms against each other, and uses optimism on the payoffs to remove the need for forced exploration. Since the resulting saddle-point algorithm approximates $T^\star(\nu)^{-1}$, it can be shown to be asymptotically optimal.

Contribution 1.8. *In Section 7.4 of Chapter 7, we propose a LεBAI, which is a game based algorithm, for ε-BAI in transductive linear bandits with Gaussian distributions. LεBAI is is asymptotically optimal and has competitive empirical performance.*

Top Two Approach

The Top Two approach [Russo, 2016] arose as an adaptation of the Thompson Sampling algorithm for regret minimization [Thompson, 1933] to best arm identification in multi-armed bandit models, for parametric families of arms. The Top Two approach is the main topic of this thesis, and many papers have been published on those algorithms since this thesis started. Therefore, we only give a historical picture here and provide extensive details in Chapter 2.

At each time n , a Top Two algorithm defines two candidate answers, a *leader* and a *challenger*, and sample the next arm among those two arms in order to verify that the leader is a better answer than the challenger. As we will see in Chapter 2, many choices are possible when defining a Top Two algorithm. Russo [2016] introduced Top Two Probability Sampling (TTPS) and Top Two Thompson Sampling (TTTS). TTTS follows a simple idea: as vanilla Thompson Sampling (TS) selects the optimal arm too much, with some probability $1 - \beta$, TTTS forces itself to select an arm which is not the one selected by TS, by re-sampling the posterior until another arm has the largest posterior sample. Adopting a Bayesian viewpoint, Russo studied the convergence rate of the posterior probability that $i^\star$ is not the best arm, under some conditions on the prior.

For Gaussian bandits, other Bayesian Top Two algorithms with frequentist components have been shown to be asymptotically β-optimal: Top Two Expected Improvement (TTEI [Qin et al., 2017]) and Top Two Transportation Cost (T3C [Shang et al., 2020]). At the beginning of this thesis, there were many open research directions as regards the Top Two approach.

Some answers are given in this thesis, others have been given by concurrent works. Notably, an adaptive choice of β to achieve asymptotic optimality was analyzed in You et al. [2023].

In Section 8.3 of Chapter 8, we highlight that Top Two algorithms can actually be seen as saddle-point algorithms. Since they also aim at solving this min-max game between the agent and the nature, they bare similarities with the game based algorithm.

Contribution 1.9. *In Part I, we present the Top Two approach for BAI in vanilla bandits for Gaussian distributions with known variance in Chapter 2, Gaussian distributions with unknown variance in Chapter 3 and bounded distributions in Chapter 4. In Chapter 2, we propose a generic definition of a Top Two algorithm which is specified by four choices and propose several instances. We provide a unified asymptotic analysis of the Top Two approach, which identifies desirable properties on the choices of the leader and challenger answers to achieve asymptotic (β-)optimality. Then, we give the first non-asymptotic analysis of a Top Two algorithm. In Chapter 5, we present the Top Two approach for ε-BAI in vanilla bandits. We propose the EB-TC$_\varepsilon$ algorithm which has near optimal asymptotic and non-asymptotic guarantees on its expected sample complexity. In Chapter 8, we present the Top Two approach for ε-BAI in transductive linear bandits, and propose LεTT as an extension of EB-TC$_\varepsilon$ which enjoys competitive empirical performance.*

Other Sampling Rules

While the algorithm described above aim at being asymptotically optimal, other approaches were designed to obtain non-asymptotic guarantees (*e.g.* LUCB [Kalyanakrishnan et al., 2012], UGapEc [Gabillon et al., 2012], KL-LUCB [Kaufmann and Kalyanakrishnan, 2013], Exp-Gap [Karnin et al., 2013], lil'UCB [Jamieson et al., 2014], etc). The first BAI algorithms were introduced and studied under the assumption that the observation have bounded support, with a known upper bound [Even-Dar et al., 2006, Kalyanakrishnan et al., 2012, Gabillon et al., 2012, Jamieson et al., 2014]. The sample complexity bounds proved for these algorithms scale as the sum of squared inverse gap, *i.e.*

$$H(\mu) := 2\Delta_{\min}^{-2} + \sum_{i \neq i^\star} 2(\mu_{i^\star} - \mu_i)^{-2} \quad \text{with} \quad \Delta_i := (\mu_{i^\star} - \mu_i) \quad \text{and} \quad \Delta_{\min} := \min_{i \neq i^\star} \Delta_i \,. \tag{1.12}$$

It satisfies $H(\mu) \leq T^\star(\nu) \leq 2H(\mu)$ where $T^\star(\nu)$ is the characteristic time for Gaussian distributions with unit variance [Garivier and Kaufmann, 2016]. The usual non-asymptotic guarantees which are obtained are of the order of $\mathcal{O}\left(H(\mu)\log(H(\mu)/\delta)\right)$. Unfortunately, the $\mathcal{O}(\cdot)$ notation often hides the large and non-explicit constants, except for LUCB [Kalyanakrishnan et al., 2012] which satisfies $\mathbb{E}_\nu[\tau_\delta] \leq 292 H(\mu)\log(H(\mu)/\delta) + 16$.

1.5 Beyond the Fixed-confidence Setting

Even though the fixed-confidence setting (see Section 1.4) is theoretically appealing as it is well understood, the budget is often limited. In the fixed-budget setting (see Section 1.5.1), the maximal budget is assumed to be known beforehand. While both the fixed-confidence and fixed-budget settings have been studied extensively, they rarely coincides with concrete experimental setups. For example, they do not cope for early stopping and continuation (see Section 1.3). Therefore, one should strive to go beyond those two major theoretical frameworks, and consider a more practical one.

Regardless of the objective of the agent, a good identification strategy should ideally come with guarantees on its current candidate answer that hold at any time. This is exactly the promise of the anytime setting (see Section 1.5.2) which solves the limitation of both the fixed-confidence and the fixed-budget settings. Such an anytime identification strategy is broadcasting (or producing) a stream of "good" recommendations $(\hat{\imath}_n)_{n\in\mathbb{N}}$.

We highlight here that, if made available to others, this stream could be used in parallel by several actors each one having a different objective. Some actors might have different budgets $\{T_k\}_{k\in[N]}$ or error parameters $\{\delta_k\}_{k\in[N]}$, other might want to answer a similar yet different question or simply maximize their cumulative gains. Taking a step back, it is actually not uncommon for an agent to internalize the sampling cost of an identification procedure, and then to broadcast its recommendations. For example, a non-profit (or governmental) organization will provide recommendations (*e.g.* on health or education) to the population for "free" since the cost is paid with donation (or taxes). For private companies, recommendations (*e.g.* latest fine-tuned parameters for a large language model) might be sold to customers or other companies in exchange of a fee or a subscription. Sometimes those recommendations are given for "free" since the customers are targeted to sell their data (*e.g.* queries to a large language model) or become client of subsidiary brands.

1.5.1 Fixed-budget Setting

In the fixed-budget setting, the agent is given a budget $T \in \mathbb{N}$. The stopping rule is deterministic since we stop after having collected T samples, and the recommendation rule needs only to be defined at time T . A fixed-budget identification strategy is judged based on its probability of error $\mathbb{P}_\nu(\mathcal{E}_\mu^{\mathrm{err}}(T))$ at time T , *i.e.* the probability that the candidate answer $\hat{\imath}_T$ at time T is not a correct answer. A "good" fixed-budget algorithm should minimize $\mathbb{P}_\nu(\mathcal{E}_\mu^{\mathrm{err}}(T))$.

Algorithms whose Error Decays Exponentially

As in Section 1.4.3, we will restrict ourselves to vanilla BAI, hence $\mathcal{E}_\mu^{\mathrm{err}}(T) = \{\hat{\imath}_T \notin i^\star(\mu)\}$. Moreover, we use the set $\mathcal{D}_{[0,1]}$ of bounded distributions to fix the ideas. The algorithmic ideas described below have been generalized to other settings.

Uniform sampling rule Before introducing fixed-budget algorithms that leverage the knowledge of T , one can first understand the performance of the uniform sampling rule. It pulls arms in a round-robin fashion and recommends the empirical best arm, *i.e.* $\hat{\imath}_n \in i^\star(\mu_n)$. Using Hoeffding's inequality, for all $n \in \mathbb{N}$,

$$\begin{aligned}\mathbb{P}_\nu(\hat{\imath}_n \notin i^\star(\mu)) \leq \sum_{i \notin i^\star(\mu)} \mathbb{P}_\nu\left(\mu_{n,i} \geq \mu_{n,i^\star}\right) &\leq \sum_{i \notin i^\star(\mu)} \exp\left(-\Delta_i^2 \left\lfloor \frac{n}{K} \right\rfloor\right) \\ &\leq (K - |i^\star(\mu)|) \exp\left(-\frac{n-K}{K\Delta_{\min}(\mu)^{-2}}\right) ,\end{aligned}$$

with $\Delta_i := \mu_\star - \mu_i$ is the gap of arm $i \notin i^\star(\mu)$ and $\Delta_i = \Delta_{\min}(\mu)$ for all $i \in i^\star(\mu)$ where $\Delta_{\min}(\mu) := \min_{i \notin i^\star(\mu)} \Delta_i$ is the smallest strictly positive gap. Asymptotically, we obtain that

$$\limsup_{n \to +\infty} \frac{n}{-\log \mathbb{P}_\nu(\hat{\imath}_n \notin i^\star(\mu))} \leq \frac{K}{\Delta_{\min}(\mu)^2} ,$$

which is referred to as the *asymptotic rate* of the exponential decay.

As for the fixed-confidence setting, we will see that better fixed-budget algorithms can be designed. Note that the uniform algorithm is the canonical example of an anytime algorithm, *i.e.* independent of any parameter (δ or T) and has guarantees at any time.

Elimination based The most popular algorithmic approach to fixed-budget BAI is the elimination based one. It has a round-based structure, and it recommends the last active answer. An algorithm is then defined by its number of rounds and, for each round, the number of active answers and the fixed allocation to the arms. The two most famous instances are Successive Reject (SR [Audibert et al., 2010]) and Sequential Halving (SH [Karnin et al., 2013]).

SR uses $K-1$ phases. Let $\overline{\log} K = 1/2 + \sum_{i \in [K]} 1/i$ and $n_k = \lceil \frac{T-K}{(K+1-k)\overline{\log} K} \rceil$ with $n_0 = 0$. During phase k , it samples all the remaining arms $n_k - n_{k-1}$ times, and rejects one answer, the one with the worst empirical mean. Audibert et al. [2010] proves that

$$\mathbb{P}_\nu(\hat{\imath}_T \notin i^\star(\mu)) \leq \frac{K(K-1)}{2} \exp\left(-\frac{T-K}{H_2(\mu)\overline{\log} K}\right) ,$$

where $H_2(\mu) = \max_{k\in[K]} i\Delta_{(k)}^{-2}$ where (k) denotes the index of the arm with the k^{th} highest mean. Therefore, a sequence of SR algorithms would achieve $H_2(\mu)\overline{\log}K$ as an asymptotic rate. Recently, Wang et al. [2024] provided an improved asymptotic rate for SR by using the large deviation principle (asymptotic arguments). Interestingly, they introduce and analyze the CR (continuous reject) algorithm which can adaptively discard an arm at any time.

In contrast, SH only uses $\log_2 K$ phases since it rejects half of the remaining answers, the ones with the worst empirical mean. During phase k , it samples all the remaining arms in A_k a number $\left\lfloor \frac{T}{|A_k|\lceil\log_2 K\rceil} \right\rfloor$ of times. Note that SH drops the observations collected in the previous phases. Karnin et al. [2013] proves that

$$\mathbb{P}_\nu(\hat{\imath}_T \notin i^\star(\mu)) \leq 3\log_2(K)\exp\left(-\frac{T}{8H_2(\mu)\log_2 K}\right) .$$

Therefore, a sequence of SH algorithms would achieve $8H_2(\mu)\log_2 K$ as an asymptotic rate. In recent years, improvements have been made on the analysis of SH, *e.g.* better rate Zhao et al. [2023] and allowing to keep past observations Kone et al. [2024]. Both SR and SH have an asymptotic rate which is better than the one achieved by uniform sampling on most instances.

Prior knowledge based Several fixed-budget BAI algorithms assume that the agent has access to some prior knowledge on unknown quantities to design upper/lower confidence bounds (UCB/LCB), *e.g.* UCB-E [Audibert et al., 2010] and UGapEb [Gabillon et al., 2012]. While this assumption is often not realistic, it yields better guarantees. For example, UCB-E uses the knowledge of $H(\mu)$ to set its exploration parameter to $a = \frac{25}{36}\frac{T-K}{H(\mu)}$ in order to achieve

$$\mathbb{P}_\nu(\hat{\imath}_T \notin i^\star(\mu)) \leq 2TK\exp\left(-\frac{T-K}{9H(\mu)}\right) .$$

where $H(\mu) = \sum_{i\in[K]} 2\Delta_i^{-2}$. Therefore, a sequence of UCB-E algorithms would achieve $18H(\mu)$ as an asymptotic rate. Compared to SR and SH, the $\log K$ term has been shaved off. In Section 1.5.1, we will see that this dependency is necessary without assuming prior knowledge.

Lower Bound on the Probability of Error

As for the fixed-confidence setting, it is interesting to understand what are the limits on the probability of error. Since lower bounds are only defined with respect to a given class of algorithms, we will assume that the sequence of algorithms $(\mathfrak{A}_T)_{T\in\mathbb{N}}$ is consistent, *i.e.* $\lim_{T\to+\infty}\mathbb{P}_\nu(\hat{\imath}_T \notin i^\star(\mu)) = 0$ for all $\nu \in \mathcal{D}^K$.

Let $\mathcal{E}_T = \mathcal{E}_\mu^{\mathrm{err}}(T)$ and $\kappa \in \mathrm{Alt}(\nu)$. Using consistency, we know that $\mathbb{P}_\kappa(\mathcal{E}_T) \to_{T\to+\infty} 1$. As in Section 1.4.1, we can use the data-processing inequality to show that

$$\sum_{i\in[K]} \mathbb{E}_\kappa[N_{T,i}]\mathrm{KL}(\kappa_i,\nu_i) \geq \mathrm{kl}(\mathbb{P}_\kappa(\mathcal{E}_T),\mathbb{P}_\nu(\mathcal{E}_T)) \geq -\mathbb{P}_\kappa(\mathcal{E}_T)\log\mathbb{P}_\nu(\mathcal{E}_T) - \log 2\,,$$

where we used that $\mathrm{kl}(x,y) \geq -x\log y - \log 2$. By re-ordering, taking the limit on T and the infimum on κ , we obtain

$$\liminf_{T\to+\infty} \frac{T}{-\log\mathbb{P}_\nu(\hat{\imath}_T \notin i^\star(\mu))} \geq \left(\inf_{\kappa\in\mathrm{Alt}(\nu)} \limsup_{T\to+\infty} \sum_{i\in[K]} \frac{\mathbb{E}_\kappa[N_{T,i}]}{T}\mathrm{KL}(\kappa_i,\nu_i)\right)^{-1}.$$

While $\mathbb{E}_\kappa[N_T]/T \in \Sigma_K$, it depends both on the alternative κ and on the budget T , hence we cannot take the supremum over Σ_K . Let $w \in \mathring{\Sigma}_K$. Considering the class of algorithms which have static proportion w asymptotically, *i.e.* $\lim_{T\to+\infty}\mathbb{E}_\kappa[N_{T,i}]/T = w_i$ for all κ , we obtain

$$\liminf_{T\to+\infty} \frac{T}{-\log\mathbb{P}_\nu(\hat{\imath}_T \notin i^\star(\mu))} \geq H_{\mathcal{C}^{sp}}(\nu,w) \quad \text{with} \quad H_{\mathcal{C}^{sp}}(\nu,w)^{-1} = \inf_{\kappa\in\mathrm{Alt}(\nu)} \sum_{i\in[K]} w_i\mathrm{KL}(\kappa_i,\nu_i)\,.$$

Let us define $H_{\mathcal{C}^{sp}}(\nu) = \min_{w\in\mathring{\Sigma}_K} H_{\mathcal{C}^{sp}}(\nu,w)$. Therefore, the best static allocation for a given instance ν seems to be $H_{\mathcal{C}^{sp}}(\nu)$. The difference between $H_{\mathcal{C}^{sp}}(\nu)$ and $T^\star(\nu)$ lies in the fact that the arguments of the KL divergence are swapped, *i.e.* $\mathrm{KL}(\kappa_i,\nu_i)$ instead of $\mathrm{KL}(\nu_i,\kappa_i)$.

While the above lower bound is instance dependent, other lower bounds have been derived in the literature with a worst-case flavor. By worst-case lower bound, we mean that, for any sequence of algorithms $(\mathfrak{A}_T)_{T\in\mathbb{N}}$, there exists an instance ν (or a sequence of instances $(\nu_T)_{T\in\mathbb{N}}$) with mean μ such that the probability of error is asymptotically lower bounded by $C(\mu) > 0$. For example, Audibert et al. [2010] show that, for Bernoulli distributions with mean $\mu \in [p,1-p]^K$

$$\liminf_{T\to+\infty} \frac{T}{-\log\max_{\sigma_T\in\mathfrak{G}_K}\mathbb{P}_{\nu^{\sigma_T}}(\hat{\imath}_T \notin i^\star(\mu^{\sigma_T}))} \geq \frac{p(1-p)}{5}H_2(\mu)\,,$$

where $\mathfrak{G}_K$ denotes the set of permutation over $[K]$ and ν^σ denotes the bandit instance with mean vector μ^σ obtained after permuting the arms according to $\sigma \in \mathfrak{G}_K$. Carpentier and Locatelli [2016] show that, for Bernoulli distributions with mean $\mu \in [1/4,3/4]^K$,

$$\exists\nu,\quad \liminf_{T\to+\infty} \frac{T}{-\log\mathbb{P}_\nu(\hat{\imath}_T \notin i^\star(\mu))} \geq \frac{\log K}{800}H(\mu)\,.$$

Komiyama et al. [2022] conjectured that no strategy might perform uniformly well under all bandit instances, and Degenne [2023] proved this conjecture by using similar techniques as in Carpentier and Locatelli [2016].

1.5.2 Anytime Setting

In the anytime setting, no fixed parameter is given to the algorithm to define its performance metric. Compared to the fixed-confidence and the fixed-budget setting where a single real value should be minimized, an infinity of real-valued metrics should be minimized. Instead of controlling the error and minimizing the budget or controlling the budget and minimizing the error, an anytime strategy is judged based on how "good" its candidate answer is at any time.

The main advantage of an anytime strategy lies in its versatility. Ideally, it should be possible to use it without modification for fixed-confidence and fixed-budget identification, *i.e.* only the stopping rule should depend on the error δ or the budget T . Therefore, when combined with the corresponding stopping rule, an anytime strategy should be judged on its expected sample complexity at any confidence level δ , *i.e.* $\{\mathbb{E}_\nu[\tau_\delta] \mid \delta \in (0,1)\}$, and on its probabilities of error at any deterministic[3] time n , *i.e.* $\{\mathbb{P}_\nu(\mathcal{E}_\mu^{\mathrm{err}}(n)) \mid n \in \mathbb{N}\}$.

Importantly, when combined with the δ-dependent stopping rule, the algorithm should still be δ-correct. Since the anytime strategy is independent of δ , it means that the stopping rule should ensure δ-correctness regardless of the sampling rule. This property is satisfied by the GLR stopping rule, hence we will use it in the following.

Pareto front on the performance Since an anytime algorithm can be used in the fixed-confidence and fixed-budget settings, the lower bounds of those settings also apply (see Sections 1.4.1 and 1.5.1). As we will see on some specific pure exploration problem, it is not possible to achieve the best performance in both settings. The impossibility of having a "best-of-both" world algorithm opens an interesting research avenue which aims at understanding the Pareto front of the performance. This would help understand what is the fundamental trade-off between optimizing for the expected sample complexities or for the probability of errors. While the lower bound aspect of this open problem is beyond the scope of this thesis, Part II presents elements of answer for the upper bound aspect both in ε-BAI and GAI.

Contribution 1.10. *In Chapter 5, we prove anytime guarantees for the EB-TC$_\varepsilon$ algorithm. This algorithm simultaneously achieve near asymptotic optimality when combined with the GLR$_\varepsilon$ stopping rule, and has an anytime exponential decay of its probability of $\tilde{\varepsilon}$-error (for any slack $\tilde{\varepsilon} \geq 0$) which is theoretically comparable to the one of uniform sampling. In Chapter 6, we prove anytime guarantees for the APGAI algorithm. When there are no good arms, APGAI simultaneously achieve asymptotic optimality when combined with the GLR stopping rule, and has an anytime exponential decay of its probability of error which is better than the one of uniform sampling. When there are good arms, the theoretical guarantees are less satisfying, and experiments suggest that better rate could be achieved.*

[3] In constrast to a random stopping time τ_δ , a deterministic stopping time n is independent of the history $\mathcal{F}_n$.

Jun and Nowak [2016] propose the *anytime exploration* setting, in which they control the error probability $\mathbb{P}_\nu(\hat{\imath}_n \neq i_\star)$ for exact best arm identification. Interestingly, the authors build on an algorithm for the fixed-confidence setting, LUCB [Kalyanakrishnan et al., 2012], whose sampling rule depends on the risk parameter δ, which they replace by a sequence δ_n. However, it has since been discovered that there is an error in their analysis.

Different Downstream Tasks

While an anytime algorithm can be straightforwardly used to tackle the same identification problem in fixed-confidence and fixed-budget settings, an external actor could also leverage the stream of recommendation and the associated guarantees for a different downstream task.

For example, if the algorithm is designed for the ε-BAI problem, external actors could also tackle $\tilde{\varepsilon}$-BAI with $\tilde{\varepsilon} \in \{\varepsilon_k\}_{k\in[N]}$. Given $\tilde{\varepsilon} \neq \varepsilon$, those external actors can be interested by the fixed-confidence setting, *i.e.* $\{\mathbb{E}_\nu[\tau_{\tilde{\varepsilon},\delta}] \mid \delta \in (0,1)\}$, or the fixed budget setting, *i.e.* $\{\mathbb{P}_\nu(\hat{\imath}_n \notin \mathcal{I}_{\tilde{\varepsilon}}(\mu)) \mid n \in \mathbb{N}\}$.

Introduced in Audibert et al. [2010], the expected *simple regret* is defined as $\mathbb{E}_\nu[\mu_\star - \mu_{\hat{\imath}_n}]$, and is independent of any parameter ε. Simple regret is typically studied in an anytime setting: Bubeck et al. [2011] contains upper bounds on the simple regret at time n for any $n \in \mathbb{N}^*$. Since $\mathbb{E}_\nu[\mu_\star - \mu_{\hat{\imath}_n}] = \int \mathbb{P}_\nu(\hat{\imath}_n \notin \mathcal{I}_\varepsilon(\mu))\mathrm{d}\,\varepsilon$, anytime guarantees on the expected simple regret can be obtained for an algorithm with anytime guarantees on the uniform ε-error probability [Zhao et al., 2023], *i.e.* $\{\mathbb{P}_\nu(\hat{\imath}_n \notin \mathcal{I}_\varepsilon(\mu)) \mid (n,\varepsilon) \in \mathbb{N}\times\mathbb{R}_+\}$.

Finally, an external actor aiming at minimizing its cumulative regret could simply pull the arm $\hat{\imath}_n$ at time n. At time T, the regret of the *induced policy* would be defined as $\sum_{n\in[T]} \mathbb{E}_\nu[\mu_\star - \mu_{\hat{\imath}_n}]$. The idea of decoupling exploration and exploitation when minimizing the regret in the multi-armed bandits literature was introduced by Avner et al. [2012]. Naturally, anytime guarantees on the expected cumulative regret of the induced policy when the recommendations have anytime guarantees on the expected simple regret.

For other identification problems (*e.g.* GAI), it might be less clear how to define relevant downstream tasks both in terms of related pure exploration problems or in terms of regret for an induced policy. The specifics highly depend on the task tackled by the algorithm which does the data collection and streams the recommendations.

Anytime Guarantees on Existing Algorithms

Before designing algorithms tailored to the anytime setting, one could attempt to leverage existing fixed-budget and fixed-confidence algorithms, *e.g.* adapt them and/or derive addi-

tional guarantees. For the sake of example, we restrict ourselves to vanilla BAI and bounded distributions $\mathcal{D}_{[0,1]}$ in the following.

Uniform sampling rule As we saw in Section 1.5.1, the uniform algorithm is the canonical example of an anytime algorithm. When combined with the GLR or fixed-budget stopping rule, it satisfies that

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\log(1/\delta)} \leq \left(\min_{i \neq i^\star} C(i^\star, i; \nu, 1_K/K) \right)^{-1} \text{ and } \limsup_{n \to +\infty} \frac{n}{-\log \mathbb{P}_\nu(\hat{\imath}_n \notin i^\star(\mu))} \leq \frac{K}{\Delta_{\min}(\mu)^2} \, .$$

In addition of being minimax optimal for the probability of error [Bubeck et al., 2011], uniform sampling is versatile and it has anytime guarantees in both settings. However, in terms of instance-dependent guarantees, uniform sampling is worse than existing fixed-confidence and fixed-budget algorithms. As such it is often the default choice for an identification strategy when the downstream objective is not clearly defined beforehand.

From fixed-budget to anytime algorithms The doubling trick [Jun and Nowak, 2016, Zhao et al., 2023] allows the conversion of any fixed-budget algorithm into an anytime algorithm, *i.e.* an algorithm that does not depend on a budget T fixed beforehand. It considers a sequence of algorithms that are run with increasing budgets $\{T_k\}_{k \in N}$, and recommends the answer outputted by the last instance. When considering $T_{k+1} = 2T_k$ for SR (resp. SH) and $T_1 = 2K\lceil \log_2 K \rceil$, it is possible to show that the rate of decays is only impacted by a multiplicative factor 4 . When combined with the fixed-budget stopping rule, it satisfies that

$$\limsup_{n \to +\infty} \frac{n}{-\log \mathbb{P}_\nu(\hat{\imath}_n \notin i^\star(\mu))} \leq 4H_2(\mu)\overline{\log} K \quad (\text{ resp. } \quad 32 H_2(\mu) \log_2 K \,) \, .$$

The major weakness of the doubling trick lies in the fact that it resets its history when it starts a new algorithm, *i.e.* it drops observations from the runs with a smaller budget.

By using the doubling trick, we only provide one side of the guarantees that are desired for an anytime algorithm. Namely, it does not provide upper bound on the expected sample complexity that this procedure entails. Using an analysis similar to the one we developed in Azize et al. [2023], it is possible to define a δ-correct stopping rule for this procedure. First, the GLR stopping rule should only be evaluated when the current instance runs out of intermediate budget. Second, instead of considering the total allocation N_n , the allocation collected by the current instance should be used in the GLR stopping rule since only those observations are used to update the candidate answer. While the analysis of this procedure is beyond the scope of this manuscript, we showed in Azize et al. [2023] that the price of doubling and forgetting for the GLR stopping rule is also a multiplicative 4 factor. Intuitively, in the asymptotic regime, the arms will be eliminated by decreasing gap value, *i.e.* the arms with the smallest true mean

will be rejected before the others. For the SR algorithm, we conjecture that

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\log(1/\delta)} \leq 4 \left(\min_{i \neq i^\star} C(i^\star, i; \nu, w_{SR}) \right)^{-1} ,$$

where $w_{SR,(k)} = (k\overline{\log} K)^{-1}$ for all $k > 1$, and $w_{SR,(1)} = (2\overline{\log} K)^{-1}$. A similar conjecture could be made for the SH algorithm.

From fixed-confidence to anytime algorithms In Section 1.4.3, we presented three approaches to design a fixed-confidence sampling rule which are independent of δ . When combined with the GLR stopping rule, they satisfy

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\log(1/\delta)} \leq T^\star(\nu) .$$

As above, this only provide one side of the desired guarantees, and we still need to upper bound the probability of error at any time. Unfortunately, it is known that algorithms that have *optimal* asymptotic guarantees in the fixed-confidence setting can be sub-optimal in terms of error probability. Indeed Komiyama et al. [2022] prove in their Appendix C that for any asymptotically optimal (exact) BAI algorithm, there exists instances in which the error probability cannot decay exponentially with the horizon, which makes them worse than the (minimax optimal) uniform sampling strategy [Bubeck et al., 2011]. Their argument relies on the sparsity of the optimal allocation also appears in BAI when considering the limit of $\Delta_{\min} \to 0$. Since the optimal allocation are not asymptotically sparse in ε-BAI, the story is different. This allowed us to obtain good anytime guarantees for EB-TC$_\varepsilon$ (see Chapter 5 for more details).

1.6 Outline of the Thesis

After this introductory Chapter 1, this thesis is organized in three parts corresponding to the type of pure exploration problems that we studied.

In Part I, we study the vanilla BAI problem in the fixed-confidence setting through the lens of the Top Two approach, and generalize it for several class of distributions. While our contributions are theoretical, they are driven by practical considerations: (1) top two algorithms are easy to understand (and implement) and most of them are computationally efficient and (2) more general distributions can better approximate real-world distributions.

- In Chapter 2, we illustrate the Top Two approach with the set $\mathcal{D}_{\mathcal{N}_\sigma}$ of Gaussian distributions with known variance σ^2 . We show how a Top Two algorithm is defined by four choices (leader answer, challenger answer, targeted allocation and mechanism to reach it) and propose several instances. We present a unified asymptotic analysis of the Top

Two approach, which identifies desirable properties for each of those four choices. Moreover, we give the first non-asymptotic analysis of a Top Two algorithm, which identifies sufficient properties of the leader (seen as a regret-minimization algorithm) for it to hold. While different Top Two algorithms have similar empirical performance, they tend to outperform other algorithms.

- In Chapter 3, we consider the set $\mathcal{D}_{\mathcal{N}}$ of Gaussian distributions with unknown variance. We introduce and analyze two approaches to deal with unknown variances, either by plugging in the empirical variance or by adapting the transportation costs. In order to calibrate our two stopping rules, we derive new time-uniform concentration inequalities, which are of independent interest. Then, we illustrate the theoretical and empirical performances of our two sampling rule wrappers on Track-and-Stop and on a Top Two algorithm. Moreover, by quantifying the impact on the sample complexity of not knowing the variances, we reveal that it is rather small.

- In Chapter 4, we tackle the set $\mathcal{D}_{[0,B]}$ of bounded distributions with support in $[0, B]$ with $B > 0$. We extend the Top Two instances proposed in Chapter 2 for this class of non-parametric distributions, and show that the desirable properties are satisfied. We illustrate the good empirical performance of those algorithms on the DSSAT real-world data.

In Part II, our goal is to study the impact of having multiple correct answers, and understand if algorithms can have good anytime guarantees. Therefore, we consider two relaxed identification problems in the anytime setting: ε-BAI and GAI. Likewise, our contributions are motivated by practical considerations: (1) easier problems which better fit the goal of practitioners and (2) "universal" algorithms which have good guarantees in different experimental setup.

- In Chapter 5, we tackle ε-BAI in the anytime setting. We propose EB-TC$_\varepsilon$, which is the first instance of Top Two algorithm analyzed for ε-BAI. EB-TC$_\varepsilon$ is an *anytime* sampling rule that can therefore be employed without modification for fixed confidence or fixed budget identification (without prior knowledge of the budget). We provide three types of theoretical guarantees for EB-TC$_\varepsilon$. We prove bounds on its expected sample complexity in the fixed confidence setting, notably showing its asymptotic optimality in combination with an adaptive tuning of its exploration parameter. We complement these findings with upper bounds on its probability of error at any time and for any error parameter, which further yield upper bounds on its simple regret at any time. Finally, we show through numerical simulations that EB-TC$_\varepsilon$ performs favorably compared to existing algorithms, in different settings.

- In Chapter 6, we study GAI in the anytime setting. We propose APGAI, an anytime and parameter-free sampling rule for GAI in stochastic bandits, which can be straightforwardly used in fixed-confidence and fixed-budget settings. Our upper bounds on its probability

of error at any time show that adaptive strategies are more efficient in detecting the absence of good arms than uniform sampling. When APGAI is combined with a stopping rule, we prove upper bounds on the expected sampling complexity, holding at any confidence level. We illustrate the good empirical performance of APGAI on real-world data from an outcome scoring problem, as well as on synthetic instances.

In Part III, we aim at understanding the impact of the structure in pure exploration problems which have multiple correct answers. To that end, we study ε-BAI for transductive linear bandits in the fixed-confidence setting. Similarly, our contributions are motivated by a practical consideration: the linear assumption can be more realistic than the independence assumption of the vanilla setting.

- In Chapter 7, we study the importance the choice of the candidate answer for ε-BAI in transductive linear bandits. Using an instantaneous furthest answer as candidate answer, we propose a simple procedure to adapt existing BAI algorithms for ε-BAI. Leveraging it in the sampling rule as well, we propose a game based algorithm (named LεBAI) which is asymptotically optimal and has competitive empirical performance.

- In Chapter 8, we extend the Top Two approach to tackle structured bandits such as ε-BAI for transductive linear bandits. Among other Structured Top Two algorithms, we propose the LεTT algorithm which recovers EB-TC$_\varepsilon$ for vanilla BAI (see Chapter 5). We highlight the challenges in the analysis of the expected sample complexity, and perform an empirical study showcasing the good empirical performance of LεTT. The contributions in Chapter 8 are currently unpublished since most challenges are not solved yet.

1.7 Publications

During my PhD thesis, I had the opportunity to conduct several research projects with different collaborators, including my supervisors, other PhD students and researchers. Based on those projects, we published several papers which are detailed below.

Publications in international conferences with proceedings

- *Choosing answers in ε-best-answer identification for linear bandits*, by **Marc Jourdan** and Rémy Degenne. *International Conference on Machine Learning* (ICML), 2022. See Chapter 7. [Jourdan and Degenne, 2022].

- *Top two algorithms revisited*, by **Marc Jourdan**, Rémy Degenne, Dorian Baudry, Rianne De Heide and Emilie Kaufmann. *Advances in Neural Information Processing Systems* (NeurIPS), 2022. See Chapters 2 and 4. [Jourdan et al., 2022].

- *Dealing with unknown variances in best-arm identification*, by **Marc Jourdan**, Rémy Degenne and Emilie Kaufmann. *International Conference on Algorithmic Learning Theory* (ALT), 2023. See Chapter 3. [Jourdan et al., 2023a].
- *Non-asymptotic analysis of a ucb-based top two algorithm*, by **Marc Jourdan** and Rémy Degenne. *Advances in Neural Information Processing Systems* (NeurIPS), 2023. See Chapter 2. [Jourdan and Degenne, 2023].
- *An ε-best-arm identification algorithm for fixed-confidence and beyond*, by **Marc Jourdan**, Rémy Degenne and Emilie Kaufmann. *Advances in Neural Information Processing Systems* (NeurIPS), 2023. See Chapters 2 and 5. [Jourdan et al., 2023b].
- *On the complexity of differentially private best-arm identification with fixed confidence*, by Achraf Azize, **Marc Jourdan**, Aymen Al Marjani and Debabrota Basu. *Advances in Neural Information Processing Systems* (NeurIPS), 2023. [Azize et al., 2023]. We do not detail this work in this thesis, as it involves the differential privacy framework. In a nutshell, we formalize the global differentially private best-arm identification problem, and design a Top Two algorithm to solve it near optimaly. Our algorithm ensure privacy by considering private estimators which can be computed by using adaptive phases. When there is no privacy constraint, our algorithm achieves asymptotic near-optimality (up to a multiplicative 4 factor) with $\mathcal{O}(K \log(T^\star(\nu) \log(1/\delta))$ rounds of adaptivity.

Preprints

- *An anytime algorithm for good arm identification*, by **Marc Jourdan** and Clémence Réda. See Chapter 6. [Jourdan and Réda, 2023].

Part I

Fixed-confidence Best Arm Identification with Top Two Algorithms

Chapter 2

A Pedagogical Example: Gaussian with Known Variances

In Chapter 2, we study the vanilla BAI problem for Gaussian distributions with known variance in the fixed-confidence setting, as described in Chapter 1. This chapter unifies results that were published in Jourdan et al. [2022], Jourdan and Degenne [2023], Jourdan et al. [2023b].

A Top Two sampling rule for bandit identification is a method which selects the next arm to sample from among two candidate arms, a leader and a challenger. Due to their simplicity and good empirical performance, they have received increased attention in recent years. In this chapter, we propose a generic definition of the Top Two approach which requires to make four choices (leader answer, challenger answer, targeted allocation and mechanism to reach it) and provide several instances for each choice. We present a unified asymptotic analysis of the Top Two approach, which identifies desirable properties for each of those four choices. Moreover, we give the first non-asymptotic analysis of a Top Two algorithm, which identifies sufficient properties of the leader (seen as a regret-minimization algorithm) for it to hold. Empirically, different Top Two algorithms have similar performance, and they tend to outperform other algorithms.

Contents

2.1 Introduction

Despite the restricted scope of its applications, studying the identification task for Gaussian distributions is a natural first step. As we shall see, the insights gained will then be generalized to wider classes of distributions.

We consider the set $\mathcal{D}_{\mathcal{N}_\sigma}$ of Gaussian distributions with variance σ , and assume that $\sigma_i = 1$ for all $i \in [K]$ by scaling, hence $\mathcal{D}^K = \mathcal{D}^K_{\mathcal{N}_1}$. Let $\nu \in \mathcal{D}^K$ which is uniquely defined by its mean vector $\mu \in \mathbb{R}^K$ such that the set of arms with largest mean $i^\star(\mu) := \arg\max_{i\in[K]} \mu_i$ is reduced to a singleton denoted by $i^\star$ (or $i^\star(\mu)$ by abusing notation), *i.e.* $\mathcal{S} = \{\mu \in \mathbb{R}^K \mid |i^\star(\mu)| = 1\}$.

A fixed-confidence algorithm is defined by a sampling rule, a recommendation rule and a stopping rule. At time n , we denote by $\hat{\imath}_n$ the candidate answer and by I_n the arm to pull. The stopping rule (and stopping time τ_δ) using a fixed confidence level $1-\delta \in (0,1)$ which ensures δ-correctness, *i.e.* $\mathbb{P}_\nu\left(\tau_\delta < +\infty,\ \hat{\imath}_{\tau_\delta} \neq i^\star(\mu)\right) \leq \delta$ for all instances $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathcal{S}$. As explained in Section 1.4.1, the δ-correctness requirement leads to a lower bound on the expected sample complexity on any instance.

Lemma 2.1 (Garivier and Kaufmann [2016]). *An algorithm which is δ-correct on all problems in $\mathcal{D}^K$ satisfies that, for all $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathcal{S}$, $\mathbb{E}_\nu[\tau_\delta] \geq T^\star(\nu)\log(1/(2.4\delta))$, where $T^\star(\nu) = \min_{\beta\in(0,1)} T^\star_\beta(\nu)$ and, for all $\beta \in (0,1)$,*

$$T^\star_\beta(\nu)^{-1} = \max_{w\in\Sigma_K, w_{i^\star}=\beta} \min_{i\neq i^\star} C(i^\star, j; \nu, w) \quad \text{with } C(i,j;\nu,w) = \mathbb{1}\left(\mu_i > \mu_j\right) \frac{(\mu_i-\mu_j)^2}{2\left(1/w_i + 1/w_j\right)} .$$

When considering the sub-class of algorithms allocating a fraction β of their sample to the best arm, we obtain a lower bound as in Lemma 2.1 with $T^\star_\beta(\nu)$ instead of $T^\star(\nu)$. An algorithm is said to be asymptotically optimal (resp. β-optimal) if its sample complexity matches that lower bound asymptotically, that is if $\limsup_{\delta\to 0} \mathbb{E}_\nu[\tau_\delta]/\log(1/\delta) \leq T^\star(\nu)$ (resp. $T^\star_\beta(\nu)$) for all $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathcal{S}$. Russo [2016] showed the worst-case inequality $T^\star_{1/2}(\nu) \leq 2T^\star(\nu)$ for any single-parameter exponential family. Therefore, the expected sample complexity of an asymptotically β-optimal algorithm with $\beta = 1/2$ is at worst twice higher than that of any asymptotically optimal algorithm. Leveraging the symmetry of Gaussian distributions with unit variance, a tighter worst-case inequality could be derived (see Lemma C.6 in Jourdan and Degenne [2023]). The allocations $w^\star(\nu)$ and $w^\star_\beta(\nu)$ realizing $T^\star(\nu)$ and $T^\star_\beta(\nu)$ are known to be unique, and satisfy $\min_{i\in[K]} \min\{w^\star(\nu)_i, w^\star_\beta(\nu)_i\} > 0$. Barrier et al. [2022] showed that $2 \leq w^\star(\nu)^{-1}_{i^\star} \leq \sqrt{K-1}+1$ (Proposition 10 in Barrier et al. [2022]). Let $H(\mu) := 2\Delta^{-2}_{\min} + \sum_{i\neq i^\star} 2\Delta_i^{-2}$ where $\Delta_i = \mu_{i^\star} - \mu_i$ and $\Delta_{\min} = \min_{i\neq i^\star} \Delta_i$, which satisfies $H(\mu) \leq T^\star(\nu) \leq 2H(\mu)$ [Garivier and Kaufmann, 2016].

Contribution 2.2. *In Chapter 2, we illustrate the Top Two approach with the set of Gaussian distributions with unit variance.*

- *We propose a unified perspective on the class of Top Two algorithms which puts forward four choices (leader answer, challenger answer, target allocation and mechanism to reach it) and propose several instances (Section 2.2).*
- *We present a unified asymptotic analysis of the Top Two approach (Section 2.3), which identifies desirable properties on the choices of the leader and challenger answers to achieve asymptotic (β-)optimality (Theorem 2.9).*
- *We give the first non-asymptotic analysis of a Top Two algorithm (Theorem 2.25), which identifies sufficient properties of the leader (seen as a regret-minimization algorithm) for it to hold (Section 2.4).*
- *While different Top Two algorithms have similar empirical performance, they tend to outperform other algorithms (Section 2.5).*

2.1.1 GLR Stopping Rule

For an arm i , we denote its number of pulls before time n by $N_{n,i} = \sum_{t\in[n-1]} \mathbb{1}(I_t = i)$ and its empirical mean by $\mu_{n,i} = \sum_{t\in[n-1]} \mathbb{1}(I_t = i) X_{t,i}$.

For the recommendation rule, we use $\hat{\imath}_n \in i^\star(\mu_n)$ as candidate answer, *i.e.* the empirical best (EB) arm. For the stopping rule, we use the GLR stopping rule (see Section 1.4.2 for more details). For Gaussian distributions, the GLR can be written as $\min_{i\neq\hat{\imath}_n} W_n(\hat{\imath}_n, i)$, where the empirical transportation cost between arm i and arm j is defined as

$$W_n(i,j) = C(i,j;\nu_n,N_n) = \mathbb{1}(\mu_{n,i} > \mu_{n,j}) \frac{(\mu_{n,i}-\mu_{n,j})^2}{2(1/N_{n,i}+1/N_{n,j})} . \tag{2.1}$$

Given a threshold function $c(n,\delta)$, the GLR stopping rule is

$$\tau_\delta = \inf\left\{ n\in\mathbb{N} \mid \min_{j\neq\hat{\imath}_n} W_n(\hat{\imath}_n, j) > c(n-1,\delta) \right\} . \tag{2.2}$$

Lemma 4.3 gives a threshold ensuring that the GLR stopping rule is δ-correct for all $\delta \in (0,1)$, independently of the sampling rule. Its proof relies on an elegant martingale construction proposed by Kaufmann and Koolen [2021], and is detailed in Appendix B.1.

Lemma 2.3. *Let $\delta \in (0,1)$. Given any sampling rule, using the threshold*

$$c(n,\delta) = 2\mathcal{C}_G(\log((K-1)/\delta)/2) + 4\log(4+\log(n/2)) \tag{2.3}$$

with the stopping rule (2.2) yields a δ-correct algorithm for Gaussian distributions with unit variance and mean in $\mathcal{S}$. The function $\mathcal{C}_G$ is defined in (B.1). It satisfies $\mathcal{C}_G(x) \approx x + \log(x)$.

2.1.2 The Greedy GLR-based Sampling Rule

Before presenting the Top Two approach for the sampling rule (see Section 2.2), we present an intuitive sampling rule inspired by the GLR stopping rule. Since it is too *greedy* in practice, we refer to it as the greedy GLR-based sampling rule. The Top Two approach is precisely based on this intuition.

Assume that we did not stop at time n , *i.e.* $n < \tau_\delta$. Then, we have a candidate (or *leader*) answer defined as $\hat{\imath}_n \in i^\star(\mu_n)$, and an alternative (or *challenger*) answer defined as $\hat{\jmath}_n \in \arg\min_{j \neq \hat{\imath}_n} W_n(\hat{\imath}_n, j)$. Since the stopping condition has not been met yet, *i.e.* $W_n(\hat{\imath}_n, \hat{\jmath}_n) \leq c(n-1,\delta)$, the collected observations do not provide enough evidence to reject the null hypothesis that $\hat{\imath}_n$ is worse than $\hat{\jmath}_n$ with probability at least $1-\delta$.

Since the algorithm has not stopped, the sampling rule should define a next arm to pull I_n . An intuitive idea is to collect additional observation to verify that $\hat{\imath}_n$ is better than $\hat{\jmath}_n$. Due to the lack of underlying structure and the fact that answers are arms, it is enough to collect observations from an arm $I_n \in \{\hat{\imath}_n, \hat{\jmath}_n\}$. There are many ways to define which arm to choose among both options, and it could be done randomly or deterministically (see Section 2.2). For example, a data-independent choice could be to pull the leader half of the time. A more sophisticated data-dependent choice could be to pull the leader a fraction of the time that depends on how often this challenger was sampled compared to the leader, *i.e.* $N_{n,\hat{\jmath}_n}/(N_{n,\hat{\imath}_n} + N_{n,\hat{\jmath}_n})$.

In regret minimization, the greedy sampling rule is also an intuitive sampling idea which simply pulls the arm with highest empirical mean. It is well known that the greedy algorithm can incur linear regret since it is not able to recover from unlucky first draws. To circumvent this problem, the two most well-known approaches are to use randomization or to apply the principle of optimism in face of uncertainty.

Due to similar reasons, this intuitive sampling rule based on the GLR stopping rule is also too greedy. To see this, let us consider hard instances $\mu = (0, -\Delta, -\Delta)$ where $\Delta > 0$. There is a non-negligible probability that the three observations collected during initialization are unlucky. By unlucky first draws, we mean that the true best arm is believed to be significantly worse than

the two sub-optimal arms, *i.e.* $X_{1,1} \ll \min\{-\Delta, X_{2,2}, X_{3,3}\}$. Then, we have $1 \notin \{\hat{\imath}_4, \hat{\jmath}_4\}$ hence the algorithm will sample one of the two sub-optimal arms. It is possible to repeat the arguments since $X_{1,1} \ll -\Delta$. There is a large probability that the additional collected observations will ensure that $\mu_{n,1} = X_{1,1} \ll \min\{-\Delta, \mu_{n,2}, \mu_{n,3}\}$ and $|\mu_{n,2} - \mu_{n,3}| \to 0$. Therefore, the best arm will never be sampled after initialization, *i.e.* $1 \notin \{\hat{\imath}_n, \hat{\jmath}_n\}$ for all $n > 3$. Both randomization and optimism can be used to add implicit exploration, hence solving the limitations of the GLR-based sampling rule (see Section 2.2).

2.2 Generic Top Two Sampling Rule

Numerous Top Two algorithms have been proposed in recent years. In this section, we propose a unified perspective on this class of sampling rules which puts forward four choices which have to be made to define a Top Two algorithm: the leader answer (Section 2.2.1), the challenger answer (Section 2.2.2), the target allocation (Section 2.2.3) and the mechanism to reach the target (Section 2.2.4). The Top Two approach is summarized in Algorithm 2.1, and we will present several instances for each choice (see Section 2.2.5 for a naming convention).

Input: Mechanisms to choose the leader answer $\mathcal{L}^B$, the challenger answer $\mathcal{L}^C$, the target allocation $\mathcal{L}^T$ and how to reach the target $\mathcal{L}^R$.
Output: Next arm to sample I_n .
Get $B_n \in [K]$ from $\mathcal{L}^B$; `// Leader answer`
Get $C_n \in [K] \setminus \{B_n\}$ from $\mathcal{L}^C$; `// Challenger answer`
Get $\beta_n(B_n, C_n) \in [0, 1]$ from $\mathcal{L}^T$; `// Target allocation`
Get $I_n \in \{B_n, C_n\}$ from $\mathcal{L}^R$ using $\beta_n(B_n, C_n)$; `// Reaching the target`

Algorithm 2.1: Generic Top Two sampling rule.

As initialization, we start by sampling each arm once. In the following, the conditioning on the history $\mathcal{F}_n$ is denoted by $\mathbb{E}_{\cdot|n}$ and $\mathbb{P}_{\cdot|n}$ for the expectations and probabilities.

2.2.1 Leader Answer

Based on the intuitive GLR-based sampling rule, the only requirement for the leader answer is to be a good estimator of the correct answer, *i.e.* an arm with highest mean $i^\star(\mu)$. As such, there are many ways to define it both with a frequentist or a Bayesian approach, and we summarize some of them in Table 2.1.

Table 2.1 – Choices for the leader answer: EB (Empirical Best) (LUCB in Kalyanakrishnan et al. [2012] and EB-TC in Jourdan et al. [2022]), UCB (Upper Confidence Bound) (TTUCB in Jourdan and Degenne [2023]), TS (Thompson Sampling) (TTTS in Russo [2016]), EI (Expected Improvement) (TTEI in Qin et al. [2017]), PS (Probability Sampling) (TTPS in [Russo, 2016]). Let $\mu_n^\star = \max_{i\in[K]} \mu_{n,i}$.

Definition	Approach	Computational requirements
$B_n^{\mathrm{EB}} \in \arg\max_{i\in[K]} \mu_{n,i}$	Frequentist	None
$B_n^{\mathrm{UCB}} \in \arg\max_{i\in[K]} U_{n,i}$	Frequentist	Efficient UCB indices
$B_n^{\mathrm{TS}} \in \arg\max_{i\in[K]} \theta_{n,i}$ with $\theta_n \sim \Pi_n$	Bayesian	Efficient sampler Π_n
$B_n^{\mathrm{EI}} \in \arg\max_{i\in[K]} \mathbb{E}_{\theta_n\sim\Pi_n\mid n}[(\theta_{n,i} - \mu_n^\star)_+]$	Bayesian	Efficient approximation $\mathbb{E}_{\Pi_n\mid n}$
$B_n^{\mathrm{PS}} \in \arg\max_{i\in[K]} \mathbb{P}_{\theta_n\sim\Pi_n\mid n}(i \in i^\star(\theta_n))$	Bayesian	Efficient approximation $\mathbb{P}_{\Pi_n\mid n}$

Frequentist approach With a frequentist perspective, one can simply consider the MLE (Maximum Likelihood Estimator) which is our candidate answer $\hat{\imath}_n$. The EB leader chooses the arm with highest empirical mean [Kalyanakrishnan et al., 2012, Jourdan et al., 2022], *i.e.* $B_n^{\mathrm{EB}} = \hat{\imath}_n \in \arg\max_{i\in[K]} \mu_{n,i}$. The EB leader is by far the least computationally demanding choice, and it is agnostic to the underlying class of distributions.

As mentioned above, combining the EB leader with a greedy challenger (*i.e.* TC challenger defined in Section 2.2.2) yields a sampling rule which is too greedy. Therefore, additional exploration could be enforced (implicitly or explicitly) when defining the leader. The principle of optimism in face of uncertainty was designed to cope for this limitation. As proxy for the unknown mean μ , we consider UCB indices such that, with high probability, $U_{n,i} \geq \mu_i$ for all $(n, i) \in \mathbb{N} \times [K]$. From a computational perspective, those indices should be efficient to compute. For Gaussian distributions with unit variance, using a bonus function $g : \mathbb{N} \to \mathbb{R}_+$, we obtain

$$U_{n,i} = \max\left\{u \in \mathbb{R} \mid N_{n,i}\mathcal{K}_{\inf}^+(\nu_{n,i}, u) \leq g(n)\right\} = \mu_{n,i} + \sqrt{2g(n)/N_{n,i}}\,, \tag{2.4}$$

where the last equality uses that $\mathcal{K}_{\inf}^+(\kappa, u) = (m(\kappa) - \max\{m(\kappa), u\})^2/2$ for $\kappa \in \mathcal{D}_{\mathcal{N}_1}$. Using concentration inequalities, we can set $g(n) = \Theta(\log n)$. The UCB leader [Jourdan and Degenne, 2023] sets $B_n^{\mathrm{UCB}} \in \arg\max_{i\in[K]} U_{n,i}$. By adding a bonus to the empirical mean, we are optimistic since we consider that the means are better than suggested by our observations.

Many other frequentist approaches can be used to define the leader answer. For example, we could build on the IMED (Indexed Minimum Empirical Divergence) algorithm [Honda and Takemura, 2015], and define the IMED leader as

$$B_n^{\mathrm{IMED}} \in \arg\min_{i\in[K]} \left\{N_{n,i}\mathcal{K}_{\inf}^+(\nu_{n,i}, \mu_n^\star) + \log N_{n,i}\right\} = \arg\min_{i\in[K]} \left\{N_{n,i}(\mu_{n,i} - \mu_n^\star)^2 + 2\log N_{n,i}\right\}\,, \tag{2.5}$$

where $\mu_n^\star = \max_{i\in[K]} \mu_{n,i}$ is the largest mean, and the second equality holds for Gaussian distributions with unit variance. The IMED leader is also efficient to compute.

Bayesian approach When adopting a Bayesian perspective, we need to have access to a posterior distribution Π_n based on the history $\mathcal{F}_n$. For vanilla bandits, the posterior distribution has a product form: $\Pi_n = \Pi_{n,1} \times \cdots \Pi_{n,K}$ where $\Pi_{n,i}$ leverages $\mathcal{H}_{n,i} := (X_{1,i}, \dots, X_{N_{n,i},i})$, which is the history of samples from arm i before time n . The choice of the prior distribution $\Pi_{1,i}$ is tightly connected to the underlying class of distributions, *e.g.* conjugate prior of the likelihood function for exponential family of distributions. For Gaussian distributions with unit variance, using the improper prior $\Pi_{1,i} = \mathcal{N}(0, +\infty)$ yields $\Pi_{n,i} = \mathcal{N}(\mu_{n,i}, 1/N_{n,i})$ as posterior distribution of arm i before time n .

The TS (Thompson Sampling) leader was introduced in Russo [2016]. It selects an arm with highest coordinate for a vector drawn from the posterior distribution, *i.e.* $B_n^{\mathrm{TS}} \in i^\star(\theta_n)$ where $\theta_n \sim \Pi_n$. The TS leader is a randomized choice, yet deterministic choices of the leader have also been proposed with a Bayesian flavor. The EI (Expected Improvement) leader [Qin et al., 2017] chooses an arm with highest expectation (with respect to the posterior) of improving upon the highest empirical mean, *i.e.* $B_n^{\mathrm{EI}} \in \arg\max_{i\in[K]} \mathbb{E}_{\theta_n\sim\Pi_n|n}\left[(\theta_{n,i} - \mu_n^\star)_+\right]$. The PS (Probability Sampling) leader [Russo, 2016] returns an arm that has the highest probability (with respect to the posterior) of having the highest mean, *i.e.* $B_n^{\mathrm{PS}} \in \arg\max_{i\in[K]} \mathbb{P}_{\theta_n\sim\Pi_n|n}(i \in i^\star(\theta_n))$.

From a computational perspective, the posterior distribution should be easy to sample from (TS leader) or yields formulas that are efficient to compute after integration (EI and PS leader). For Gaussian distributions with unit variance, it is efficient to sample from Π_n and we have explicit formulas (except for the PS leader), *e.g.*

$$\mathbb{E}_{\theta_n\sim\Pi_n|n}\left[(\theta_{n,i} - \mu_n^\star)_+\right] = N_{n,i}^{-1/2} f\left(\sqrt{N_{n,i}}\,(\mu_n^\star - \mu_{n,i})\right) , \tag{2.6}$$

where $f(x) = -x\Phi(-x) + \phi(-x) \approx \exp(-x^2/2)$, Φ and ϕ are the cdf and pdf of $\mathcal{N}(0,1)$.

Other Bayesian approaches, which have been considered for regret minimization, could be used to define the leader answer. For example, we could build on the VBOS (Variational Bayesian Optimistic Sampling) algorithm [O'Donoghue and Lattimore, 2021] which was introduced for regret minimization, and define the VBOS leader as

$$B_n^{\mathrm{VBOS}} \sim \pi_n \quad \text{with} \quad \pi_n \in \arg\max_{w\in\Sigma_K} \sum_{i\in[K]} w_i \left(\mu_{n,i} + (\Phi_{n,i}^\star)^{-1}(-\log w_i)\right) , \tag{2.7}$$

where $(\Phi_{n,i}^\star)^{-1}$ is the inverse of the convex conjugate of the empirical cumulant generating function. Recall that the cumulant generating function of a distribution $\kappa \in \mathcal{D}^K$ is $\Phi_\kappa(\lambda) = \log \mathbb{E}_{X\sim\kappa} \exp(\langle \lambda, X - m(\kappa)\rangle)$ and the convex conjugate of a function f is $f^\star(y) = \sup_x\{\langle x, y\rangle - f(x)\}$. While there is no explicit formulas for π_n , it is the solution of a concave maximization

problem, which is computationally tractable so long as each $\Phi_{n,i}$ is readily accessible (which is the case for Gaussian distributions).

Regret minimization approach The terminology Top Two was first used to describe the TTTS algorithm, which arose as an adaptation of Thompson sampling to best arm identification in multi-armed bandit models [Russo, 2016]. In the above choices of leader answer, we recognize other regret-minimization algorithms, *e.g.* UCB, EI, IMED, VBOS. Based on those remarks, we can clearly see that: *the Top Two method can be used as a generic wrapper to convert any regret minimization algorithm into a best arm identification strategy*.

Let Alg be a regret minimization algorithm which is independent of the horizon n . The Alg leader [Jourdan and Degenne, 2023] is defined as the sampling recommendation of Alg at time n (*i.e.* given the history $\mathcal{F}_n$). While this allows to design many Top Two algorithms, there are Top Two algorithms which do not fall into this category, *e.g.* PS and EB leaders.

2.2.2 Challenger Answer

Based on the intuitive GLR-based sampling rule, the challenger answer should be taken as an alternative (or confusing) answer which is challenging our belief that the leader is a correct answer, *i.e.* an arm which might have a higher mean. There are many ways to define it both with a frequentist or a Bayesian approach, and we summarize some of them in Table 2.2. When the aim is to design an asymptotically (β-)optimal algorithm, the choice of the challenger becomes quite restrictive. Since we know that there exists a unique (resp. β-)optimal allocation $w^\star(\nu)$ (resp. $w^\star_\beta(\nu)$), the challenger should be chosen to ensure convergence towards those allocations, without computing them explicitly. While the leader should simply identify $i^\star$, the choice of the challenger should optimally balance between all the remaining arms $[K] \setminus \{i^\star\}$.

Frequentist approach With a frequentist perspective, one can consider the most confusing alternative answer compared to the leader answer. The TC challenger [Shang et al., 2020] chooses the arm with minimal empirical transportation cost compared to the leader, *i.e.* $C_n^{\mathrm{TC}} \in \arg\min_{i \neq B_n} W_n(B_n, i)$.

As mentioned above, the TC challenger might be too greedy when no additional exploration is enforced (implicitly or explicitly) in the choice of the leader. Similarly as for frequentist choices of leader, optimism can be used by adding a bonus to the empirical transportation cost, *i.e.* our belief that the leader answer is the best arm will be larger than suggested by the observations. There are several ways to define the bonus. To foster exploration implicitly, the bonus should penalize arms that are sampled the most. Intuitively, the penalization should be large enough to cope for the randomness of the empirical transportation cost (*i.e.* be an upper

Table 2.2 – Choices for the challenger answer given a leader B_n: TC (Transportation Cost) (T3C in Shang et al. [2020]), TCI (TC Improved) (EB-TCI in Jourdan et al. [2022]), KKT (Karush-Kuhn-Tucker) (TS-KKT(ρ) in You et al. [2023], RS (Re-Sampling) (TTTS in Russo [2016]), PPS (Posterior PS) (TS-PPS in You et al. [2023]), EI (Expected Improvement) (TTEI in Qin et al. [2017]), PS (Probability Sampling) (TTPS in Russo [2016]).

Definition	Approach	Computational requirements
$C_n^{\mathrm{TC}} \in \arg\min_{i \neq B_n} W_n(B_n, i)$	Frequentist	Efficient empirical TC
$C_n^{\mathrm{TCI}} \in \arg\min_{i \neq B_n} \{W_n(B_n, i) + \log N_{n,i}\}$	Frequentist	Efficient empirical TC
$C_n^{\mathrm{KKT}} \in \arg\min_{i \neq B_n} \{W_n(B_n, i)$ $-\rho \log\left(1/N_{n,B_n} + 1/N_{n,i}\right)/n\}$ with $\rho > 0$	Frequentist	Efficient empirical TC
$C_n^{\mathrm{RS}} \in \arg\max_{i \in [K]} \theta_{n,i}$ with $\theta_n \sim \Pi_n$ until $B_n \notin i^\star(\theta_n)$	Bayesian	Highly efficient sampler Π_n
$C_n^{\mathrm{PPS}} \sim (p_{n,i})_{i \in [K]}$ with $p_{n,B_n} = 0$ and $p_{n,i} \propto \mathbb{P}_{\theta_n \sim \Pi_n \mid n}\left(\theta_{n,i} > \theta_{n,B_n}\right)$	Bayesian	Efficient approximation $\mathbb{P}_{\Pi_n \mid n}$
$C_n^{\mathrm{EI}} \in \arg\max_{i \neq B_n} \mathbb{E}_{\theta_n \sim \Pi_n \mid n}\left[(\theta_{n,i} - \theta_{n,B_n})_+\right]$	Bayesian	Efficient approximation $\mathbb{E}_{\Pi_n \mid n}$
$C_n^{\mathrm{PS}} \in \arg\max_{i \neq B_n} \mathbb{P}_{\theta_n \sim \Pi_n \mid n}(i \in i^\star(\theta_n))$	Bayesian	Efficient approximation $\mathbb{P}_{\Pi_n \mid n}$

confidence bound), yet not too large to prevent the choice of the challenger answer to be close to the uniform one. Using concentration inequalities, we can show that, with high probability,

$$\left| \sqrt{W_n(i,j)} - \sqrt{C(i,j;\nu,N_n)} \right| = \mathcal{O}(\sqrt{\log n}) \,,$$

hence it is natural to use a logarithmic penalization. Based on those considerations and inspired by the IMED algorithm, the TCI challenger [Jourdan et al., 2022] chooses $C_n^{\mathrm{TCI}} \in \arg\min_{i \neq B_n} \{W_n(B_n, i) + \log N_{n,i}\}$. A slightly different choice was made by You et al. [2023] with the KKT challenger, *i.e.* $C_n^{\mathrm{KKT}} \in \arg\min_{i \neq B_n} \{W_n(B_n, i) - \frac{\rho}{n} \log\left(1/N_{n,B_n} + 1/N_{n,i}\right)\}$.

While the Top Two terminology was introduced in Russo [2016], the first sampling rule to have a Top Two structure is the greedy sampling strategy in LUCB1 [Kalyanakrishnan et al., 2012], which selects the EB leader and the UCB challenger, *i.e.* $C_n^{\mathrm{UCB}} \in \arg\min_{i \neq B_n} U_{n,i}$. Then, it samples them both. Instead of using the GLR stopping rule, LUCB1 stops when the LCB (lower confidence bound) of the leader exceeds the UCB of the challenger.

Bayesian approach Using a Bayesian perspective, one can re-sample vectors from the posterior distribution until the leader is not among the arms with highest mean, *i.e.* we are sampling with rejection. The RS challenger [Russo, 2016] is based on this resampling strategy, *i.e.* $C_n^{\mathrm{RS}} \in \arg\max_{i \in [K]} \theta_{n,i}$ with $\theta_n \sim \Pi_n$ until $B_n \notin i^\star(\theta_n)$. Given that Π_n concentrates towards the Dirac in $\{\mu\}$, it will become computationally expensive to sample until $B_n \notin i^\star(\theta_n)$ when $B_n = i^\star$

(since the probability of this event converges towards zero). Therefore, the sampler should be highly efficient for the RS challenger to be tractable for large time n .

Other choices of challenger have been proposed with a Bayesian flavor. The EI challenger [Qin et al., 2017] chooses an arm with highest expectation (with respect to the posterior) of improving upon the leader answer, *i.e.* $C_n^{\text{EI}} \in \arg\max_{i \neq B_n} \mathbb{E}_{\theta_n \sim \Pi_n|n} \left[(\theta_{n,i} - \theta_{n,B_n})_+\right]$. Given the posterior distribution defined in Section 2.2.1, we have

$$\mathbb{E}_{\theta_n \sim \Pi_n|n} \left[(\theta_{n,i} - \theta_{n,B_n})_+\right] = \sqrt{1/N_{n,B_n} + 1/N_{n,i}} f\left(\frac{\mu_{n,B_n} - \mu_{n,i}}{\sqrt{1/N_{n,B_n} + 1/N_{n,i}}}\right) , \tag{2.8}$$

where $f(x) = -x\Phi(-x) + \phi(-x) \approx \exp(-x^2/2)$, Φ and ϕ are the cdf and pdf of the standard normal distribution. The PS challenger [Russo, 2016] returns an arm that has the highest probability (with respect to the posterior) of having the highest mean (excluding B_n), *i.e.* $C_n^{\text{PS}} \in \arg\max_{i \neq B_n} \mathbb{P}_{\theta_n \sim \Pi_n|n}(i \in i^\star(\theta_n))$. The PPS challenger [You et al., 2023] samples an arm proportionally to the probability (with respect to the posterior) that its mean exceed the one of the leader, *i.e.* $B_n^{\text{PPS}} \sim (p_{n,i})_{i \in [K]}$ with $p_{n,B_n} = 0$ and $p_{n,i} \propto \mathbb{P}_{\theta_n \sim \Pi_n|n} (\theta_{n,i} > \theta_{n,B_n})$.

Similarly, we could define other Bayesian challengers. For example, a VBOS challenger could ne defined as $C_n^{\text{VBOS}} \sim (\tilde{\pi}_{n,i})_{i \in [K]}$ with $\tilde{\pi}_{n,B_n} = 0$ and $\tilde{\pi}_{n,i} = \pi_{n,i}/(1 - \pi_{n,B_n})$ where π_n is the VBOS policy [O'Donoghue and Lattimore, 2021] defined as in (2.7).

Why will the allocation be balanced optimally ? Assume that the leader answer identifies the best arm, *i.e.* $B_n = i^\star$ for n large enough. Given the posterior distribution defined in Section 2.2.1, we can approximately show that, for n large enough,

$$\forall i \neq i^\star, \quad \mathbb{P}_{\theta_n \sim \Pi_n|n}(i \in i^\star(\theta_n) \setminus \{i^\star\}) \approx \mathbb{P}_{\theta_n \sim \Pi_n|n}(i \in i^\star(\theta_n)) \approx \mathbb{P}_{\theta_n \sim \Pi_n|n} (\theta_{n,i} > \theta_{n,i^\star}) .$$

A crucial step in the analysis of Bayesian approach for the challenger is to control the probability that the drawn vector is such that $i^\star$ is believed to be worse than another arm i , *i.e.* $\mathbb{P}_{\theta_n \sim \Pi_n|n} (\theta_{n,i} > \theta_{n,i^\star})$. Therefore, we need both concentration results (*i.e.* upper bounds) and anti-concentration results (*i.e.* lower bounds), the former being often easier to derive than the later. Qin et al. [2017] showed that $\mathbb{P}_{\theta_n \sim \Pi_n|n} (\theta_{n,i} > \theta_{n,i^\star}) \approx \exp(-W_n(i^\star, i))$ for Gaussian distributions with unit variance. Directly using (2.8), we see that $\mathbb{E}_{\theta_n \sim \Pi_n|n} \left[(\theta_{n,i} - \theta_{n,i^\star})_+\right] \approx \exp(-W_n(i^\star, i))$. Asymptotically, this implies that the Bayesian approach for the challenger will coincide with the TC challenger.

Similarly, all the frequentist challengers proposed above are approximately equivalent to the TC challenger, except for the UCB challenger which is an asymptotically sub-optimal choice. Therefore, it is enough to understand why the TC challenger balances optimally between the remaining arms $[K] \setminus \{i^\star\}$.

Let us assume that there is sufficient exploration to ensure that $\mu_n \approx \mu$ for n large enough. Let $w_n = N_n/(n-1) \in \Sigma_K$ be the empirical proportions. Then, we have

$$W_n(i^\star, i) \approx (n-1)C(i^\star, i, \nu, w_n) \quad \text{hence} \quad C_n \in \arg\min_{i \neq i^\star} C(i^\star, i, \nu, w_n) \,.$$

At this point, we need to make an assumption as regards the choice $I_n \in \{i^\star, C_n\}$, *i.e.* the target allocation over arms which will be discussed in Section 2.2.3. For the sake of exposure, we discuss the fixed design of sampling the leader a fixed fraction $\beta \in (0,1)$ of the time, hence $w_{n,i^\star} \approx \beta$. Asymptotically, the empirical proportions will reach the β-equilibrium, *i.e.* $w_n \approx w^\star_\beta(\nu)$, which is the unique allocation $w \in \Sigma_K$ with $w_{i^\star} = \beta$ such that

$$\forall i \neq i^\star, \quad C(i^\star, i, \nu, w) = T^\star_\beta(\nu)^{-1} \,.$$

To see that, we can suppose towards contradiction that, there exists $\varepsilon > 0$ and $i \neq i^\star$, such that $w_{n,i} > w^\star_\beta(\nu)_i + \varepsilon$. Then, there exists $j \notin \{i, i^\star\}$ such that $w_{n,i} < w^\star_\beta(\nu)_i - \varepsilon$, and we can show that $C(i^\star, i, \nu, w_n) > C(i^\star, j, \nu, w_n)$ by using the monotonicity of transportation costs as a function of their allocation. Intuitively, an arm that overshoots its β-optimal allocation will cease to be sampled until it gets close enough to it. By formalizing this intuition, we can show that the empirical allocation converges towards the β-optimal allocation. Therefore, by going back to the GLR stopping rule, we have $\tau_\delta \lesssim T^\star_\beta(\nu)\log(1/\delta)$ which yields asymptotic β-optimality (see Section 1.4.2).

To reach asymptotic optimality, we should ensure that the best arm is sampled with the optimal proportion, *i.e.* $w_{n,i^\star} \approx \beta^\star = w^\star(\nu)_{i^\star}$. This was one of the main open problem for Top Two algorithms until it was solved recently by You et al. [2023] (see Section 2.2.3).

2.2.3 Target Allocation Over Arms

To define the target allocation over arms, we leverage the fact that there is no underlying structure and that answers are arms, hence it is enough to collect observations from an arm $I_n \in \{B_n, C_n\}$. Conditioned on the leader/challenger pair (B_n, C_n) , defining a target allocation over those two arms is equivalent to define a target for the leader B_n , which we will denote by $\beta_n(B_n, C_n) \in [0,1]$. A running average of those targets represents the fraction of time the leader B_n should be pulled when selecting (B_n, C_n) as leader/challenger pair.

Fixed design approach While they are doomed to be sub-optimal, choosing the target allocation according to a fixed design arose out of convenience, until a provably better choice was discovered. The first approach is to fix the target allocation to a given proportion $\beta \in (0,1)$ [Russo, 2016], *i.e.* $\beta_n(B_n, C_n) = \beta$ for all n . While this design will be at most asymptotically β-optimal

(*i.e.* reaching $T^\star_\beta(\nu)$), it enjoys good empirical performance on most instances for moderate value of δ . Theoretically, the fixed design is near optimal and, for $\beta = 1/2$, it is at most twice worse than the asymptotic optimal algorithm, *i.e.* $T^\star_{1/2}(\nu) \leq 2T^\star(\nu)$. LUCB1 [Kalyanakrishnan et al., 2012] pulls both arms, which approximately corresponds to taking $\beta = 1/2$.

While considering a fixed β has some benefits, it has the main weakness of being agnostic to the considered challenger. To cope for this limitation while constraining the design, another natural choice is to simply pull the least sampled arm, *i.e.* $\beta_n(B_n, C_n) = \mathbb{1}\left(N_{n,B_n} \leq N_{n,C_n}\right)$. For Gaussian with unit variance, this is equivalent to the best challenger heuristic policy (*e.g.* BC [Garivier and Kaufmann, 2016, Ménard, 2019]) which selects the arm with largest gradient of the empirical transportation cost. It is also equivalent to taking the arm which, if sampled, increases the empirical transportation cost the most would the estimator μ_n be unchanged (*e.g.* greedy choice in LinGapE [Xu et al., 2018]), *i.e.*

$$\operatorname*{arg\,min}_{i\in\{B_n,C_n\}} N_{n,i} = \operatorname*{arg\,max}_{i\in[K]} C(B_n, C_n; \nu_n, N_n + 1_i) = \operatorname*{arg\,max}_{i\in[K]} \frac{\partial C(B_n, C_n; \nu_n, N_n)}{\partial w_i} ,$$

where $1_i = (\mathbb{1}(j = i))_{j\in[K]}$ and ties are broken arbitrarily at random. Asymptotically, the BC-TE algorithm [Lee et al., 2023] is equivalent to using the TS leader and the TC challenger with this rule, even though it is not a Top Two algorithm per se. Lee et al. [2023] shows that BC-TE achieves another notion of asymptotic near-optimality, which is problem-dependent instead of being problem-independent as β-optimality. It involves a characteristic time $\underline{T}(\nu)$ defined as

$$\underline{T}(\nu)^{-1} = \sup_{w\in\Sigma_K, \frac{w_{(2)}}{w_{i^\star}+w_{(2)}}=\gamma} \min_{i\neq i^\star} C(i^\star, i; \nu, w) , \tag{2.9}$$

where (2) is the second best arm and γ is is the best ratio between the best arm and the second best arm to distinguish them, *i.e.*

$$\frac{\partial C(i^\star, (2); \nu, (1-\gamma)1_{i^\star} + \gamma 1_{(2)})}{\partial w_{i^\star}} = \frac{\partial C(i^\star, (2); \nu, (1-\gamma)1_{i^\star} + \gamma 1_{(2)})}{\partial w_{(2)}} . \tag{2.10}$$

For Gaussian distribution with unit variance, it simplifies to $\gamma = 1/2$.

Even though fixed designs allow to achieve asymptotic near-optimality, finding the optimal design for the target allocation was an active area of research.

Optimal design IDS You et al. [2023] propose the IDS (Information Directed Selection) choice for the target allocation, defined as $\beta_n(B_n, C_n)$ with $\beta_n(i, j) = 1/2$ when $\mu_{n,i} \leq \mu_{n,j}$ and otherwise

$$\beta_n(i, j) = \frac{N_{n,i}\frac{\partial C(i,j;\nu_n,N_n)}{\partial w_i}}{W_n(i, j)} = \frac{N_{n,i}\mathcal{K}^-_{\inf}(\nu_{n,i}, u_{i,j}(\nu_n, N_n))}{W_n(i, j)} , \tag{2.11}$$

where we used Lemma 2.4 for the second equality with $u_{i,j}(\nu_n, N_n)$ is a minimizer realizing $W_n(i,j)$ defined therein. The proof of Lemma 2.4 is detailed in Appendix B.2.

Lemma 2.4. *Let* $u_{i,j}(\nu, w) \in \arg\min_{u\in\mathbb{R}}\{w_i \mathcal{K}_{\inf}^{-}(\nu_i, u) + w_j \mathcal{K}_{\inf}^{+}(\nu_j, u)\}$ *be a minimizer of* $C(i,j;\nu,w)$ *for all* $i \neq j$ *. Then, for all* $j \neq i^\star$ *,*

$$\frac{\partial C(i^\star, j; \nu, w)}{\partial w_{i^\star}} = \mathcal{K}_{\inf}^{-}(\nu_{i^\star}, u_{i^\star,j}(\nu, w)) \quad \text{and} \quad \frac{\partial C(i^\star, j; \nu, w)}{\partial w_j} = \mathcal{K}_{\inf}^{+}(\nu_j, u_{i^\star,j}(\nu, w)) \,. \tag{2.12}$$

For Gaussian with unit variance, we have

$$\beta_n(i,j) = N_{n,j}/(N_{n,i} + N_{n,j}) \quad \text{if} \quad \mu_{n,i} > \mu_{n,j} \quad \text{and} \quad \beta_n(i,j) = 1/2 \quad \text{otherwise} \,. \tag{2.13}$$

Importantly, the IDS proportions are independent of the empirical means for Gaussian with known variance.

The IDS proportions are obtained by simplifying the dual formulation of the optimization problem $T^\star(\nu)^{-1} = \max_{w\in\Sigma_K} \min_{i\neq i^\star} C(i^\star, i; \nu, w)$ which can be seen as the following convex optimization problem

$$T^\star(\nu)^{-1} = \max\left\{ \phi \mid \sum_{i\in[K]} w_i = 1,\ \forall i \in [K],\ w_i \geq 0,\ \forall i \neq i^\star,\ \phi - C(i^\star, i; \nu, w) \leq 0 \right\} \,. \tag{2.14}$$

Lemma 2.5 gives a necessary and sufficient condition for optimality in (2.14), which features a dual allocation vector $\gamma \in \Sigma_{K-1}$. Intuitively, the dual variable γ_i should be thought as the conditional probability of selecting arm i as challenger given that the leader is $i^\star$. Moreover, $\beta(i^\star, i; \nu, w)$ represents the conditional probability of pulling arm i given that the leader/challenger pair of answers is $(i^\star, i)$. Therefore, it is intuitive to take $\beta_n(i,j) = \beta(i^\star, i; \nu_n, N_n)$. The proof of Lemma 2.5 is detailed in Appendix B.3, and it was known in the literature (*e.g.* Proposition 4 in Qin and You [2023]).

Lemma 2.5. *A feasible solution* (ϕ, w) *is optimal for* (2.14) *if and only if* $\phi = T^\star(\nu)^{-1}$ *and there exists a dual variable* $\gamma \in \Sigma_{K-1}$ *such that,* $\gamma_i(\phi - C(i^\star, i; \nu, w)) = 0$ *for all* $i \neq i^\star$ *, and*

$$w_i = \begin{cases} \sum_{i\neq i^\star} \gamma_i \beta(i^\star, i; \nu, w) & \text{if } i = i^\star \\ \gamma_i \left(1 - \beta(i^\star, i; \nu, w)\right) & \text{otherwise} \end{cases} \,, \quad \text{where} \quad \beta(i^\star, i; \nu, w) = \frac{w_{i^\star} \frac{\partial C(i^\star, i; \nu, w)}{\partial w_{i^\star}}}{C(i^\star, i; \nu, w)} \,.$$

Optimal design BOLD In the fixed-budget setting, Chen and Ryzhov [2023] proposed the BOLD (Balancing Optimal Large Deviations) choice for the target allocation. Adapted to the fixed-confidence setting (by swapping the arguments in the KL), BOLD selects

$$I_n = B_n \quad \text{if} \quad \sum_{i \neq B_n} \frac{\mathcal{K}_{\inf}^{-}(\nu_{B_n}, u_{B_n,i}(\nu_n, N_n))}{\mathcal{K}_{\inf}^{+}(\nu_i, u_{B_n,i}(\nu_n, N_n))} > 1 \quad \text{and} \quad I_n = C_n \quad \text{otherwise}\,, \tag{2.15}$$

where $u_{i,j}(\nu, w)$ is defined in Lemma 2.4. For Gaussian distributions with unit variance, the BOLD choice can be written as

$$I_n = B_n \quad \text{if} \quad N_{n,B_n}^2 < \sum_{i \neq B_n} N_{n,i}^2\,, \quad \text{and} \quad I_n = C_n \quad \text{otherwise}\,, \tag{2.16}$$

which coincides with the approach proposed in Shin et al. [2018] for the fixed-budget setting.

Lemma 2.6 gives a necessary and sufficient condition for optimality in (2.14). While it still features the information balance from Lemma 2.5, it introduces the concept of overall balance between arms. Since the *l.h.s.* of the overall balance is an decreasing function of $w_{i^\star}$, it is intuitive to sample the leader when the empirical version of the *l.h.s.* is larger than 1 . The proof of Lemma 2.6 is detailed in Appendix B.4, and it was known in the literature (*e.g.* Proposition 3 in Qin and You [2023]).

Lemma 2.6. *An allocation w is optimal for $T^\star(\nu)^{-1}$ if and only if it satisfies*

$$\begin{aligned} &\textit{Information balance:} && \forall i \neq i^\star,\ C(i^\star, i; \nu, w) = T^\star(\nu)^{-1}\,, \\ &\textit{Overall balance:} && \sum_{i \neq i^\star} \frac{\mathcal{K}_{\inf}^{-}(\nu_{i^\star}, u_{i^\star,i}(\nu, w))}{\mathcal{K}_{\inf}^{+}(\nu_i, u_{i^\star,i}(\nu, w))} = 1\,, \end{aligned}$$

where $u_{i^\star,i}(\nu, w)$ is defined in Lemma 2.4.

Recall that the dual variable γ_i in Lemma 2.5 represents the conditional probability of selecting arm i as challenger given that the leader is $i^\star$. The proof of Lemma 2.5 implies that it is inversely proportional to $\frac{\partial C(i^\star,i;\nu,w)}{\partial w_i}$.

Compared to IDS, BOLD is independent of the challenger answer. Compared to IDS, BOLD is a binary decision (*i.e.* $\beta_n(B_n, C_n) \in \{0, 1\}$). This makes the proof of sufficient exploration harder (see Section 2.3.3), and the generalization to linear bandits less clear (see Section 8.2.1 of Chapter 8).

Recently, Bandyopadhyay et al. [2024] show that combining the leader/challenger pair EB-TC (or EB-TCI) with the optimal design BOLD yields an asymptotically optimal BAI algorithm for any one-parameter exponential family of distributions $\mathcal{D}_{\exp}$. Their algorithms

(AT2 and IAT2) relies on an additional polynomial forced exploration step. Compared to other lower bound based algorithms (see Section 1.4.3), the Top Two approach is amenable to analysis without a forced exploration step since it ensures sufficient exploration provided some properties hold on the leader, the challenger and the target (see Section 2.3.3). We conjecture that the BOLD target also satisfies such a property. When studying the Top Two approach (without forced exploration) for one-parameter exponential family [Jourdan et al., 2022], an assumption on the tail distributions (*e.g.* sub-exponential distributions) is needed to prove sufficient exploration.

2.2.4 Mechanism to Reach the Target Allocation

Equipped with a target $\beta_n(B_n, C_n) \in [0,1]$ for the leader, we need a mechanism to reach this target. When $\beta_n(B_n, C_n) \in \{0,1\}$, one can simply sample the corresponding arm. When $\beta_n(B_n, C_n) \in (0,1)$, one needs to convert a proportion $\beta_n(B_n, C_n) \in (0,1)$ into a decision to pull either $I_n = B_n$ or $I_n = C_n$.

Randomized approach Historically, the TTTS algorithm used randomization to make this choice. Namely, it will set $I_n = B_n$ with probability $\beta_n(B_n, C_n)$, and set $I_n = C_n$ otherwise, *i.e.* $\mathbb{P}_{|n}(I_n = i \mid (B_n, C_n) = (i,j)) = \beta_n(i,j)$ and $\mathbb{P}_{|n}(I_n = j \mid (B_n, C_n) = (i,j)) = 1 - \beta_n(i,j)$. Therefore, we have

$$\mathbb{P}_{|n}(I_n = i) = \sum_{j \neq i} \beta_n(i,j)\mathbb{P}_{|n}((B_n, C_n) = (i,j)) + \sum_{j \neq i}(1 - \beta_n(j,i))\mathbb{P}_{|n}((B_n, C_n) = (j,i)) \,. \quad (2.17)$$

Tracking approach Inspired by the success of C-Tracking Garivier and Kaufmann [2016] for Track-and-Stop, in Jourdan and Degenne [2023] we replace this randomization step by a tracking step. Given that the target allocation is defined conditioned on the leader/challenger pair, we use $K(K-1)$ independent tracking procedures, *i.e.* one per pair $(i,j) \in [K]^2$ such that $i \neq j$. As we will see, the tracking procedure is relevant mostly when the leader/challenger pair is deterministic. We do not study tracking for randomized leader/challenger choices.

All the following notation use the event that the leader/challenger pair is (i,j) , which is obtained by a deterministic mechanism (*e.g.* EB-TCI). Let $T_n(i,j) := \sum_{t \in [n-1]} \mathbb{1}\left((B_t, C_t) = (i,j)\right)$ be the number of times it occurs, the averaged target allocation of the leader be $\bar{\beta}_n(i,j) := T_n(i,j)^{-1} \sum_{t \in [n-1]} \mathbb{1}\left((B_t, C_t) = (i,j)\right) \beta_t(i,j)$ and $N^i_{n,j} := \sum_{t \in [n-1]} \mathbb{1}\left((B_t, C_t) = (i,j),\ I_t = j\right)$ be the number of pulls of arm j when i is the leader. Then,

$$I_n = C_n \quad \text{if} \quad N^{B_n}_{n,C_n} \leq (1 - \bar{\beta}_{n+1}(B_n, C_n))T_{n+1}(B_n, C_n), \quad \text{and} \quad I_n = B_n \quad \text{otherwise}. \quad (2.18)$$

Using Theorem 6 in Degenne et al. [2020b] for each tracking procedure yields Lemma 2.7.

Lemma 2.7. *Using the $K(K-1)$ tracking procedures as in (2.18), we obtain that $-1/2 \leq N^i_{n,j} - (1-\bar{\beta}_n(i,j))T_n(i,j) \leq 1$ for all $n > K$, all $i \in [K]$, all $j \neq i$.*

In Jourdan et al. [2023b], we proposed the tracking procedure in (2.18) to cope for an adaptive target which depends on $\{B_n, C_n\}$. When the target is agnostic to the leader/challenger pair, *i.e.* $\beta_n(B_n, C_n) = \beta$, it is sufficient to define K tracking procedure, *i.e.* one per possible leader. In Jourdan and Degenne [2023], we set

$$I_n = B_n \quad \text{if} \quad N^{B_n}_{n,B_n} \leq \beta \sum_{i \neq B_n} T_{n+1}(B_n, i)\,, \quad \text{and} \quad I_n = C_n \quad \text{otherwise}\,. \tag{2.19}$$

Similarly, using Theorem 6 in Degenne et al. [2020b], we obtain Lemma 2.8.

Lemma 2.8. *Using the K tracking procedures as in (2.19), we obtain that $-1/2 \leq N^i_{n,i} - \beta \sum_{j \neq i} T_n(i,j) \leq 1$ for all $n > K$ and all $i \in [K]$.*

Deterministic or randomized Among the four choices to define a Top Two sampling rule, the mechanism to reach the target allocation has the least impact by far. Empirically, we observe very little difference when using one over the other. At the exception of a fully deterministic Top Two algorithm, I_n is a random variable due to the randomness in B_n , C_n or the random selection between both. It is direct to show that $N_{n,i} - \mathbb{E}[N_{n,i}]$ is a martingale with mean 0 and sub-Gaussian increments and, with high probability, $\|N_n - \mathbb{E}[N_n]\|_\infty = \tilde{\mathcal{O}}(\sqrt{n})$.

Using a fully deterministic Top Two sampling rule over a randomized one is motivated by practical and theoretical reasons. First, the practitioner might be only willing to use a deterministic algorithm. While this holds for some specific applications, this is not the case in clinical trials. Second, in the analysis, it is easier to control deterministic counts since it removes the need for martingales arguments to cope for the randomness of the algorithm itself (*i.e.* random counts). Therefore, it simplifies the non-asymptotic analysis.

2.2.5 Naming Convention

Since a Top Two algorithm is defined by four choices and each one has several possibilities, there is a combinatorial number of possible instances. To ease the understanding, we propose to use `{leader}-{challenger}-{target}` as naming convention, which extends the ones from Jourdan et al. [2022] and You et al. [2023]. Importantly, the fourth choice is not mentioned since it is a

direct consequence of the three previous choices. Namely, we use randomization when either the leader or the challenger is randomized. Otherwise, we use the tracking (2.19) for fixed design β , and the tracking (2.18) for optimal design IDS. To refer to a pair of leader/challenger answers (regardless of the target considered), we use `{leader}-{challenger}` as naming convention.

To familiarize the reader with this naming convention, we detail some examples. The greedy GLR-based sampling rule from Section 2.1.2 refers to EB-TC. As regards the Top Two algorithms which were introduced with a different naming convention, we have to be name a few: LUCB [Kalyanakrishnan et al., 2012] is (almost) EB-UCB-$1/2$, TTTS [Russo, 2016] is TS-RS-β , TTPS [Russo, 2016] is PS-PS-β , TTEI [Qin et al., 2017] is EI-EI-β , T3C [Shang et al., 2020] is TS-TC-β , TTUCB [Jourdan and Degenne, 2023] is UCB-TC-β , AT2 (resp. IAT2) [Bandyopadhyay et al., 2024] is EB-TC-BOLD (resp. EB-TCI-BOLD) with forced exploration.

2.2.6 A Simple, Interpretable and Generalizable Approach

There are many reasons behind the success of Top Two algorithms. For statisticians, the Top Two algorithms are appealing since they have simultaneously good theoretical guarantees and good empirical performance. Throughout this thesis, we will illustrate those two characteristics with theorems and experiments. For practitioners, the Top Two algorithms are attractive since they are simple, interpretable, generalizable, and versatile. While the versatility of this approach will be made clearer in Section 5.4 of Chapter 5, we can already motivate the first three characteristics.

Simple The Top Two approach (see Algorithm 2.1) is simply defined by four choices having a simple intuition. While the leader answer aims at identifying $i^\star$, the challenger answer attempts to confuses the belief of the agent that the leader is truly $i^\star$. The target corresponds to the fraction of samples we want to allocate to the leader answer. Then, we need a mechanism to convert this targeted proportion into an actual arm to pull among the leader and challenger.

As a direct consequence of its simplicity, the Top Two approach is also easy to implement (four functions to code). The computational (and memory) cost of Top Two approach mostly depends on the choices of the leader answer (see Table 2.1) and the challenger answer (see Table 2.2). We will not delve into the pros and cons of each choices. Importantly, we note that the EB-TC algorithm has no additional computational cost since the EB leader is the recommendation rule and the TC challenger is already computed by the GLR stopping rule. However, Bayesian approaches require either an efficient sampler (TS leader, RS challenger when highly efficient) or explicit formulas (EI leader and EI challenger).

Interpretable Since we can grasp the role of each of those choices, the sampling strategy enforced by a Top Two approach is interpretable. While additional samples are always collected to gather more information, the status of the next arm to be pulled is clear with the Top Two approach. It is believed to be a good arm when sampling the leader, and a sub-optimal arm when sampling the challenger.

In domains fraught with ethical implications such as phase III of clinical trials, it is even possible to enforce some constraints on the sampling strategy a priori. For example, a Top Two approach with fixed design β will attempt to cure a fraction β of the volunteers. Therefore, the target β could be chosen by an ethical board a priori.

Generalizable When describing the possible instances for each choice (leader, challenger and target), we used a generic notation which is amenable to changes of the goals and assumptions with limited modifications. For example, the empirical transportation costs are easily modified when considering BAI for Gaussian with unknown variance (Chapter 3) or bounded distributions (Chapter 4). Similarly, it is direct to adapt the Top Two approach to tackle ε-BAI (Chapter 5). Finally, the Structured Top Two generalize the Top Two approach for structured settings (*e.g.* linear bandits).

2.3 Asymptotic Sample Complexity Upper Bound

In BAI for Gaussian distributions with known variance, Theorem 2.9 shows the asymptotic (β-)optimality of many Top Two sampling rules when combined with the GLR stopping rule. We present a unified asymptotic analysis of the Top Two approach, which identifies desirable properties on the choices of the leader and challenger answers to achieve asymptotic (β-)optimality. The proof is detailed in the rest of this section.

Theorem 2.9. *Let $(\beta, \delta) \in (0,1)^2$. Combined with the GLR stopping rule (2.2) using the threshold (2.3), the Top Two sampling rule (Algorithm 2.1) using (i) any leader in Table 2.1, (ii) any challenger in Table 2.2, (iii) the fixed design β or the optimal design IDS (2.11), and (iv) randomization (or tracking for a deterministic leader/challenger pair, see Section 2.2.4) yields an algorithm which is δ-correct and satisfies that, for all $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathbb{R}^K$ such that $\overline{\Delta}_{\min}(\mu) := \min_{i \neq j} |\mu_i - \mu_j| > 0$,*

$$\textbf{[IDS]} \ \limsup_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\log(1/\delta)} \leq T^\star(\nu) \quad \textit{and} \quad \textbf{[fixed } \beta\textbf{]} \ \limsup_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\log(1/\delta)} \leq T^\star_\beta(\nu) \, ,$$

Distinct means Restricting to instances such that $\overline{\Delta}_{\min}(\mu) > 0$ (which implies $|i^\star(\mu)| = 1$) is an uncommon assumption in BAI. However, prior Top Two algorithms [Russo, 2016, Qin et al., 2017, Shang et al., 2020] have been analyzed on those instances. Our generic analysis reveals that it is solely used to prove sufficient exploration (see Section 2.3.3). Experiments suggests that the greedy EB-TC is the only Top Two algorithm which has poor empirical performance when $\overline{\Delta}_{\min}(\mu) = 0$ and $|i^\star(\mu)| = 1$. For ε-BAI with $\varepsilon > 0$ (see Chapter 5), the EB-TC$_\varepsilon$ will not suffer from this issue thank to the implicit exploration of the TC$_\varepsilon$ challenger. As mentioned in Section 2.1.2, the greediness of EB-TC in BAI can be alleviated by considering randomization or optimism which foster implicit exploration. In the asymptotic analysis, we will not formalize this concept explicitly with a sufficient property on the leader or on the challenger. However, for the non-asymptotic analysis of the UCB-TC-β algorithm (see Section 2.4), the UCB leader has the property of identifying the best arm except for logarithmic number of rounds with high probability (Lemma 2.28).

Beyond Gaussian distributions Theorem 2.9 can be readily adapted for Gaussian with known (non-unit) variance $\sigma^2 > 0$. Moreover, it also holds for σ-sub-Gaussian random variables thanks to direct adaptations of concentration results. However, it is (β-)optimal in a distribution sense only for Gaussian distributions.

While it is an open problem to analyze IDS for more general distributions than Gaussian with known variance, our analysis of the fixed design β can be generalized. In Jourdan et al. [2022], we show the asymptotic β-optimality of deterministic Top Two algorithms for one-parameter exponential family of sub-exponential distributions, as well as the one of randomized Top Two algorithms for Bernoulli distributions. In Chapter 3, we prove the asymptotic β-optimality of deterministic Top Two algorithms for Gaussian distributions with unknown variance. In Chapter 4, we show the asymptotic β-optimality of deterministic and randomized Top Two algorithms for bounded distributions.

Proof outline After restating known regularities on the characteristic times (Section 2.3.1), the asymptotic analysis is split into three parts. In Section 2.3.2, we derive a sufficient condition on any sampling rule to obtain the desired asymptotic upper bound when combined with the GLR stopping rule. In Section 2.3.3, we highlight how the Top Two sampling rule is exploring the arms sufficiently. In Section 2.3.4, we show how the Top Two sampling rule is balancing the empirical allocation to ensure its converge towards the (β-)optimal allocation. In each step of the analysis, we focus on the analysis of IDS, then highlight how to cope for a fixed design β with tracking as in (2.19). The asymptotic analysis follows the unified analysis of Jourdan et al. [2022], which was inspired by Qin et al. [2017] and Shang et al. [2020], then extended by Jourdan and Degenne [2023] for tracking and by You et al. [2023] for IDS. For

notational simplicity, we omit the conditioning on the history $\mathcal{F}_n$ when defining all the values, expectations and probabilities in the rest of this section.

The randomness comes from the observations and, potentially, the use of randomization in the leader, challenger or the mechanism to reach the target. Lemma 2.10 is a standard concentration result of the empirical mean for sub-Gaussian observations, which has been used for the asymptotic analysis of Top Two algorithms (*e.g.* Lemma 3 in Qin et al. [2017], Lemma 5 in Shang et al. [2020], Lemmas 5 and 14 in Jourdan et al. [2022]) hence we omit its proof.

Lemma 2.10. *There exists two independent sub-Gaussian random variables W_μ and W_K such that almost surely, for all $i \in [K]$ and all n such that $N_{n,i} \geq 1$,*

$$|\mu_{n,i} - \mu_i| \leq W_\mu \sqrt{\frac{\log(e + N_{n,i})}{N_{n,i}}} \quad \text{and} \quad \|N_n - \mathbb{E}[N_n]\|_\infty \leq W_K \sqrt{(n+1)\log(e+n)} \, .$$

In particular, any random variable which is polynomial in (W_μ, W_K) has a finite expectation.

2.3.1 Characteristic Times

Lemma 2.11 restates some fundamental results on the characteristic time and the optimal allocation, which were shown by Russo [2016] for any one-parameter exponential family.

Lemma 2.11 ([Russo, 2016])**.** *If $i^\star(\mu)$ is a singleton and $\beta \in (0,1)$, then $w^\star(\nu)$ and $w^\star_\beta(\nu)$ are singletons,* i.e. *the optimal allocations are unique, and $w^\star(\nu)_i > 0$ and $w^\star_\beta(\nu)_i > 0$ for all $i \in [K]$. $T^\star_{1/2}(\nu) \leq 2T^\star(\nu)$ and with $\beta^\star = w^\star_{i^\star}(\nu)$, $\frac{T^\star_\beta(\nu)}{T^\star(\nu)} \leq \max\left\{\frac{\beta^\star}{\beta}, \frac{1-\beta^\star}{1-\beta}\right\}$. Moreover, for all $i \neq i^\star$,*

$$T^\star(\nu)^{-1} = C(i^\star, i; \nu, w^\star(\nu)) \quad \text{and} \quad T^\star_\beta(\nu)^{-1} = C(i^\star, i; \nu, w^\star_\beta(\nu)) \, . \tag{2.20}$$

The functions $(\kappa, w) \to C(i, j; \kappa, w)$ and $\kappa \to i^\star(m(\kappa))$ are continuous. The function $\kappa \to C(i, j; \kappa, 1_K)$ is continuous with strictly positive value for (i, j) such that $m(\kappa)_i > m(\kappa)_j$. The function $w \to C(i^\star, i; \kappa, w)$ is increasing.

Let $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathbb{R}^K$ such that $i^\star(\mu) = \{i^\star\}$. Let $\beta \in (0,1)$. Let $w^\star$ and $w^\star_\beta$ be the unique optimal allocation and β-optimal allocation, *i.e.* $w^\star(\nu) = \{w^\star\}$ and $w^\star_\beta(\nu) = \{w^\star_\beta\}$ where $w^\star(\nu)$ and $w^\star_\beta(\nu)$ are the maximizer of $T^\star(\nu)^{-1}$ and $T^\star_\beta(\nu)^{-1}$ defined in Lemma 2.1.

The property (2.20) is a necessary condition for optimality, it means that there is equality of the transportation cost at the equilibrium, hence is referred to as the *information balance* [You et al., 2023]. Another necessary condition for optimality is the *overall balance*, which means

that the allocations are balancing the contribution of each arm to the transportation cost, *i.e.* $\sum_{i \neq i^\star} \frac{\mathcal{K}_{\inf}^-(\nu_{i^\star}, u(i^\star, j, \nu, w^\star))}{\mathcal{K}_{\inf}^+(\nu_i, u(i^\star, i, \nu, w^\star))} = 1$ where $u(i^\star, j, \nu, w^\star)$ is the minimizer of $C(i^\star, i; \nu, w^\star)$. For Gaussian distribution with known variance, the overall balance has a convenient expression, which is problem-independent since it does not depend explicitly on ν , *i.e.* $(w_{i^\star}^\star)^2 = \sum_{i \neq i^\star} (w_i^\star)^2$.

2.3.2 Asymptotic (β-)Optimality

The first step of the asymptotic analysis is to show that the convergence towards the (β-)optimal allocation is a sufficient condition to obtain asymptotic (β-)optimality when using the GLR stopping rule (Lemma 2.12), as discussed in Section 1.4.2.

Convergence time Let $\gamma > 0$ and $w \in \mathring{\Sigma}_K$. Let us define the *convergence time* T_γ , which is a random variable quantifies the number of samples required for the empirical ratios of allocation $(N_{n,i}/N_{n,i^\star})_{i \neq i^\star}$ to stay γ-close to $(w_i/w_{i^\star})_{i \neq i^\star}$ forever, *i.e.*

$$T_\gamma(w) := \inf \left\{ T \geq 1 \mid \forall n \geq T, \max_{i \neq i^\star} \left| \frac{N_{n,i}}{N_{n,i^\star}} - \frac{w_i}{w_{i^\star}} \right| \leq \gamma \right\} . \tag{2.21}$$

Proven in Appendix B.5, Lemma 2.12 also holds true for a broader class of thresholds (see the *asymptotically tight* threshold defined therein), whose (n, δ) dependencies is sufficient to ensure asymptotic (β-)optimality.

Lemma 2.12. *Assume that the sampling rule satisfies that there exists $\gamma_\nu > 0$ such that: for all $\gamma \in (0, \gamma_\nu]$, $\mathbb{E}_\nu[T_\gamma(w^\star)] < +\infty$ (resp. $\mathbb{E}_\nu[T_\gamma(w_\beta^\star)] < +\infty$) with $T_\gamma(w)$ as in (2.21). Using the threshold $c(n, \delta)$ as in (2.3) in the GLR stopping rule (2.2), this sampling rule yields an algorithm such that, for all $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathbb{R}^K$ such that $|i^\star(\mu)| = 1$,*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\log(1/\delta)} \leq T^\star(\nu) \quad (\textit{resp. } T_\beta^\star(\nu)) .$$

Using Lemma 2.12, the proof of Theorem 2.9 for the optimal design IDS (resp. fixed design β) boils down to showing that $\mathbb{E}_\nu[T_\gamma(w^\star)] < +\infty$ (resp. $\mathbb{E}_\nu[T_\gamma(w_\beta^\star)] < +\infty$). In Section 2.3.3, we will show sufficient exploration for all the arms, *e.g.* $\min_{i \in [K]} N_{n,i} \geq \sqrt{n/K}$ for n large enough. In Section 2.3.4, we will prove that the expectation of the convergence time is finite.

2.3.3 Sufficient Exploration

In this section, we identify properties on the leader (Property 2.15) and the challenger (Property 2.16) under which the Top Two algorithm ensures sufficient exploration (Lemma 2.17)

$$\exists N_0 \text{ s.t. } \mathbb{E}_\nu[N_0] < +\infty,\ \forall n \geq N_0, \quad \min_{i\in[K]} N_{n,i} \geq \sqrt{n/K} \quad \text{and} \quad \arg\max_{i\in[K]} \mu_{n,i} = \{i^\star\}\,. \tag{2.22}$$

Effective leader and challenger For an algorithm to alleviate under-sampling some arms, it should have a strictly positive probability of sampling them. In Top Two algorithms, the choice of the arm to pull I_n is defined by the leader B_n and the challenger C_n . When there is randomization, it is not trivial to manipulate B_n and C_n . Therefore, we define the *effective* leader answer $\widehat{B}_n$ and the *effective* challenger answer $\widehat{C}_n$ as the answers maximizing the respective probability of being selected, *i.e.*

$$\widehat{B}_n \in \arg\max_{i\in[K]} \mathbb{P}_{|n}(B_n = i) \quad \text{and} \quad \widehat{C}_n \in \arg\max_{i\neq\widehat{B}_n} \mathbb{P}_{|n}(C_n = i | B_n = \widehat{B}_n)\,, \tag{2.23}$$

where $\widehat{C}_n$ is defined conditioned on the effective leader $\widehat{B}_n$. We assume that ties are broken uniformly at random. Note that they are fully determined by the mechanisms to choose the leader and the challenger. When the choices of the leader/challenger pair are deterministic ones, we have $(\widehat{B}_n, \widehat{C}_n) = (B_n, C_n)$ for all n .

Target allocation A *good* target allocation should ensure that the allocations are balanced between arms. Lemma 2.13 states that the target allocation of the least sampled arm in the effective leader/challenger pair is strictly positive.

Lemma 2.13. *There exists $\beta_{\min} > 0$ such that, for all $n > K$, $\beta_n(\widehat{B}_n, \widehat{C}_n) \geq \beta_{\min}$ if $N_{n,\widehat{B}_n} \leq N_{n,\widehat{C}_n}$, and $1 - \beta_n(\widehat{B}_n, \widehat{C}_n) \geq \beta_{\min}$ otherwise.*

Proof. It is direct for fixed design $\beta \in (0,1)$ by taking $\beta_{\min} = \min\{\beta, 1-\beta\}$ and for optimal design IDS by taking $\beta_{\min} = 1/2$ respectively. ■

While Lemma 2.13 holds for fixed design β and optimal design IDS, it seems difficult to prove it for the optimal design BOLD as defined in (2.16). When both the leader and the challenger are deterministic, we have $(\widehat{B}_n, \widehat{C}_n) = (B_n, C_n)$. Since $N_{n,B_n} \leq N_{n,C_n}$ implies that $N^2_{n,B_n} < \sum_{i\neq B_n} N^2_{n,i}$, the leader will be sampled if it is the least sampled arm among those two. However, the condition $N_{n,B_n} > N_{n,C_n}$ does not necessarily implies that the challenger will be sampled with a strictly positive probability. For randomized leader or challenger, the situation

is even more intricate. We conjecture that the optimal design BOLD satisfies another type of property to obtain sufficient exploration. As it does not fit perfectly in our unified analysis, we leave this question for future work. In Bandyopadhyay et al. [2024], the forced exploration step removes the need to show that the optimal design BOLD ensures sufficient exploration.

Proof strategy Given Lemma 2.13, the sufficient exploration can be proven if we show that either $\widehat{B}_n$ or $\widehat{C}_n$ is among the under-sampled arms if some still exists. Before formalizing the properties required by the leader and challenger pair to ensure sufficient exploration, we introduce the relevant notation.

Given an arbitrary threshold $L \in \mathbb{R}_+^*$, we define the sampled enough set and its arms with highest mean (when not empty) as

$$S_n^L := \{i \in [K] \mid N_{n,i} \geq L\} \quad \text{and} \quad \mathcal{I}_n^\star := \arg\max_{i \in S_n^L} \mu_i \,, \tag{2.24}$$

where $\mathcal{I}_n^\star$ is a set with potentially multiple values. At time n , S_n^L can only be non-empty for $L \leq n$, hence it depends explicitly on n . Assumption 2.14 ensures that $\mathcal{I}_n^\star$ is unique. As mentioned above, this distinct means assumption will only be used for the proof of sufficient exploration, and it could be lifted with a more detailed analysis leveraging additional properties on the leader/challenger pair (*e.g.* as for UCB-TC-β) or in ε-BAI (thanks to the TC$_\varepsilon$ challenger).

Assumption 2.14. *All the arms have distinct means*, i.e. $\overline{\Delta}_{\min}(\mu) := \min_{i \neq j} |\mu_i - \mu_j| > 0$.

To prove sufficient exploration, we aim at finding a time N_1 with $\mathbb{E}_\nu[N_1] < +\infty$ and a threshold $L(n)$ such that $S_n^{L(n)} = [K]$ for $n \geq N_1$, *i.e.* $L(n) = \sqrt{n/K}$ in (2.22). We proceed by contradiction. The idea is to show that if some arms are still highly under-sampled, then either $\widehat{B}_n$ or $\widehat{C}_n$ will be mildly under-sampled. Since they have a strictly positive probability of being sampled (Lemma B.3), this will yield a contradiction by the pigeonhole principle. We define the highly and the mildly under-sampled sets

$$U_n^L := \{i \in [K] \mid N_{n,i} < \sqrt{L}\} \quad \text{and} \quad V_n^L := \{i \in [K] \mid N_{n,i} < L^{3/4}\} \,. \tag{2.25}$$

The choice of $\sqrt{L}$ and $L^{3/4}$ is arbitrary, and we could take L^{α_1} and L^{α_2} with $0 < \alpha_1 < \alpha_2 < 1$.

Leader answer A *good* leader should first identify the best arm among the arms that are sampled enough (Property 2.15). Property 2.15 states that if $\widehat{B}_n$ is sampled enough, then $\widehat{B}_n$ is an arm with highest mean among the sampled enough arms.

Property 2.15. *There exists L_0 with $\mathbb{E}_\nu[(L_0)^\alpha] < +\infty$ for all $\alpha > 0$ such that if $L \geq L_0$, for all n such that $S_n^L \neq \emptyset$, $\widehat{B}_n \in S_n^L$ implies $\widehat{B}_n \in \mathcal{I}_n^\star$.*

Challenger answer Given a good leader, a *good* challenger should enforce exploration on the arms that are not sampled enough yet when the leader does not do it already (Property 2.16). Property 2.16 states that if some arms are still highly under-sampled, *i.e.* $U_n^L \neq \emptyset$, then having sampled $\widehat{B}_n$ enough implies that $\widehat{C}_n$ is mildly under-sampled or has highest true mean among the sampled enough arms. In Appendix B.13, we show Property 2.16 for the TC challenger, and explain why it also holds for other choices of challenger answer as in Table 2.2.

Property 2.16. *Let $\widehat{B}_n$ be an effective leader satisfying Property 2.15 and $\widehat{C}_n$ the associated effective challenger. Let $\mathcal{J}_n^\star = \arg\max_{i \in \overline{V_n^L}} \mu_i$. There exists L_1 with $\mathbb{E}_\nu[L_1] < +\infty$ such that if $L \geq L_1$, for all n such that $U_n^L \neq \emptyset$, $\widehat{B}_n \notin V_n^L$ implies $\widehat{C}_n \in V_n^L \cup \left(\mathcal{J}_n^\star \setminus \left\{\widehat{B}_n\right\}\right)$.*

Lemma 2.17 shows (2.22), *i.e.* all arms are sufficiently explored for n large enough.

Lemma 2.17. *Suppose that Assumption 2.14 holds. Using randomization (or tracking for a deterministic leader/challenger, see Section 2.2.4), a Top Two algorithm with leader and challenger satisfying Properties 2.15 and 2.16 is such that there exists N_0 with $\mathbb{E}_\nu[N_0] < +\infty$ such that, for all $n \geq N_0$, $\min_{i \in [K]} N_{n,i} \geq \sqrt{n/K}$ and $\arg\max_{i \in [K]} \mu_{n,i} = \{i^\star\}$.*

The proof of Lemma 2.17 is detailed in Appendix B.6, and the intuition is sketched below.

Proof. Suppose towards contradiction that there are still arms which are sampled less than $\sqrt{n/K}$. Combining Properties 2.15 and 2.16, we obtain that either the effective leader or the effective challenger is still an undersampled arm. Using Lemma 2.13, we know that there is a strictly positive probability of sampling the least pulled arm among those two arms. Therefore, after a large enough time, this undersampled arm will be sampled enough. Repeating this argument, it is possible to show that all arms are sampled enough, otherwise we will have a contradiction. ■

2.3.4 Convergence Towards the (β-)Optimal Allocation

In this section, we identify properties on the leader (Property 2.18) and the challenger (Property 2.21) under which we prove that the Top Two algorithm ensures that the convergence time

have finite expectation (Lemma 2.22), *i.e.*

$$\exists \gamma_\nu > 0,\ \forall \gamma \in (0, \gamma_\nu], \quad \mathbb{E}_\nu[T_\gamma(w^\star)] < +\infty \quad (\text{resp. } \mathbb{E}_\nu[T_\gamma(w^\star_\beta)] < +\infty\ ) \ , \tag{2.26}$$

with $T_\gamma(w)$ as in (2.21). In the following, we suppose that (2.22) holds, *i.e.* all arms are sufficiently explored for n large enough. While Lemma 2.17 shows it under certain conditions, we "forget" them to highlight specific properties and show that Assumption 2.14 is not needed once we have sufficient exploration. Since there is a unique best arm $i^\star$, (2.22) implies that $\arg\max_{i \in S_n^{\sqrt{n/K}}} \mu_i = \{i^\star\}$ for all $n \geq N_0$, with S_n^L as in (2.24) and N_0 as in (2.22).

Leader answer A *good* leader should simply identify the best arm $i^\star$. Property 2.18 states that the probability that the leader is not the unique best arm vanishes. In Appendix B.12, we show Property 2.18 for the EB leader, and explain why it also holds for other choices of leader answer as in Table 2.1.

Property 2.18. *Suppose that* (2.22) *holds. Let* g_1 *be a function* s.t. $\sum_{t \in [n-1]} g_1(t) =_{+\infty} \mathcal{O}(n^{1-\alpha_1})$ *with* $\alpha_1 > 0$ *. There exists* N_1 *with* $\mathbb{E}_\nu[N_1] < +\infty$ *such that, for all* $n \geq N_1$ *,* $\mathbb{P}_{|n}(B_n \neq i^\star) \leq g_1(n)$ *.*

Target allocation A *good* target allocation should ensure that the allocations are (β-)optimaly balanced between arms. For the optimal design IDS, it means that the overall balance equation is approximately satisfied by the empirical allocation.

Lemma 2.19. *Suppose that* (2.22) *holds. Assume that the leader satisfies Property 2.18. Let* $\gamma > 0$ *. There exists* N_2 *with* $\mathbb{E}_\nu[N_2] < +\infty$ *such that, for all* $n \geq N_2$ *,*

$$\textbf{[IDS]} \quad \frac{1}{(n-1)^2} \left| N_{n,i^\star}^2 - \sum_{j \neq i^\star} N_{n,j}^2 \right| \leq \gamma \quad \textit{and} \quad \textbf{[fixed } \beta\textbf{]} \quad |N_{n,i^\star}/(n-1) - \beta| \leq \gamma \ .$$

For fixed design $\beta \in (0,1)$, it is straightforward to see that Lemma 2.19 holds true both when using randomization (or tracking for a deterministic leader/challenger). For the optimal design IDS defined in (2.13), we can also show that Lemma 2.13 holds true when using tracking or randomization. Intuitively, the property is satisfied if the increments are small enough for their summation to be small enough. Using informally the differential d notation, we have

$\mathrm{d}N_{n,i^\star} = \beta_n(i^\star, C_n)$ and $\mathrm{d}N_{n,j} = \mathbb{1}(j = C_n)(1 - \beta_n(i^\star, C_n))$. Therefore, we obtain

$$\mathrm{d}\left(N_{n,i^\star}^2 - \sum_{j \neq i^\star} N_{n,j}^2\right) = 2N_{n,i^\star}\beta_n(i^\star, C_n) - 2N_{n,C_n}(1 - \beta_n(i^\star, C_n)) = 0\,,$$

where the last equality holds since $\beta_n(i^\star, j) = N_{n,j}/(N_{n,i^\star} + N_{n,j})$ as $\mu_{n,i^\star} > \max_{i \neq i^\star} \mu_{n,i}$. Summing those small increments will yield the result. The proof of Lemma 2.19 is detailed in Appendix B.7 for randomization (or tracking for a deterministic leader/challenger, see Section 2.2.4).

The optimal design BOLD in (2.16) is precisely defined to ensure that $|N_{n,i^\star}^2 - \sum_{j \neq i^\star} N_{n,j}^2|$ remains small. Therefore, it is straightforward to show that the property of IDS in Lemma 2.19 also holds for BOLD. The convergence towards the optimal allocation when using BOLD has been shown in Bandyopadhyay et al. [2024]. In addition of holding for any one-parameter exponential family of distributions, their analysis greatly differs from the one presented in this thesis. The asymptotic path followed by the algorithm is described by a series of ordinary differential equations satisfied by a limiting fluid dynamics of the allocations.

Lemma 2.20 is a corollary of Lemma 2.19: the best arm is sampled linearly, and the reweighted overall balance equation is approximately satisfied empirically when using the optimal design IDS. The proof of Lemma 2.20 is detailed in Appendix B.8.

Lemma 2.20. *Suppose that* (2.22) *holds. Let* $\gamma > 0$. *Using optimal design IDS, there exists* N_4 *with* $\mathbb{E}_\nu[N_4] < +\infty$ *such that for all* $n \geq N_4$, $(4\sqrt{2(K-1)})^{-1} \leq N_{n,i^\star}/(n-1) \leq 3/4$ *and* $|1 - \sum_{i \neq i^\star} (N_{n,i}/N_{n,i^\star})^2| \leq \gamma$.

Challenger answer Given a good leader, a *good* challenger should balance each sub-optimal arm with the (β-)optimal allocation. In other words, when the ratio of their empirical proportions exceeds the ratio of their (β-)optimal allocation, this arm should have a small probability of being sampled again (Property 2.21). In Appendix B.13, we show Property 2.21 for the TC challenger, and explain why it also holds for other choices of challenger answer as in Table 2.2.

Property 2.21. *Suppose that* (2.22) *holds. Assume that the leader satisfies Property 2.18. Let* $\gamma \in (0, \gamma_0]$ *where* $\gamma_0 > 0$ *is a problem dependent constant. Let* g_3 *be a function* s.t. *such that* $\sum_{t \in [n-1]} g_3(t) =_{+\infty} o(n^{1-\alpha_3})$ *with* $\alpha_3 > 0$. *There exists* N_3 *with* $\mathbb{E}_\nu[N_3] < +\infty$ *such that, for all* $n \geq N_3$ *and all* $i \neq i^\star$,

$$\frac{N_{n,i}}{N_{n,i^\star}} \geq \gamma + \begin{cases} w_i^\star/w_{i^\star}^\star & \textbf{[IDS]} \\ w_{\beta,i}^\star/w_{\beta,i^\star}^\star & \textbf{[fixed } \beta\textbf{]} \end{cases} \quad \Longrightarrow \quad \mathbb{P}_{|n}(C_n = i \mid B_n = i^\star) \leq g_3(n)\,. \tag{2.27}$$

Lemma 2.22 shows (2.26), *i.e.* the convergence time have finite expectation.

Lemma 2.22. *Suppose that* (2.22) *holds. Using randomization* (*or tracking for a deterministic leader/challenger, see Section 2.2.4*)*, a Top Two algorithm with leader and challenger satisfying Properties 2.18 and 2.21 is such that there exists* $\gamma_\nu > 0$ *such that, for all* $\gamma \in (0, \gamma_\nu]$ *,*

$$\textbf{[IDS]}\ \mathbb{E}_\nu[T_\gamma(w^\star)] < +\infty \quad \textit{and} \quad \textbf{[fixed } \beta\textbf{]}\ \mathbb{E}_\nu[T_\gamma(w^\star_\beta)] < +\infty\,,$$

The proof of Lemma 2.22 is detailed in Appendix B.9, and the intuition is sketched below.

Proof. Using Property 2.18, we know that the leader will be $i^\star$. Using Property 2.21, we also know that an arm $i \neq i^\star$ will not be chosen as challenger if the ratio of the empirical allocation $N_{n,i}/N_{n,i^\star}$ exceeds the ratio of the (resp. β-)optimal allocation (by some constant). Since it is not a challenger anymore, it will not be sampled, hence the empirical ratio will decrease again. Leveraging Lemma 2.19, the allocation of the arms $i \neq i^\star$ will balance themselves (β-)optimally for n large enough. Therefore, the convergence time have finite expectation. ■

Suppose that Assumption 2.14 holds, combining Lemmas 2.12, 2.17 and 2.22, we obtain Theorem 2.9 for a Top Two algorithm with leader satisfying Properties 2.15 and 2.18 and challenger satisfying Properties 2.16 and 2.21.

2.3.5 A Pedagogical Example: EB-TC

In the unified analysis presented above, the proofs were only done for two choices of the target allocation, *i.e.* fixed β and optimal design IDS, and the randomized and the tracking approach (for deterministic leader/challenger pair) to reach the target. When studying a specific Top Two algorithms, it remains to show that the leader satisfies Properties 2.15 and 2.18 and that the challenger satisfies Properties 2.16 and 2.21. For the sake of space, we will not detail the analysis of each possible choices of the leader and challenger as detailed in Tables 2.1 and 2.2. We present the proof for the EB leader and the TC challenger, and refer to Appendix D in Jourdan et al. [2022] for more details on the other choices.

EB leader answer Using concentration arguments (Lemma 2.10), it is straightforward to show that the EB leader answer will be the arm with the highest true mean among arms that are sampled enough (*i.e.* Property 2.15 holds true). Provided that (2.22) holds, the EB leader

answer will be $i^\star$, hence $\mathbb{P}_{|n}(B_n \neq i^\star) = 0$ (*i.e.* Property 2.18 holds true). Both proofs are detailed in Appendix B.12.

Other leader answers Given the formulas of the UCB leader in (2.4) and the EI leader in (2.6), it is straightforward to show that Properties 2.15 and 2.18 also hold for those leader answers. For randomized leaders, the effective TS leader answer matches the effective PS leader hence the proof is the same, *i.e.* $\arg\max_i \mathbb{P}_{|n}(B_n^{TS} = i) = \arg\max_i \mathbb{P}_{\theta_n \sim \Pi_n|n}(i \in i^\star(\theta_n)) = \arg\max_i \mathbb{P}_{|n}(B_n^{PS} = i)$. Let $i \neq j$, then $\mathbb{P}_{\theta_n \sim \Pi_n|n}(i \in i^\star(\theta_n)) \leq \mathbb{P}_{\theta_n \sim \Pi_n|n}(\theta_{n,i} \geq \theta_{n,j})$. For Gaussian distributions, we need to use the following concentration result (*e.g.* Qin et al. [2017])

$$\mathbb{P}_{\theta_n \sim \Pi_n|n}(\theta_{n,i} \geq \theta_{n,j}) \leq \exp(-W_n(i,j))/2 \quad \text{when} \quad \mu_{n,j} > \mu_{n,i} \,. \tag{2.28}$$

This allows to conclude that Properties 2.15 and 2.18 hold for the TS and PS leader answers.

TC challenger answer The proof of Property 2.15 relies on comparing the rate of growth of the empirical transportation costs. Given two arms (i, j) which are sampled enough and satisfies that $\mu_i > \mu_j$, the empirical transportation cost is strictly positive and increases linearly (Lemma 2.23 proved in Appendix B.10).

Lemma 2.23. *Let S_n^L and $\mathcal{I}_n^\star$ as in (2.24). There exists L_4 with $\mathbb{E}_\nu[(L_4)^\alpha] < +\infty$ for all $\alpha > 0$ such that, for all $L \geq L_4$ and all n with $S_n^L \neq \emptyset$, $W_n(i,j) \geq LC_\nu$ for all $(i,j) \in \mathcal{I}_n^\star \times \left(S_n^L \setminus \mathcal{I}_n^\star\right)$, where $C_\nu > 0$ is a problem dependent constant.*

Given two arms (i, j) such that only arm i is sampled enough, the empirical transportation cost is linearly upper bounded (Lemma 2.24 proved in Appendix B.11).

Lemma 2.24. *Let S_n^L as in (2.24). There exists L_5 with $\mathbb{E}_\nu[(L_5)^\alpha] < +\infty$ for all $\alpha > 0$ such that for all $L \geq L_5$ and all $n \in \mathbb{N}$, $W_n(i,j) \leq L(D_\nu + D_0 W_\mu)^2$ for all $(i,j) \in S_n^L \times \overline{S_n^L}$, where $D_\nu > 0$ (resp. $D_0 > 0$)is a problem (resp. in)dependent constant and W_μ as in Lemma 2.10.*

Combining Lemmas 2.23 and 2.24, it is direct to see that all the empirical transportation costs between two sampled enough arms will exceed the any empirical transportation costs between a sampled enough leader and an undersampled challenger. Therefore, using Property 2.15, the TC challenger answer will be an undersampled when the leader is sampled enough (*i.e.* Property 2.15 holds).

To prove Property 2.21, we will compare empirical transportation costs more precisely. We only sketch the proof of the optimal design IDS (similar proof for fixed design β). Let i be

an arm such that $N_{n,i}/N_{n,i^\star} \geq \gamma + w_i^\star/w_{i^\star}^\star$. Using Lemma 2.19, there exists an arm $j \notin \{i^\star, i\}$ such that $N_{n,j}/N_{n,i^\star} \leq w_j^\star/w_{i^\star}^\star$. Using that $w \to C(i^\star, i; \kappa, w)$ is increasing and the equality at equilibrium (Lemma 2.11), we obtain that

$$\frac{W_n(i^\star, i)}{W_n(i^\star, j)} \geq \left(\frac{\mu_{n,i^\star} - \mu_{n,i}}{\mu_{i^\star} - \mu_i} \frac{\mu_{i^\star} - \mu_j}{\mu_{n,i^\star} - \mu_{n,j}}\right)^2 \frac{1 + w_{i^\star}^\star/w_i^\star}{1 + (w_i^\star/w_{i^\star}^\star + \gamma)^{-1}} \underset{n\to+\infty}{\approx} \frac{1 + w_{i^\star}^\star/w_i^\star}{1 + (w_i^\star/w_{i^\star}^\star + \gamma)^{-1}} > 1 \,.$$

Since $W_n(i^\star, i) > W_n(i^\star, j)$, we conclude that $C_n \neq i$ (*i.e.* Property 2.21 holds true).

Both proofs are detailed in Appendix B.13.

Other challenger answers Given the formulas of the TCI challenger, the KKT challenger and the EI challenger in (2.8), it is straightforward to show that Properties 2.16 and 2.21 also hold for those challenger answers. For randomized leaders, the effective RS challenger answer matches the effective PS leader hence the proof is the same, *i.e.* $\arg\max_{i \neq \widehat{B}_n} \mathbb{P}_{|n}(C_n^{RS} = i \mid B_n = \widehat{B}_n) = \arg\max_{i \neq \widehat{B}_n} \mathbb{P}_{\theta_n \sim \Pi_n|n}(i \in i^\star(\theta_n)) = \arg\max_{i \neq \widehat{B}_n} \mathbb{P}_{|n}(C_n^{PS} = i \mid B_n = \widehat{B}_n)$. While the effective PPS challenger can be different, the proof is also similar. The proof of Property 2.16 relies on (2.28), and on a coarse anti-concentration result, *i.e.* $\mathbb{P}_{\Pi_n|n}(J_n \in i^\star(\theta_n)) \geq \mathbb{P}_{\Pi_n|n}(\theta_{n,J_n} \geq u_n)/2^{K-1}$ for a well chosen u_n and J_n . The proof of Property 2.21 relies on (2.28), and on a tight anti-concentration result, *i.e.* $\mathbb{P}_{\theta_n \sim \Pi_n|n}(\theta_{n,i} \geq \theta_{n,i^\star}) \gtrapprox \exp(-W_n(i^\star, i))$.

2.4 Non-asymptotic Sample Complexity Upper Bound

While the literature provides a detailed understanding of the asymptotic regime, many interesting questions are unanswered in the non-asymptotic regime. Recent works [Chen et al., 2017c, Simchowitz et al., 2017, Mason et al., 2020, Marjani et al., 2022] have shown that the sample complexity is affected by large moderate confidence terms (independent of δ). The asymptotic analysis presented above applies to EB-TC algorithm whose empirical stopping times is order of magnitude larger than its competitors for $\delta = 0.01$. Since the proof of asymptotic optimality hides design flaws, non-asymptotic guarantees should be derived to understand which Top Two algorithms will perform well in practice for any reasonable choice of δ (not necessarily close to 0). We tackle this problem in this section.

The UCB-TC-β algorithm (see Algorithm 2.2) is a specific instance of Algorithm 2.1 which uses the GLR stopping rule and combines the UCB leader, the TC challenger, the fixed β target and K tracking procedures as in (2.19) to reach it. In Jourdan and Degenne [2023], we proposed TTUCB which is the first fully deterministic Top Two algorithm. Adopting the naming convention from Section 2.2.5, TTUCB is exactly UCB-TC-β.

Input: $(\beta, \delta) \in (0,1)^2$, threshold $c : \mathbb{N} \times (0,1) \to \mathbb{R}^+$ and function $g : \mathbb{N} \to \mathbb{R}^+$.
Pull once each arm $i \in [K]$; **for** $n > K$ **do**
  Set $\hat{\imath}_n \in \arg\max_{i \in [K]} \mu_{n,i}$; // Candidate answer
  If $\min_{i \neq \hat{\imath}_n} \frac{\mu_{n,\hat{\imath}_n} - \mu_{n,i}}{\sqrt{1/N_{n,\hat{\imath}_n} + 1/N_{n,i}}} \geq \sqrt{2c(n-1,\delta)}$ **then return** $\hat{\imath}_n$; // GLR stopping
  Set $B_n \in \arg\max_{i \in [K]} \left\{ \mu_{n,i} + \sqrt{g(n)/N_{n,i}} \right\}$; // Leader
  Set $C_n \in \arg\min_{i \neq B_n} \frac{(\mu_{n,B_n} - \mu_{n,i})_+}{\sqrt{1/N_{n,B_n} + 1/N_{n,i}}}$; // Challenger
  Set $I_n = B_n$ if $N^{B_n}_{n,B_n} \leq \beta L_{n+1,B_n}$, otherwise $I_n = C_n$; // Tracking
  Pull I_n , observe X_{n,I_n} and update (μ_n, N_n) ;
end for

Algorithm 2.2: UCB-TC-β (or TTUCB) algorithm.

In Section 2.3, the guarantees hold only for arms having distinct means. Moreover, an asymptotic result provides no guarantees on the performance in moderate regime of δ . We address those two limitations. We prove a non-asymptotic upper bound on the expected sample complexity holding for any instance having a unique best arm.

Bonus for the UCB leader For notational simplicity, we denote by $L_{n,i} = \sum_{j \neq i} T_n(i,j)$ the number of times arm i was leader. Let $\alpha > 1$ and $s > 1$ be two concentration parameters. The choice of $g(n)$ in the UCB leader should ensure that we have an upper confidence bound on μ_i holding with high probability: with probability $1 - Kn^{-s}$, for all $t \in [n^{1/\alpha}, n]$ and all arms $i \in [K]$, $\mu_i \in [\mu_{t,i} \pm \sqrt{g(t)/N_{t,i}}]$. For Gaussian distribution with unit variance, a function g which is sufficient for the purpose of our proof can be obtained by a union bound over time, giving $g_u(n) = 2\alpha(1+s)\log n$. We can improve on g_u with mixtures of martingales, yielding $g_m(n) = \overline{W}_{-1}\left(2s\alpha \log(n) + 2\log(2 + \alpha \log n) + 2\right)$ with $\overline{W}_{-1}(x) = -W_{-1}(-e^{-x})$ for all $x \geq 1$, where W_{-1} is the negative branch of the Lambert W function, and $\overline{W}_{-1}(x) \approx x + \log(x)$ (see Appendix A). A UCB leader with $g_0(n) = 0$ recovers the Empirical Best (EB) leader. Choosing g is central for empirical performance and non-asymptotic guarantees, but not for asymptotic ones. Lower g yields better empirical performance since larger g are more conservative. In our experiments with $\alpha = s = 1.2$, we use g_m since $g_m(n) \leq g_u(n)$ for $n \geq 50$.

Non-asymptotic upper bound Theorem 2.25 gives an upper bound on the expected sample complexity holding for any δ and any instance having a unique best arm. The proof of Theorem 2.25 is sketched in Section 2.4.1, and we refer the reader to Appendix D in Jourdan and Degenne [2023] for more details.

Theorem 2.25. *Let $\delta \in (0,1)$. Using the threshold $c(n,\delta)$ as in (2.3) in the GLR stopping rule (2.2) and g_u with $s = \alpha = 1.2$, the UCB-TC-1/2 algorithm satisfies that, for all $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathbb{R}^K$ such that $|i^\star(\mu)| = 1$,*

$$\mathbb{E}_\nu[\tau_\delta] \leq \inf_{(w_0,\varepsilon)\in[0,(K-1)^{-1}]\times(0,1]} \max\left\{T_\varepsilon(\delta,w_0), C_\mu^{1.2}, C_\varepsilon(w_0)^6, (2/\varepsilon)^{1.2}\right\} + 12K\,,$$

where $C_\mu = h_1\left(26H(\mu)\right)$, $C_\varepsilon(w_0) = 2/(\varepsilon a_\nu(w_0)) + 1$,

$$T_\varepsilon(\delta,w_0) = \sup\left\{n \mid n-1 \leq 2T^\star_{1/2}(\nu)(1+\varepsilon)^2(1-w_0)^{-d_\nu(w_0)}(\sqrt{c(n-1,\delta)} + \sqrt{4\log n})^2\right\}\,,$$

with $a_\nu(w_0) = (1-w_0)^{d_\nu(w_0)} \max\{\min_{i\neq i^\star(\mu)} w^\star_{1/2}(\nu)_i, w_0/2\}$ and $d_\nu(w_0) = |\{i \neq i^\star(\mu) \mid w^\star_{1/2}(\nu)_i < w_0/2\}|$. The function $h_1(x) := x\overline{W}_{-1}\left(\log(x) + \frac{2+2K}{x}\right)$ is positive, increasing for $x \geq 2+2K$, and satisfies $h_1(x) \approx x(\log x + \log\log x)$.

The UCB-TC-1/2 sampling rule using g_m satisfies a similar upper bound. Since Theorem 2.25 holds for any instance having a unique best arm, this corroborates the intuition that assuming $\min_{i\neq j}|\mu_i - \mu_j| > 0$ is an artifact of the existing proof to obtain asymptotic β-optimality.

The upper bound on $\mathbb{E}_\nu[\tau_\delta]$ involves several terms. The δ-dependent term is $T_\varepsilon(\delta, w_0)$. In the asymptotic regime, we can show that $\limsup_{\delta\to 0} T_0(\delta,0)/\log(1/\delta) \leq 2T^\star_{1/2}(\nu)$, *i.e.* taking $w_0 = 0$ and letting ε go to zero. While there is (sub-optimal) factor 2 in $T_0(\delta,0)$, Section 2.3 shows that UCB-TC-1/2 is asymptotically $1/2$-optimal. This factor is a price we paid to obtain more explicit non-asymptotic terms, and removing it would require more sophisticated arguments in order to control the convergence of the empirical proportions $N_n/(n-1)$ towards $w^\star_{1/2}(\nu)$.

In the regime where $H(\mu) \to +\infty$, the upper bound is dominated by the δ-independent term $C_\mu^{1.2}$ (when $\alpha = 1.2$) with satisfies $C_\mu = \mathcal{O}(H(\mu)\log H(\mu))$. Compared to the best known upper and lower bounds in this regime (see discussion below), our non-asymptotic term has a sub-optimal scaling in $\mathcal{O}((H(\mu)\log H(\mu))^\alpha)$ with $\alpha > 1$. While taking $\alpha \approx 1$ would mitigate this sub-optimality, it would yield a larger dependency in $C_\varepsilon(w_0)^{\alpha/(\alpha-1)}$. Empirically, Figures 2.2(b) hints that the empirical performance of UCB-TC-1/2 has a better scaling with $H(\mu)$ than $H(\mu)^\alpha$.

For instances such that $\min_{i\neq i^\star} w^\star_{1/2}(\nu)_i$ is arbitrarily small, taking $w_0 = 0$ yields an arbitrarily large $C_\varepsilon(0)$. By clipping with $w_0/2 > 0$, we circumvent this pitfall and ensure that $C_\varepsilon(w_0) = \mathcal{O}(K/\varepsilon)$. Since it yields a larger $T_\varepsilon(\delta,w_0)$, we are trading-off asymptotic terms for improved non-asymptotic ones. We illustrate this with two archetypal instances. For the "1-sparse" instance, in which $\mu_1 > 0$ and $\mu_i = 0$ for all $i \neq 1$, we have by symmetry that $2w^\star_{1/2}(\nu)_i = 1/(K-1)$ for all $i \neq 1$. Therefore, we have $C_\varepsilon(w_0) = \mathcal{O}(K/\varepsilon)$ since $d_\nu(w_0) = 0$ for all $w_0 \in [0, 1/(K-1)]$. The "almost dense" instance is such that $\mu_1 = 1$, $\mu_K = 0$ and

Table 2.3 – Upper bound on the sample complexity τ_δ in probability (§) or in expectation (†). The notation $\mathcal{O}$ displays the dominating term when $\delta \to 0$ for the asymptotic regime, and when $H(\mu) \to +\infty$ (or $\Delta_i \to 0$) for the finite-confidence one. The notation $\tilde{\mathcal{O}}$ hides polylogarithmic factors. (**) Upper bound on $\mathbb{E}_\nu[\tau_\delta \mathbb{1}(\mathcal{E})]$ where $\mathbb{P}[\mathcal{E}^\complement] \leq \gamma$. (*) The asymptotic upper bound holds for instances having all distinct means, while the non-asymptotic one doesn't require this assumption. Ordered references: Kalyanakrishnan et al. [2012], Karnin et al. [2013], Jamieson et al. [2014], Degenne et al. [2019], Katz-Samuels et al. [2020], Wang et al. [2021], Barrier et al. [2022], Jourdan and Degenne [2023].

Algorithm	Asymptotic behavior	Finite-confidence behavior
LUCB1†	$\mathcal{O}\left(H(\mu)\log(1/\delta)\right)$	$\mathcal{O}\left(H(\mu)\log H(\mu)\right)$
Exp-Gap§	$\mathcal{O}\left(H(\mu)\log(1/\delta)\right)$	$\mathcal{O}(\sum_{i\neq i^\star}\Delta_i^{-2}\log\log\Delta_i^{-1})$
lil' UCB§	$\mathcal{O}\left(H(\mu)\log(1/\delta)\right)$	$\mathcal{O}(\sum_{i\neq i^\star}\Delta_i^{-2}\log\log\Delta_i^{-1})$
DKM†	$T^\star(\nu)\log(1/\delta)+\tilde{\mathcal{O}}(\sqrt{\log(1/\delta)})$	$\tilde{\mathcal{O}}\left(KT^\star(\nu)^2\right)$
Peace§	$\mathcal{O}\left(T^\star(\nu)\log(1/\delta)\right)$	$\mathcal{O}\left(H(\mu)\log(K/\Delta_{\min})\right)$
FWS†	$T^\star(\nu)\log(1/\delta)+\mathcal{O}(\log\log(1/\delta))$	$\mathcal{O}\left(e^K H(\mu)^{19/2}\right)$
EBS† **	$T^\star(\nu)\log(1/\delta)+o(1)$	$\mathcal{O}\left(KH(\mu)^4/w_{\min}^2\right)$
UCB-TC†*	$T^\star_\beta(\nu)\log(1/\delta)+\mathcal{O}(\log\log(1/\delta))$	$\mathcal{O}\left(\max\{H(\mu)\log H(\mu), K^{\frac{1}{\alpha-1}}\}^\alpha\right)$ for $\alpha>1$

$\mu_i = 1-\gamma$ for all $i \notin \{1, K\}$. By symmetry, there exists a function $h : [0,1) \to [0,(K-1)^{-1})$ with $\lim_{\gamma\to 0} h(\gamma) = 0$, such that $2w^\star_{1/2}(\nu)_K = h(\gamma)$ and $2w^\star_{1/2}(\nu)_i = (1-h(\gamma))/(K-2)$ for all $i \notin \{1, K\}$. While $\lim_{\gamma\to 0} C_\varepsilon(0) = +\infty$, we obtain $\lim_{\gamma\to 0} C_\varepsilon(w_0) = \mathcal{O}(K/\varepsilon)$ by taking $w_0 = (1-h(\gamma))/(K-2)$ since $d_\nu(w_0) = 1$.

Comparison with existing upper bounds Table 2.3 summarizes the asymptotic and non-asymptotic scaling of the upper bound on the sample complexity of existing BAI algorithms. Among the class of asymptotically (β-)optimal algorithms, very few of them also enjoy non-asymptotic guarantees, *e.g.* the analyses of Track-and-Stop and prior Top Two algorithms are asymptotic. The gamification approach of Degenne et al. [2019] is the first attempt to provide both. Their non-asymptotic upper bound on $\mathbb{E}_\nu[\tau_\delta]$ involves an implicit time $T_1(\delta)$ which scales with $KT^\star(\nu)^2$ and is only valid for $\log(1/\delta) \gtrsim KT^\star(\nu)$ (see Lemma 2, with constants in Appendix D.7). Let $T^\star_\delta := T^\star(\nu)\log(1/\delta)$. As a first order approximation, they obtain $T_1(\delta) \approx T^\star_\delta + \Theta\left(\sqrt{T^\star_\delta \log T^\star_\delta}\right)$, and we obtain $T_0(\delta) \approx \Theta\left(T^\star_\delta + \log T^\star_\delta\right)$. Wang et al. [2021] were the first to obtain an upper bound on $\mathbb{E}_\mu[\tau_\delta]$ of the form $\Theta(T^\star_\delta + \log\log(1/\delta))$. While they improved the second-order δ-dependent term, the δ-independent term scales with $e^K H(\mu)^{19/2}$ (see their Theorem 2 for $\varepsilon^{-1} \gtrsim T^\star(\nu)$, with constants given by Appendix N). The algorithm proposed by Barrier et al. [2022] has a non-asymptotic upper bound on $\mathbb{E}_\nu[\tau_\delta \mathbb{1}(\mathcal{E})]$ of the form $(1+\varepsilon)T^\star_\delta + f(\mu,\delta)$ which is valid for $\log(1/\delta) \gtrsim w_{\min}^{-2} K/\Delta_{\min}$, where $\mathcal{E}$ is such that $\mathbb{P}_\nu(\mathcal{E}^\complement) \leq \gamma$. Since $f(\mu,\delta) =_{\delta\to 0} o(1)$, they obtain a better δ-dependency. However, $f(\mu,\delta)$ is arbitrarily large when $w_{\min} := \min_{i\in[K]} w^\star(\nu)_i$ is arbitrarily small since it scales with $KH(\mu)^4/w_{\min}^2$. Therefore,

they suffer from the pitfall which we avoided by clipping. In light of Table 2.3, UCB-TC-β enjoys the best scaling when $H(\mu) \to +\infty$ in the class of asymptotically (β-)optimal BAI algorithms.

The LUCB1 algorithm [Kalyanakrishnan et al., 2012] is the first algorithm which resembles a Top Two algorithm, with the difference that LUCB samples both the leader and the challenger instead of choosing one. As LUCB1 satisfies $\mathbb{E}_\nu[\tau_\delta] \leq 292H(\mu)\log(H(\mu)/\delta) + 16$, it enjoys better scaling when $H(\mu) \to +\infty$ than UCB-TC-β. Since the empirical allocation of LUCB1 is not converging towards $w^\star_{1/2}(\nu)$, it is not asymptotically $1/2$-optimal. The Peace algorithm [Katz-Samuels et al., 2020] has a non-asymptotic upper bound on τ_δ of the form $\mathcal{O}((T^\star_\delta + \gamma^\star(\mu))\log(K/\Delta_{\min}))$ holding with probability $1-\delta$. The term $\gamma^\star(\mu)$ is a Gaussian-width which originates from concentration on the suprema of Gaussian processes, and $\gamma^\star(\mu) = \mathcal{O}(H(\mu))$.

Another class of BAI algorithms focus on the dependency in the gaps $\Delta_i := \mu_{i^\star} - \mu_i$, and derive non-asymptotic upper bound on τ_δ holding with high probability. Karnin et al. [2013], Jamieson et al. [2014], Chen et al. [2017b,c] gives δ-PAC algorithms with an upper bound of the form $\mathcal{O}(H(\mu)\log(1/\delta) + \sum_{i\neq i^\star}\Delta_i^{-2}\log\log\Delta_i^{-1})$, and Jamieson et al. [2014] shows that for two arms the dependency $\Delta^{-2}\log\log\Delta^{-1}$ is optimal when $\Delta \to 0$. While those algorithms obtain the best scaling when $H(\mu) \to +\infty$, they are not asymptotically (β-)optimal.

2.4.1 Proof Sketch of Theorem 2.25

Existing analyses of Top Two algorithms are asymptotic in nature and requires too much control on the empirical means and proportions to yield any meaningful information in the finite-confidence regime. Therefore, we adopt a different approach which resembles the non-asymptotic analysis of Degenne et al. [2019]. We first define concentration events to control the deviations of the random variables used in the UCB leader and the TC challenger. For all $n > K$, let $\mathcal{E}_n := \bigcap_{i\in[K]}\bigcap_{t\in[n^{5/6},n]}(\mathcal{E}^1_{t,i}\cap\mathcal{E}^2_{t,i})$ where

$$\mathcal{E}^1_{t,i} := \left\{\sqrt{N_{t,i}}|\mu_{t,i}-\mu_i| < \sqrt{6\log t}\right\} \text{ and } \mathcal{E}^2_{t,i} := \left\{\frac{(\mu_{t,i^\star}-\mu_{t,i})-(\mu_{i^\star}-\mu_i)}{\sqrt{1/N_{t,i^\star}+1/N_{t,i}}} > -\sqrt{8\log t}\right\} .$$

Using concentration results, it is straightforward to show that $\mathbb{P}_\nu(\mathcal{E}_n) \leq (2K-1)n^{-1.2}$ for all $n > K$. Using Lemma 2.26, the proof boils down to constructing a time $T(\delta)$ after which $\mathcal{E}_n \subseteq \{\tau_\delta \leq n\}$ for $n > T(\delta)$ since it would yield that $\mathbb{E}_\nu[\tau_\delta] \leq T(\delta) + 12K$.

Lemma 2.26. *Let $(\mathcal{E}_n)_{n>K}$ be a sequence of events and $T(\delta) > K$ be such that for $n \geq T(\delta)$, $\mathcal{E}_n \subseteq \{\tau_\delta \leq n\}$. Then, $\mathbb{E}_\nu[\tau_\delta] \leq T(\delta) + \sum_{n>K}\mathbb{P}_\mu(\mathcal{E}_n^\complement)$.*

Let $n > K$ such that $\mathcal{E}_n \cap \{n < \tau_\delta\}$ holds true, and $t \in [n^{5/6}, n]$ such that $B_t^{\mathrm{UCB}} = i^\star$. Using that $t \leq n < \tau_\delta$, under $\bigcap_{i \neq i^\star} \mathcal{E}_{t,i}^2$, the stopping condition yields that

$$\sqrt{2c(n-1,\delta)} \geq ((\mu_{i^\star} - \mu_{C_t^{\mathrm{TC}}})(1/N_{t,i^\star}^{i^\star} + 1/N_{t,C_t^{\mathrm{TC}}}^{i^\star})^{-1/2} - \sqrt{8 \log n})_+ .$$

Let $w_{1/2}^\star$ be the unique element of $w_{1/2}^\star(\nu)$. Lemma 2.27 links the empirical proportions $N_{t,i}^{i^\star}/(t-1)$ to $w_{1/2,i}^\star$ for $i \in \{i^\star, C_t^{\mathrm{TC}}\}$. It is the key technical challenge of our non-asymptotic proof strategy. Its proof is sketched below.

Lemma 2.27. *Let $\varepsilon \in (0,1]$. There exists $T_\mu > 0$ such that for all $n > T_\mu$ such that $\mathcal{E}_n \cap \{n < \tau_\delta\}$ holds true, there exists $t \in [n^{5/6}, n]$ with $B_t^{UCB} = i^\star$, which satisfies*

$$(n-1)(1/N_{t,i^\star}^{i^\star} + 1/N_{t,C_t^{\mathrm{TC}}}^{i^\star}) \leq (1+\varepsilon)^2 (2 + 1/w_{1/2,C_t^{\mathrm{TC}}}^\star)/\beta .$$

First, we conclude the proof of Theorem 2.25. Let ε, T_μ and t as in Lemma 2.27 and

$$T(\delta) := \sup \left\{ n \mid \; n - 1 \leq T_{1/2}^\star(\mu)(1+\varepsilon)^2 \left(\sqrt{c(n-1,\delta)} + \sqrt{4 \log n} \right)^2 / \beta \right\} .$$

For all $n > \max\{T_\mu, T(1)\}$, we have

$$\sqrt{c(n-1,\delta)} + \sqrt{4 \log n} \geq \sqrt{\beta (n-1) T_{1/2}^\star(\mu)^{-1} (1+\varepsilon)^{-2}} .$$

Therefore, we have proven the result since $\mathcal{E}_n \cap \{n < \tau_\delta\} = \emptyset$ for all $n > \max\{T_\mu, T(\delta)\}$.

Provided that $B_t = i^\star$, the above only used the stopping condition and the TC challenger, and no other properties of the leader. Lemma 2.28 shows that $B_t^{\mathrm{UCB}} = i^\star$, except for a sublinear number of times. Its proof in Appendix B.14 uses classical tools from the regret analysis of the UCB algorithms. Section 2.4.1 exhibits sufficient conditions on a regret minimization algorithm to obtain a result similar to Lemma 2.28, hence achieving a non-asymptotic upper bound.

Lemma 2.28. *Under the event $\bigcap_{k \in [K]} \bigcap_{t \in [n^{5/6}, n]} \mathcal{E}_{t,k}^1$, we have $L_{n,i^\star} \geq n - 1 - 24 H(\mu) \log n - 2K$.*

Proof sketch of Lemma 2.27 The key technical challenge is to link $N_{t,C_t^{\mathrm{TC}}}^{i^\star}/(n-1)$ with $w_{1/2,C_t^{\mathrm{TC}}}^\star$. We adopt the approach used in the analysis of the APT (Anytime Parameter-free Thresholding) algorithm [Locatelli et al., 2016]: consider an arm being over-sampled and study the last time this arm was pulled. Its empirical transportation cost will be simultaneously large (over-sampled) and the smallest (chosen as TC challenger), and can be related to the one at time n

(last time it was sampled). By the pigeonhole principle, at time n ,

$$\exists k_1 \neq i^\star, \text{ s.t. } \quad N^{i^\star}_{n,k_1} \geq 2(L_{n,i^\star} - N^{i^\star}_{n,i^\star}) w^\star_{1/2,k_1} \, . \tag{2.29}$$

Let t_1 be the last time at which $B^{\mathrm{UCB}}_t = i^\star$ and $C^{\mathrm{TC}}_t = k_1$, hence $N^{i^\star}_{t_1,k_1} \geq N^{i^\star}_{n,k_1} - 1$. Using Lemmas 2.8 and 2.28, we show that $N^{i^\star}_{t_1,k_1} \gtrapprox w^\star_{1/2,k_1}(n-1)$, hence $t_1 \geq n^{5/6}$ for n large enough. Then, we need to link $N^{i^\star}_{t_1,i^\star}$ with $(n-1)/2$. When $w^\star_{1/2,k_1}$ is small, (2.29) can be true at $t_1 = n^{5/6}$, hence there is no hope to show that $t_1 = n - o(n)$. To circumvent this problem, we link $N^{i^\star}_{t_1,i^\star}$ with $N^{i^\star}_{t_1,k_1}$ thanks to Lemma 2.8, and use that

$$\frac{n-1}{N^{i^\star}_{t_1,i^\star}} + \frac{n-1}{N^{i^\star}_{t_1,k_1}} \leq \left(2 + \frac{n-1}{N^{i^\star}_{t_1,k_1}}\right)\left(\frac{N^{i^\star}_{t_1,k_1}}{N^{i^\star}_{t_1,i^\star}} + 1\right) \leq 2(1+\varepsilon)^2(2 + 1/w^\star_{1/2,k_1}) \, ,$$

for $n > T_\mu(w_-)$ with $T_\mu(w_-) \leq \max\{C^{1.2}_\mu, (2/(\varepsilon w_-)+1)^6, (2/\varepsilon)^{1.2}\}$, where $w_- = \min_{i \neq i^\star} w^\star_{1/2,i}$ is a strictly positive lower bound on $w^\star_{1/2,k_1}$. This concludes the proof for $w_0 = 0$. The (sub-optimal) multiplicative factor 2 in $T_0(\delta, 0)$ comes from the inequality (2.30). To remove it, we need to control the deviation between the empirical proportion of arm i and $w^\star_{1/2,i}$ for all $i \in [K]$. Nevertheless, UCB-TC-1/2 is asymptotically $1/2$-optimal (see Section 2.3).

Refined analysis For $w_0 \in (0, (K-1)^{-1}]$, we clip $\min_{i \neq i^\star} w^\star_{1/2,i}$ by $w_0/2$ (see Jourdan and Degenne [2023] for more details). Our method could be used to analyze APT with threshold $\gamma \in \mathbb{R}$ and precision $\varepsilon = 0$ [Locatelli et al., 2016] in the fixed-confidence setting, while clipping $|\mu_i - \gamma|$ when this gap is too small. It can also be used for the non-asymptotic analysis of other algorithms, *e.g.* EB-TC$_\varepsilon$ in Chapter 5 and APGAI in Chapter 6.

Beyond Gaussian distributions Theorem 2.25 hold for 1-sub-Gaussian random variables thanks to direct adaptations of concentration results. The situation is akin to the regret bound of UCB: it holds for the class of 1-sub-Gaussian distributions, but it is close to optimality in a distribution-dependent sense only for Gaussian distributions. However, if the focus is on asymptotically β-optimal algorithms, then it is challenging to express the characteristic time $T^\star(\nu)$ for the non-parametric class of σ-sub-Gaussian distributions.

The UCB-TC-β algorithm can also be defined for more general distributions such as single-parameter exponential families or bounded distributions. It is only a matter of adapting the definition of the UCB leader and the TC challenger. For bounded distributions, we refer the reader to Chapter 4. We believe that non-asymptotic guaranties could be obtained for more general distributions, but it will come at the price of more technical arguments and less explicit non-asymptotic terms. In particular, it is challenging to prove a counterpart of Lemma 2.27.

Regret Minimization Leader

Our non-asymptotic analysis highlights that any regret minimization algorithm that selects the arm $i^\star$ except for a sublinear number of times (Property 2.29) can be used as leader with the TC challenger. The UCB leader satisfies Property 2.29 by Lemma 2.28.

Property 2.29. *There exists $(\tilde{\mathcal{E}}_n)_n$ with $\sum_n \mathbb{P}_\mu(\tilde{\mathcal{E}}_n^\complement) < +\infty$ and a function h with $h(n) = \mathcal{O}(n^\gamma)$ for some $\gamma \in (0,1)$ such that under event $\tilde{\mathcal{E}}_n$, $L_{n,i^\star} \geq n - 1 - h(n)$.*

Since the concentration events $(\tilde{\mathcal{E}}_n)_n$ are only controlling the choice of the leader, we still need to control the random variables used in the TC challenger. This is straightforwardly done by considering $\mathcal{E}_n = \tilde{\mathcal{E}}_n \cap (\bigcap_{i\in[K]} \bigcap_{t\in[n^{5/6},n]} \mathcal{E}_{t,i}^2)$ which satisfies that $\sum_n \mathbb{P}_\mu(\mathcal{E}_n^\complement) < +\infty$.

For asymptotic guarantees, the sufficient properties on the leader described in Section 2.3 are weaker since they are even satisfied by the greedy choice $B_n = \hat{\imath}_n$. While Top Two algorithms were introduced by Russo [2016] to adapt Thompson Sampling to BAI, we have shown that other regret minimization algorithms can be used for the choice of the leader. *The Top Two approach can be used as a wrapper to convert any regret minimization algorithm into a best arm identification strategy* by combining it with the TC challenger.

The regret of an algorithm at time n , $\bar{R}_n = \sum_{i\neq i^\star} \Delta_i N_{n,i}$, is almost always studied through its expectation $\mathbb{E}[\bar{R}_n]$. This is however not sufficient for our application. We need to prove that with high probability, $N_{n,i}$ is small for all arm $i \neq i^\star$. We showed those guarantees for UCB and they are known for ETC [Lattimore and Szepesvari, 2019], yet unknown for Thompson Sampling. We cannot in general obtain a good enough bound on $N_{n,i}$ from a bound on $\mathbb{E}[\bar{R}_n]$. However, we can if we have high probability bounds on $\bar{R}_n$. Suppose that a regret minimization algorithm Alg_1 satisfies Property 2.30 and is independent of the horizon n .

Property 2.30. *There exists $s > 1$, $\gamma \in (0,1)$, $(\mathcal{E}_{n,\delta})_{(n,\delta)}$ with $\sum_n \mathbb{P}_\mu[\mathcal{E}_{n,n^{-s}}^\complement] < +\infty$ and a function h with $h(n, n^{-s}) = \mathcal{O}(n^\gamma)$ such that under event $\mathcal{E}_{n,\delta}$, $\bar{R}_n \leq h(n,\delta)$.*

Let Alg_2 be the algorithm Alg_1 used in a Top Two procedure, but which uses only the observations obtained at times n such that $I_n = B_n$ and discards the rest. Let $\tilde{\mathcal{E}}_n = \mathcal{E}_{n,n^{-s}}$ and $\Delta_{\min} = \min_{i\neq i^\star} \Delta_i$. Then, under $\tilde{\mathcal{E}}_n$, Alg_2 satisfies $\sum_{i\neq i^\star} N_{n,i}^i \leq h(n,n^{-s})/\Delta_{\min}$ and Lemma 2.8 yields $N_{n,i^\star}^{i^\star} \geq \beta(n-1) - h(n,n^{-s})/\Delta_{\min} - K/2$. Therefore, Property 2.29 holds for $\tilde{\mathcal{E}}_n$ and $h(n) = (h(n,n^{-s})/\Delta_{\min} + K/2 + 1)/\beta$. Given a specific algorithm, a finer analysis could avoid discarding information by using Alg_1 with every observations.

2.5 Experiments

In the moderate regime ($\delta = 0.1$), we assess the empirical performance of several instances of the Top Two approach. For the leader/challenger pair, we consider the following benchmarks: TTTS [Russo, 2016] (*i.e.* TS-RS), TTEI [Qin et al., 2017] (*i.e.* EI-EI), T3C [Shang et al., 2020] (*i.e.* TS-TC), EB-TCI [Jourdan et al., 2022], TTUCB [Jourdan and Degenne, 2023] (*i.e.* UCB-TC) with bonus g_m and concentration parameters $s = \alpha = 1.2$. For the target allocation, we compare the fixed design $\beta = 1/2$ with the optimal designs IDS and BOLD. We recall that the mechanism to reach the target is defined implicitly by the first three choices (see Section 2.2.5). While TS-RS and TS-TC will use randomization, the three other algorithms will use tracking (2.18) (resp. (2.19)) for optimal (resp. fixed) design. In addition, we consider Track-and-Stop (TaS) [Garivier and Kaufmann, 2016], FWS [Wang et al., 2021], DKM [Degenne et al., 2019], LUCB [Kalyanakrishnan et al., 2012] and uniform sampling. At the exception of LUCB, all algorithms uses the stopping rule (2.2) with the heuristic threshold $c(n, \delta) = \log((1 + \log n)/\delta)$. Even though this choice is not sufficient to prove δ-correctness, it yields an empirical error which is several orders of magnitude lower than δ . To allow for a fair numerical comparison, LUCB uses $\sqrt{2c(n-1,\delta)/N_{n,i}}$ as bonus, which is too tight to yield valid confidence intervals. For supplementary experiments and implementation details, we refer the reader to Appendices I in Jourdan et al. [2022] and G in Jourdan and Degenne [2023].

"1-sparse" instances The ratio $T^\star_{1/2}(\nu)/T^\star(\nu)$ seems to reach its highest value $r_K = 2K/(1 + \sqrt{K-1})^2$ for "1-sparse" instances (Lemma C.6 in Jourdan and Degenne [2023]), *i.e.* $\mu_i = \mu_1 - \Delta$ for all $i \neq 1$ with $\Delta > 0$. To best observe differences between the Top Two algorithms using fixed design $\beta = 1/2$ and the ones using optimal designs IDS and BOLD, we consider such instances with $K = 35$ ($r_K \approx 3/2$) and $(\mu_1, \Delta) = (0, 0.5)$. We average our results on 1000 runs.

Figure 2.1(a) showcases that Top Two algorithms are significantly better than other existing BAI algorithms. The best performance among the Top Two algorithm is reached by TS-TC since EI-EI suffers from numerous outliers. UCB-TC is slightly better than TS-RS, which outperforms EB-TCI. In Figure 2.1(b), we observe that Top Two algorithms using the optimal design IDS instead of the fixed design $\beta = 1/2$ have a smaller empirical stopping time. In Figure 2.1(c), we see similar performance between the Top Two algorithms using the optimal design IDS or the optimal design BOLD.

Random instances We assess the performance on 1000 random Gaussian instances with $K = 10$ such that $\mu_1 = 0.6$ and $\mu_i \sim \mathcal{U}([0.2, 0.5])$ for all $i \neq 1$.

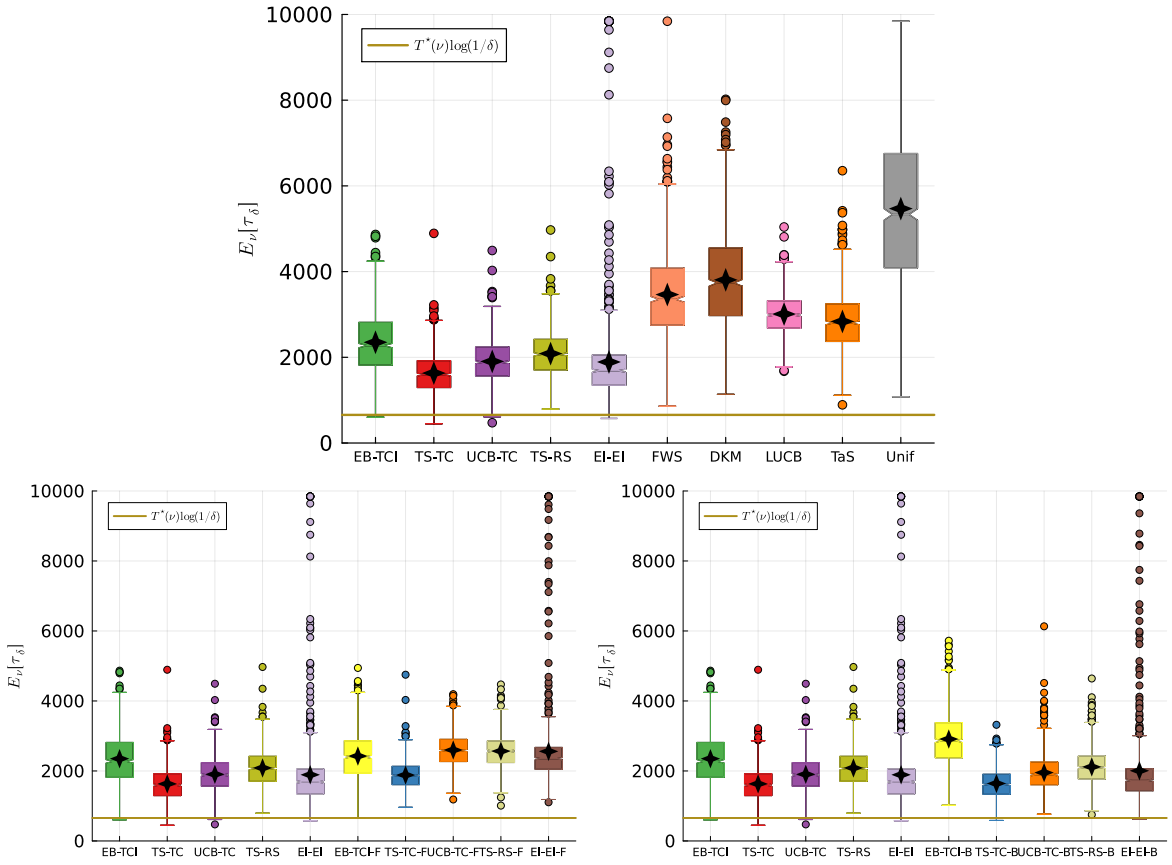

Figure 2.1 – Empirical stopping time on the "1-sparse" instance $(K, \mu^\star, \Delta) = (35, 0, 0.5)$. Top Two algorithms use optimal design IDS unless specified otherwise: "-F" is the fixed design $\beta = 1/2$ and "-B" is the optimal design BOLD. The lower bound is $T^\star(\nu)\log(1/\delta)$.

Figure 2.2(a) confirms the ranking between algorithms observed in Figure 2.1(a). Top Two algorithms have similar performance, and they outperform their competitors. On random instances, EB-TCI is better than UCB-TC and TS-RS.

Larger sets of arms We evaluate the impact of larger number of arms. As in Jamieson and Nowak [2014], we use the "1-sparse" instances $(\mu_1, \Delta) = (1/4, 1/4)$ with varying K for which $H(\mu) = 32(K-1)$. We restrict ourselves to algorithms with low computational cost and good empirical performance, average our results on 100 runs.

In Figure 2.2(b), all algorithms have the same linear scaling in K (*i.e.* in $H(\mu)$). Faced with an increase in the number of arms, the TS leader used in TS-TC appears to be more robust than the UCB leader in UCB-TC. This is a common feature of UCB algorithms which have to overcome the bonus of sub-optimal arms.

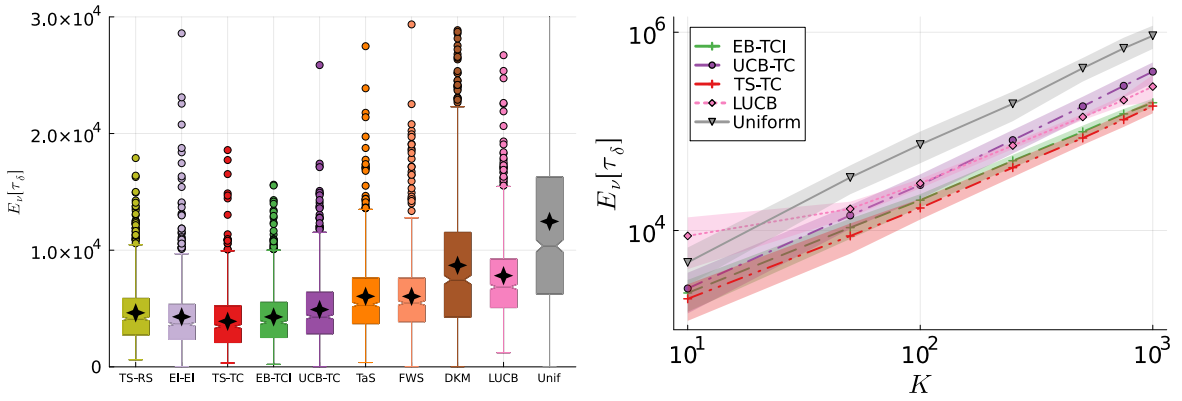

Figure 2.2 – Empirical stopping time on (a) random instances ($K = 10$) and (b) "1-Sparse" instances.

EB-TC algorithm As hinted in Section 2.1.2, the EB-TC algorithm perform poorly empirically. Its greediness is reflected by the large empirical stopping time, *i.e.* it gets stuck due to unlucky first draws. While this situation is particularly apparent for the "1-sparse" instances (*i.e.* the algorithm does not stop), we can also observe it for random instances. In the setting of Figure 2.2(a), EB-TC-IDS has an average empirical stopping time of $5.7 \cdot 10^3$, a standard deviation of 12535 , and a maximum at 95426 .

2.6 Discussion

In Chapter 2, we provided a detailed overview of the Top Two approach. We proposed a unified perspective on the class of Top Two algorithm which are defined by four choices (leader answer, challenger answer, target allocation and mechanism to reach it), and presented several instances (Section 2.2). We introduced a unified asymptotic analysis of the Top Two approach (Section 2.3), which identifies desirable properties on the choices of the leader and challenger answers to achieve asymptotic (β-)optimality (Theorem 2.9). Moreover, we gave the first non-asymptotic analysis of a Top Two algorithm (Theorem 2.25), which identifies sufficient properties of the leader (seen as a regret-minimization algorithm) for it to hold (Section 2.4). While different Top Two algorithms had similar empirical performance, they outperformed other algorithms (Section 2.5).

While asymptotic β-optimality was first shown for the Top Two approach in Shang et al. [2020], the proof was only done for TS-TC-β and TS-RS-β for Gaussian with known variance. While this chapter proposed a unified analysis for other Top Two algorithm, the presentation was only done for this specific class of distributions. Extension to more general class of distributions are also possible. We note that the extension to other classes of one-parameter exponential family is relatively straightforward (see Appendix H in Jourdan et al. [2022]). Therefore, we

will focus on more challenging classes, *e.g.* Gaussian with unknown variance (see Chapter 3) or non-parametric distributions (*e.g.* bounded distributions as in Chapter 4).

Even though IDS has been introduced by You et al. [2023] for general one-parameter exponential families, it is still an open problem to show asymptotic optimality for distributions other than Gaussian distributions. While an extension of IDS was proposed for more general classes of distributions, deriving guarantees for them is also challenging. Even though those extensions are only heuristics (at the moment), they tend to perform better than using a fixed β . Using BOLD as target and a forced exploration step, Bandyopadhyay et al. [2024] have shown asymptotic optimality of some Top Two algorithms for general one-parameter exponential families.

Chapter 3

Dealing with Unknown Variances

In Chapter 3, we study the vanilla BAI problem for Gaussian distributions with unknown variance in the fixed-confidence setting, as studied in Chapter 2 for known variance. The presented results were published in Jourdan et al. [2023a].

The problem of identifying the best arm among a collection of items having Gaussian rewards distribution is well understood when the variances are known. Despite its practical relevance for many applications, few works studied it for unknown variances. In this chapter we introduce and analyze two approaches to deal with unknown variances, either by plugging in the empirical variance or by adapting the transportation costs. In order to calibrate our two stopping rules, we derive new time-uniform concentration inequalities, which are of independent interest. Then, we illustrate the theoretical and empirical performances of our two sampling rule wrappers on Track-and-Stop and on a Top Two algorithm. Moreover, by quantifying the impact on the sample complexity of not knowing the variances, we reveal that it is rather small.

Contents

3.1 Introduction

As detailed in Chapter 1, the motivation to study BAI for Gaussian distributions with unknown variance comes from practical consideration. Quite surprisingly, and despite its practical relevance, this problem has received little attention in the bandit literature. Gaussian distributions could indeed be used to model the revenue generated by different versions of a website in the context of A/B testing, or some biological indicator of the efficiency of a treatment in the context of an adaptive clinical trial comparing several treatments. In both case, assuming known variances is a limitation.

We consider the set $\mathcal{D}_{\mathcal{N}}$ of Gaussian distributions with unknown variance, hence $\mathcal{D}^K = \mathcal{D}_{\mathcal{N}}^K$. The distributions are denoted by $\nu_{x,\sigma^2} = \mathcal{N}(x, \sigma^2)$. Recall that the set of Gaussian distributions with known variance is denoted by $\mathcal{D}_{\mathcal{N}_\sigma}$. We denote the Kullback-Leibler (KL) divergence between ν_{x_1,σ_1^2} and ν_{x_2,σ_2^2} by $\mathrm{KL}((x_1,\sigma_1^2),(x_2,\sigma_2^2))$. Let $\nu \in \mathcal{D}^K$ which is uniquely defined by its mean and variance vectors $(\mu, \sigma^2) \in \mathbb{R}^K \times (\mathbb{R}_+^\star)^K$ such that the set of arms with largest mean $i^\star(\mu) := \arg\max_{i\in[K]} \mu_i$ is reduced to a singleton denoted by $i^\star$ (or $i^\star(\mu)$ by abusing notation), *i.e.* $\mathcal{S} = \{\mu \in \mathbb{R}^K \mid |i^\star(\mu)| = 1\}$.

A fixed-confidence algorithm is defined by a sampling rule, a recommendation rule and a stopping rule. At time n, we denote by $\hat{\imath}_n$ the candidate answer and by I_n the arm to pull. The stopping rule (and stopping time τ_δ) using a fixed confidence level $1-\delta \in (0,1)$ which ensures δ-correctness, *i.e.* $\mathbb{P}_\nu(\tau_\delta < +\infty,\ \hat{\imath}_{\tau_\delta} \neq i^\star(\mu)) \leq \delta$ for all instances $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathcal{S}$.

In this chapter, we follow the approach pioneered by Garivier and Kaufmann [2016] and initially introduced for one-dimensional parametric models (*e.g.* Gaussian with known variance). They derived lower bounds on the expected sample complexity of δ-correct algorithms and introduced algorithms inspired by the maximization of those lower bounds. As it is common in previous work for the stopping rule, we will compare a Generalized Likelihood Ratio (GLR) to a well chosen threshold [Kaufmann and Koolen, 2021].

Related work Algorithms based on GLR stopping rules and aimed at matching a sample complexity lower bound were either studied for one-parameter exponential families [Degenne et al., 2019] or under generic heavy tails assumption [Agrawal et al., 2020]. Other algorithms are either based on eliminations or on confidence intervals and have been mostly analyzed for sub-Gaussian distributions with a known variance proxy [Even-Dar et al., 2006, Kalyanakrishnan et al., 2012, Jamieson et al., 2014]. For the special case of bounded distributions, confidence intervals based on the empirical variance have been used [Gabillon et al., 2012, Lu et al., 2021] but the resulting algorithms cannot be applied to unbounded distributions as they rely on the empirical Bernstein inequality [Maurer and Pontil, 2009]. In the fixed budget setting,

in which the size of the exploration phase is fixed in advance, it is possible to upper bound the error probability of the Successive Reject algorithm of Audibert et al. [2010] when the variances are unknown, as we only need to upper bound the probability that one empirical mean is smaller than another, see also Faella et al. [2020]. However, in the fixed-confidence setting elimination thresholds, confidence intervals or GLR tests need to be calibrated in a data-dependent way, which calls for the development of new time-uniform concentration inequalities, that we provide in this work.

In the related literature on ranking and selection [Hong et al., 2021], the problem of finding the Gaussian distribution with largest mean has been studied for unknown variances. This literature mostly seek to design algorithm that are δ-correct whenever the gap between the best and second best arm is larger than some specified indifference zone [Kim and Nelson, 2001]. However the work of Fan et al. [2016] does not consider an indifference zone and their algorithm is therefore comparable to ours. They propose an elimination strategy which features the empirical variances and whose calibration is done based on simulation arguments (resorting to continuous-time approximations) and justified in an asymptotic regime only (when δ goes to zero). Our algorithms have better empirical performance and stronger theoretical guarantees.

Contribution 3.1. *In Chapter 3, we propose two approaches to deal with unknown variances: plugging in the empirical variance or considering the transportation costs for unknown variance. This allows to easily adapt existing algorithms.*

- *By extending the lower bound of Garivier and Kaufmann [2016] to our two-parameters setting, we quantify the impact on the expected sample complexity of not knowing the variances.*
- *Our two approaches yield the Empirical Variance GLR (EV-GLR) stopping rule, which plugs in the empirical variance in a GLR assuming known variance, and the GLR stopping rule, which corresponds to a GLR assuming unknown variance. Our main technical contribution lies in the derivation of (near) optimal stopping thresholds which ensure the δ-correctness of both the GLR and the EV-GLR stopping rules, regardless of the sampling rule.*
- *When considering the sampling rule, each approach yields a wrapper which is a simple procedure that can be applied to any BAI algorithm for known variances. We illustrate each wrapper with the Track-and-Stop and the Top Two approach. We show that algorithms obtained by adapting the transportation costs enjoy stronger theoretical guarantees than the ones plugging in the empirical variance, and obtain the first asymptotically optimal algorithms for Gaussian bandits with unknown variances.*

Empirically, both wrappers have comparable performance when applied to multiple BAI algorithms. Our findings show that not knowing the variances has a small impact on the expected sample complexity.

3.2 Lower Bound and GLR Stopping Rule

3.2.1 Lower Bounds

As discussed in Section 1.4.1, the δ-correctness requirement leads to a lower bound on the expected sample complexity on any instance.

Lemma 3.2 (Garivier and Kaufmann [2016]). *An algorithm which is δ-correct on all problems in $\mathcal{D}^K_{\mathcal{N}_\sigma}$ satisfies that, for all $\nu \in \mathcal{D}^K_{\mathcal{N}_\sigma}$ with mean $\mu \in \mathcal{S}$, $\mathbb{E}_\nu[\tau_\delta] \geq T^\star(\nu;\sigma)\log(1/(2.4\delta))$. An algorithm which is δ-correct on all problems in $\mathcal{D}^K_{\mathcal{N}}$ satisfies that, for all $\nu \in \mathcal{D}^K_{\mathcal{N}}$ with mean $\mu \in \mathcal{S}$, $\mathbb{E}_\nu[\tau_\delta] \geq T^\star(\nu)\log(1/(2.4\delta))$.*

For Gaussian with unknown (resp. known) variances, Lemma 3.2 shows that $T^\star(\nu)$ (resp. $T^\star(\nu;\sigma)$) is the asymptotic complexity of the BAI problem on the instance $\nu := (\nu_{\mu_a,\sigma_i^2})_a$, where

$$
\begin{aligned}
&T^\star(\nu;\sigma)^{-1} = \sup_{w\in\Sigma_K} \min_{j\neq i} C(i,j;\nu,w;\sigma) \quad \text{with} \\
&C(i,j;\nu,w;\sigma) = \frac{1}{2}\mathbb{1}\left(\mu_i \geq \mu_j\right) \inf_{u\in\mathbb{R}} \left\{ w_i \frac{(\mu_i-u)^2}{\sigma_i^2} + w_j \frac{(\mu_j-u)^2}{\sigma_j^2} \right\} , \\
&T^\star(\nu)^{-1} = \sup_{w\in\Sigma_K} \min_{j\neq i} C(i,j;\nu,w) \quad \text{with} \\
&C(i,j;\nu,w) = \mathbb{1}\left(\mu_i > \mu_j\right) \inf_{u\in\mathbb{R}} \left\{ w_i \mathcal{K}^-_{\inf}(\nu_i,u) + w_j \mathcal{K}^+_{\inf}(\nu_j,u) \right\} ,
\end{aligned}
$$

where the function $\mathcal{K}_{\inf}$ over the class $\mathcal{D}_{\mathcal{N}}$ has the following closed-form expression

$$
\begin{aligned}
\mathcal{K}^-_{\inf}(\nu_{x,\sigma^2},u) &= \frac{1}{2}\mathbb{1}\left(x>u\right)\log\left(1+\frac{(x-u)^2}{\sigma^2}\right) \quad \text{and} \\
\mathcal{K}^+_{\inf}(\nu_{x,\sigma^2},u) &= \frac{1}{2}\mathbb{1}\left(x<u\right)\log\left(1+\frac{(x-u)^2}{\sigma^2}\right) .
\end{aligned}
$$

The maximizer over the simplex Σ_K in these complexities is denoted by $w^\star(\nu)$ and $w^\star(\nu;\sigma)$. The rationale for the difference between the $T^\star(\nu)$ and $T^\star(\nu;\sigma)$ is that when the variances are unknown, there exists instances $\nu_{\lambda,\tilde{\sigma}^2}$ for $\tilde{\sigma} \neq \sigma$ that are harder to differentiate from (μ,σ^2) than instances ν_{λ,σ^2} with respect to an information criterion. Namely, the minimizer in $\tilde{\sigma}$ is $\tilde{\sigma}_i = \sigma_i^2 + (\mu_i-\lambda_i)^2$, thus even if we want to identify the arm with largest mean, the closest alternatives have an increased variance. When the variances are known, the transportation cost has a convenient closed form, *i.e.* $C(i,j;\nu,w;\sigma) = \mathbb{1}\left(\mu_i>\mu_j\right)\frac{1}{2}\frac{(\mu_i-\mu_j)^2}{\sigma_i^2/w_i+\sigma_j^2/w_j}$. On the contrary, the infimum in the alternative mean parameter λ in the transportation cost for unknown variance has no simple analytic form since it is a real root of a cubic polynomial. Still, comparing the two

types of transportation costs (and using properties of the mapping $x \mapsto \log(1+x)/x$) permits to establish a link between $T^\star(\mu;\sigma)$ and $T^\star(\nu)$ (resp. $T^\star_\beta(\mu;\sigma)$ and $T^\star_\beta(\nu)$), hence to quantify the impact of not knowing the variances.

Lemma 3.3. *Let* $d(\nu) = \max_{i \neq i^\star(\mu)} \frac{(\mu_{i^\star(\mu)} - \mu_i)^2}{\min\{\sigma_i^2, \sigma_{i^\star(\mu)}^2\}}$. *Then,*

$$1 < \frac{T^\star(\nu)}{T^\star(\mu;\sigma)} \leq \frac{d(\nu)}{\log(1+d(\nu))} \quad \textit{and} \quad 1 < \frac{T^\star_\beta(\nu)}{T^\star_\beta(\nu;\sigma)} \leq \frac{d(\nu)}{\log(1+d(\nu)} \,. \tag{3.1}$$

Proof. Using that $x \to \log(1+x)/x$ is decreasing for $x > 0$ and $x \to \log(1+x)$ is concave, we can obtain the result directly. See Lemma 4 in Jourdan et al. [2023a]. ■

When $d(\nu)$ is small, say $d(\nu) \leq 1$, the two complexities are close since we then have $T^\star(\nu;\sigma)/T^\star(\nu) \in [\log 2, 1)$. Observe that a small $d(\nu)$ also implies that the BAI problem is hard: if $d(\nu) \leq c \in \mathbb{R}_+$ then for all $i \in [K]$, $\frac{\min\{\sigma_i^2, \sigma_{i^\star(\mu)}^2\}}{(\mu_{i^\star(\mu)} - \mu_i)^2} \geq c^{-1}$. Since that ratio is roughly the number of samples needed to distinguish the two arms, the problem is hard when it is large. Still, there exists instances with an arbitrarily large complexity ratio $T^\star(\nu)/T^\star(\nu;\sigma)$. We conjecture that they always correspond to easy problems, for which both $T^\star(\nu;\sigma)$ and $T^\star(\nu)$ are small. Lemma 3.3 is not sufficient to prove this conjecture as there exists hard instances with a large value of $d(\nu)$ and instances for which the upper bound in (3.1) is not tight.

Asymptotic (β-)optimality We say that an algorithm is asymptotically optimal (resp. β-optimal) on $\mathcal{D}^K_{\mathcal{N}}$ if it is δ-correct and its sample complexity matches that lower bound, *i.e.* for all $\nu \in \mathcal{D}^K_{\mathcal{N}}$ such that $\mu \in \mathcal{S}$, $\limsup_{\delta \to 0} \mathbb{E}_\nu[\tau_\delta]/\log(1/\delta) \leq T^\star(\nu)$ (resp. $T^\star_\beta(\nu)$). For $\beta \in (0,1)$, the definition of $T^\star_\beta(\nu)$ is the same as $T^\star(\nu)$ with the additional constraint on the outer maximization that $w_{i^\star} = \beta$, hence $T^\star(\nu) = \min_{\beta \in (0,1)} T^\star_\beta(\nu)$. Recall that $T^\star_{1/2}(\nu) \leq 2T^\star(\nu)$ [Russo, 2016]. The β-optimality on $\mathcal{D}^K_{\mathcal{N}_\sigma}$ involves $T^\star_\beta(\mu;\sigma)$, which is similarly related to $T^\star(\mu;\sigma)$. While there is a rich literature on asymptotically (β-)optimal algorithms for Gaussian with known variance, we are the first to derive algorithms with those guarantees when the variances are unknown.

3.2.2 GLR Stopping Rules

For all $i \in [K]$, let $N_{n,i} = \sum_{t \in [n-1]} \mathbb{1}(I_t = i)$, $\nu_{n,i} = \mathcal{N}(\mu_{n,i}, \sigma^2_{n,i})$, $\mu_{n,i}$ and $\sigma^2_{n,i}$ be the empirical count, distribution, mean and variance of arm i after time n , defined as

$$\mu_{n,i} := N_{n,i}^{-1} \sum_{t \in [n-1]} \mathbb{1}(I_t = i) X_{t,i} \qquad \text{and} \qquad \sigma^2_{n,i} := N_{n,i}^{-1} \sum_{t \in [n-1]} \mathbb{1}(I_t = i)(X_{t,i} - \mu_{n,i})^2 \,.$$

As candidate answer, we use $\hat{\imath}_n \in i^\star(\mu_n)$, *i.e.* the empirical best arm (EB). For the stopping rule, we use GLR stopping rules (see Section 1.4.2 for more details). For Gaussian with unknown variances, the GLR can be written as $\min_{i \neq \hat{\imath}_n} W_n(\hat{\imath}_n, i)$, where the empirical transportation cost between arm i and arm j is defined as

$$W_n(i,j) = C(i,j;\nu_n,N_n) = \mathbb{1}\left(\mu_{n,i} > \mu_{n,j}\right) \inf_{u \in \mathbb{R}} \sum_{k \in \{i,j\}} \frac{N_{n,k}}{2} \log\left(1 + \frac{(\mu_{n,k}-u)^2}{\sigma^2_{n,k}}\right) .$$

For Gaussian with known variances, replacing the variance vector σ^2 by its empirical estimate σ^2_n yields the empirical transportation cost for known variance, which is defined as

$$W_n^{\mathrm{EV}}(i,j) = C(i,j;\nu_n,N_n;\sigma_n) = \mathbb{1}\left(\mu_{n,i} > \mu_{n,j}\right) \frac{(\mu_{n,i}-\mu_{n,j})^2}{2(\sigma^2_{n,i}/N_{n,i} + \sigma^2_{n,j}/N_{n,j})} .$$

Let $(c_{i,j})_{(i,j)\in[K]^2}$ such that $c_{i,j} : \mathbb{N}^K \times (0,1] \to \mathbb{R}_+$. The GLR stopping rule is defined as

$$\tau_\delta := \inf\left\{n \in \mathbb{N} \mid \forall i \neq \hat{\imath}_n,\ W_n(\hat{\imath}_n, i) > c_{\hat{\imath}_n,i}(N_n,\delta)\right\} . \tag{3.2}$$

The EV-GLR stopping rule given a family of thresholds $(c_{i,j})_{(i,j)\in[K]^2}$ is defined as

$$\tau_\delta^{\mathrm{EV}} := \inf\left\{n \in \mathbb{N} \mid \forall i \neq \hat{\imath}_n,\ W_n^{\mathrm{EV}}(\hat{\imath}_n, i) > c_{\hat{\imath}_n,i}(N_n,\delta)\right\} . \tag{3.3}$$

The (resp. EV-)GLR stopping rule is a good candidate to match $T^\star(\nu)$ (resp. $T^\star(\nu;\sigma)$). Indeed, it is easy to prove that sampling arms from $w^\star(\nu)$ (resp. $w^\star(\nu;\sigma)$) and using the threshold $c_{i,j}(N_n,\delta) = \log(1/\delta)$, the lower bound would be matched. However, such a threshold is too good to be δ-correct (Section 3.3). Moreover, $w^\star(\nu)$ (resp. $w^\star(\nu;\sigma)$) needs to be estimated since it is unknown (Section 3.4).

3.3 Calibration of the Stopping Thresholds

We present ways of calibrating the thresholds used by the GLR stopping rule, by leveraging concentration arguments. Under any sampling rule, to obtain a δ-correct GLR stopping rule it suffices to show that the family of thresholds is such that the following time-uniform concentration inequality holds for all $\nu \in \mathcal{D}_{\mathcal{N}}^K$ with mean $\mu \in \mathcal{S}$: with probability $1-\delta$,

$$\forall n \in \mathbb{N}, \forall i \neq i^\star(\mu), \quad \sum_{k \in \{i, i^\star(\mu)\}} \frac{N_{n,k}}{2} \log\left(1 + \frac{(\mu_{n,k}-\mu_k)^2}{\sigma^2_{n,k}}\right) \leq c_{i,i^\star(\mu)}(N_n,\delta) . \tag{3.4}$$

Aiming at matching the lower bound, we want to derive a family of thresholds satisfying $c_{i,j}(w,\delta) \sim_{\delta\to 0} \log(1/\delta)$. As regards the time dependency, generalizations of the law of the

iterated logarithm suggest we could achieve $\mathcal{O}(\log\log n)$. Both dependencies are achieved for known variances [Kaufmann and Koolen, 2021], and we are the first to show it for unknown variances (Theorem 3.6). While simple ideas yield δ-correct thresholds (Section 3.3.1), obtaining the ideal dependency in δ requires sophisticated concentration arguments (Section 3.3.2).

Similar arguments can be used to calibrate the thresholds used by the EV-GLR stopping rule. Moreover, δ-correct thresholds for the EV-GLR stopping rule can be obtained by using the ones calibrated for GLR stopping rule, and vice-versa (see Jourdan et al. [2023a]).

3.3.1 Simple Ideas

As per-arm concentration results are easier to obtain, we control each term of the sum in (3.4).

Student thresholds Since $(\mu_{n,i} - \mu_i)/\sigma_{n,i}$ is an observation of the Student distribution $\mathcal{T}_{N_{n,i}-1}$, a first simple approach involves the quantiles of Student distributions with n degrees of freedom. A direct union bound over time and arms yield a δ-correct family of thresholds.

Lemma 3.4. *Let $s > 1$ and ζ be the Riemann ζ function. Let a family of thresholds $c_{i,j}(N_n, \delta)$ with value $+\infty$ if $n < \max_{k\in\{i,j\}} t_k^S(\delta)$ and otherwise $c_{i,j}^S(N_n, \delta) = \max\left\{c^S(N_{n,i}, \delta), c^S(N_{n,j}, \delta)\right\}$. Taking*

$$c^S(n,\delta) = n\log\left(1 + \frac{1}{n-1}Q\left(1 - \frac{\delta}{4(K-1)\zeta(s)n^s}; \mathcal{T}_{n-1}\right)^2\right) \tag{3.5}$$

yields a δ-correct family of thresholds for the GLR stopping rule on instances $\nu \in \mathcal{D}_{\mathcal{N}}^K$ with mean $\mu \in \mathcal{S}$. The stochastic initial times are

$$\forall i \in [K], \quad t_i^S(\delta) := \inf\left\{n \in \mathbb{N} \mid N_{n,i} \geq \max\left\{2, \left(\frac{\delta}{4(K-1)\zeta(s)}\right)^{1/s}\right\}\right\} . \tag{3.6}$$

Box thresholds As illustrated in Figure 3.1, the Student threshold suffers from a probably sub-optimal dependence in both $\log(1/\delta)$ and n . This is why we propose an alternative method where the union bound is replaced by time-uniform concentration (which has proved useful to improve both dependencies in different contexts) and the Student concentration by concentration on the mean and the variance separately. The resulting time-uniform upper and lower tail concentration inequalities for the empirical variance (Corollary C.8) are of independent interest. Thanks to these "box" confidence regions on (μ_n, σ_n^2) , Lemma 3.5 yields a δ-correct family of thresholds. The proof of Lemma 3.5 is detailed in Appendix C.1.

Lemma 3.5. *Let $\eta_0 > 0$, $s > 1$, ζ be the Riemann ζ function and, for $i \in \{0, -1\}$, $\overline{W}_i(x) = -W_i(-e^{-x})$ for $x \geq 1$ where $(W_i)_{i\in\{0,-1\}}$ are the branches of the Lambert W function. Define*

$$\varepsilon_\mu(n,\delta) = \frac{1}{n}\overline{W}_{-1}\left(1 + 2\log\left(\frac{4(K-1)\zeta(s)}{\delta}\right) + 2s + 2s\log\left(1 + \frac{\log n}{2s}\right)\right),$$
$$1 - \varepsilon_{-,\sigma}(n,\delta) = \overline{W}_0\left(1 + \frac{2(1+\eta_0)}{n}\left(\log\left(\frac{4(K-1)\zeta(s)}{\delta}\right) + s\log\left(1 + \log_{1+\eta_0}(n)\right)\right)\right) - \frac{1}{n}.$$

The family of thresholds $c^{Box}_{i,j}(N_n,\delta)$ with value $+\infty$ if $n < \max_{k\in\{i,j\}} t^{Box}_k(\delta)$ and otherwise

$$c^{Box}_{i,j}(N_n,\delta) = \sum_{k\in\{i,j\}} \frac{N_{n,k}}{2}\log\left(1 + \frac{\varepsilon_\mu(N_{n,k},\delta)}{1 - \varepsilon_{-,\sigma}(N_{n,k}-1,\delta)}\right) \tag{3.7}$$

yields a δ-correct family of thresholds for the GLR stopping rule on instances $\nu \in \mathcal{D}^K_{\mathcal{N}}$ with mean $\mu \in \mathcal{S}$. The stochastic initial times are

$$\forall i,\ t^{Box}_i(\delta) = \inf\left\{ n \mid N_{n,i} > 1 + e^{1+W_0\left(\frac{2(1+\eta_0)}{e}\left(\log\left(\frac{4(K-1)\zeta(s)}{\delta}\right) + s\log\left(1 + \frac{\log(N_{n,i}-1)}{\log(1+\eta_0)}\right)\right) - e^{-1}\right)} \right\}. \tag{3.8}$$

To derive the Box threshold, we leverage a lower bound on the empirical variance which is ensured to be strictly positive (hence informative) thanks to the initial time condition (3.8). As $W_0(x) \in [-1, +\infty)$, it also yields that $N_{n,i} > 2$. Using that $W_0(x) \approx \log(x) - \log\log(x)$, it is asymptotically equivalent to $\frac{2(1+\eta_0)\log(1/\delta)}{\log\log(1/\delta)}$. Since the lower bound in Lemma 3.2 suggests that the stopping time is asymptotically equivalent to $T^\star(\nu)\log(1/\delta)$, the condition (3.8) has a vanishing influence compared to the stopping time. For the parameters used in our simulations (see Section 3.3.3), (3.8) is empirically satisfied after sampling each arm 16 times for $\delta = 0.1$ and 20 times for $\delta = 0.001$. Recall that $\overline{W}_{-1}(x) \approx x + \log x$ and $\overline{W}_0(x) \approx e^{-x+e^{-x}}$ (see Appendix A).

3.3.2 Beyond Box

While being simpler to derive by controlling each arm independently, the above thresholds have a worse δ dependency than more sophisticated approach controlling the joint term (3.4). Since it is challenging to deal with (3.4), we consider as a proxy the KL divergences for which is is easier to construct martingales, which can improve on the δ dependency. To do so, we "remove" the minimization step over variances, *i.e.* consider KL instead of $\mathcal{K}_{\inf}$. Then, we apply the arguments used to obtain (3.4). Under any sampling rule, to obtain a δ-correct GLR stopping rule it suffices to show that the family of thresholds is such that the following time-uniform

concentration inequality holds for all $\nu \in \mathcal{D}_{\mathcal{N}}^K$ with mean $\mu \in \mathcal{S}$: with probability $1-\delta$,

$$\forall n \in \mathbb{N}, \forall i \neq i^\star(\mu), \quad \sum_{j \in \{i, i^\star(\mu)\}} N_{n,j} \mathrm{KL}((\mu_{n,j}, \sigma^2_{n,j}), (\mu_j, \sigma^2_j)) \leq c_{i,i^\star(\mu)}(N_n, \delta) \,. \tag{3.9}$$

KL thresholds First, we derive time-uniform concentration results on the summation of KL divergences. Then, applied to our setting, it yields a δ-correct family of thresholds (Theorem 3.6).

Theorem 3.6. *Let $\eta_1 > 0$, $\gamma, s > 1$. Let ε_μ , $\varepsilon_{-,\sigma}$ as in Lemma 3.5 with $\tilde{\delta} = \frac{\delta}{3}$ and $(t_i^{Box})_i$ as in (3.8), and define $1 + \varepsilon_{+,\sigma}(n,\delta) =$*

$$\overline{W}_{-1}\left(1 + \frac{2(1+\eta_1)}{n}\left(\log\left(\frac{12(K-1)\zeta(s)}{\delta}\right) + s\log\left(1 + \log_{1+\eta_1}(n)\right)\right)\right) - \frac{1}{n}.$$

For all n and all $i \in [K]$, define $n_{n,i} = \gamma^{\lfloor \log_\gamma N_{n,i} \rfloor}$, $\bar{t}_{n,i} = \inf\{n \mid N_{n,i} = n_{n,i}\}$,

$$\mu^2_{++,n,i} = \max_{\pm}\left(\mu_{\bar{t}_{n,i},i} \pm 2\sigma_{\bar{t}_{n,i},i}\sqrt{\frac{\varepsilon_\mu(n_{n,i},\delta)}{1-\varepsilon_{-,\sigma}(n_{n,i}-1,\delta)}}\right)^2,$$

$$\sigma^2_{\pm,n,i} = \sigma^2_{\bar{t}_{n,i},i}\frac{1 \pm \varepsilon_{\pm,\sigma}(n_{n,i}-1,\delta)}{1 \mp \varepsilon_{\mp,\sigma}(n_{n,i}-1,\delta)} \quad \text{and} \quad R_{n,i}(\delta) = \frac{\sigma^3_{+,n,i} f_+\left(g(\sigma^2_{+,n,i}, \mu^2_{++,n,i})\right)}{\sigma^3_{-,n,i} f_-\left(g(\sigma^2_{-,n,i}, \mu^2_{++,n,i})\right)},$$

where $f_\pm(x) = \frac{1 \pm \sqrt{1-x}}{\sqrt{x}}$ and $g(x,y) = \frac{2x}{(x+2y+\frac{1}{2})^2}$. The family of thresholds $c^{\mathrm{KL}}_{i,j}(N_n,\delta)$ with value $+\infty$ if $n < \max_{k\in\{i,j\}} \max\{t^{Box}_k(\delta/3), t^m_k(\delta)\}$ and otherwise

$$c^{\mathrm{KL}}_{i,j}(N_n,\delta) = 4\overline{W}_{-1}\left(1 + \frac{\log\frac{2\zeta(s)^2}{\delta}}{4} + \sum_{k\in\{i,j\}}\left(\frac{s}{4}\log(1+\log_\gamma N_{n,k}) + \frac{1}{2}\log\left(\gamma R_{n,k}(\delta)\right)\right)\right) \tag{3.10}$$

yields a δ-correct family of thresholds for the GLR stopping rule on instances $\nu \in \mathcal{D}_{\mathcal{N}}^K$ with mean $\mu \in \mathcal{S}$. The stochastic initial times are

$$\forall i,\ t^m_i(\delta) = \inf\left\{n \mid N_{n,i} > 1 + \max\left\{\frac{e^{s/\log\left(\frac{12(K-1)\zeta(s)}{\delta}\right)}}{1+\eta_0}, \frac{e^{s/\left(\log\left(\frac{12(K-1)\zeta(s)}{\delta}\right) - \frac{1}{2(1+\eta_1)}\right)}}{1+\eta_1}\right\}\right\}. \tag{3.11}$$

As $\overline{W}_{-1}(x) \approx x + \log(x)$, Theorem 3.6 proves that we can obtain δ-correct threshold with the dependencies $c(N_n,\delta) \sim_{\delta\to 0} \log(1/\delta)$ and $c(N_n,\delta) \sim_{n\to+\infty} C\log\log n$, which are widely used in practice for BAI problems. While this dependency is justified for known variances [Kaufmann and Koolen, 2021], Theorem 3.6 legitimates its use for unknown variances.

To control the KL divergence between the true parameter and the MLE for Gaussian with unknown variances, our threshold combines two concentration results and is obtained by covering $\mathbb{N}$ with slices of times with geometrically increasing size (referred to as the "peeling" method). First, we use a crude per-arm concentration step to restrict the estimated parameters to a region around the true mean and variance. Then, a second result uses the knowledge of the restriction to get a finer concentration on the weighted sum of KL. It is proved for generic exponential families by approximating the KL divergence by a quadratic function on this crude confidence region. In (3.10), $R_{n,i}(\delta)$ represents the cost of this approximation, while $\log(1 + \log_\gamma N_{n,k})$ is the cost of time-uniform. The initial time condition (3.11) ensures the monotonicity of the preliminary concentration, and it is of the form $N_{n,i} > 1 + c_0(\delta)$ where $c_0(\delta) > 0$. In our simulations (see Section 3.3.3), (3.11) is empirically satisfied after sampling each arm twice for all considered δ . The detailed proof is omitted for the sake of space, and we refer the reader to Appendices F and G in Jourdan et al. [2023a] for more details.

Degenne [2019] derives concentration on the KL divergence of sub-Gaussian d-dimensional exponential families defined on the natural parameter space $\Theta_D = \mathbb{R}^d$. This doesn't include Gaussian with unknown variance, but our proof builds on his method. The main challenge was to tackle $\Theta_D \neq \mathbb{R}^d$, and we solved it by truncation on the sequence of crude confidence regions. In generalized linear bandits, truncated Gaussian distributions were also used to derive tail-inequalities for martingales "re-normalized" by their quadratic variation [Faury, 2021]. For general d-dimensional exponential families, Chowdhury et al. [2023] derives concentrations on the KL divergence between the true parameter and a linear combination of the MLE and the true parameter. As we are interested in the KL divergence between the true parameter and the MLE, we cannot leverage their result.

BoB thresholds While the KL thresholds reach the desired dependency in (t, δ) , using (3.9) instead of (3.4) yields larger thresholds due to additive constants. To overcome this hurdle, we maximize (3.4) under the per-arm box constraints (Lemma 3.5) and the pairwise non-linear constraint (Theorem 3.6). The resulting family of thresholds is denoted by BoB (Best of Both) thresholds. While the BoB thresholds have no closed-form solution, they can be approximated with non-linear solvers, *e.g.* Ipopt [Wächter and Biegler, 2006].

Corollary 3.7. *Let $f(x, y) = (1 + y)x - 1 - \log(x)$ for all $(x, y) \in (\mathbb{R}_+^\star)^2$. Let $(t_i^{Box})_i$ and $(t_i^m)_i$ as in (3.8,3.11). Let $\varepsilon_\mu, \varepsilon_{-,\sigma}$ as in Lemma 3.5 and $(c_{i,j}^{\mathrm{KL}})_{(i,j)\in[K]^2}$ as in (3.10). The family of thresholds $c_{i,j}^{BoB}(N_n, \delta)$ with value $+\infty$ if $n < \max_{k\in\{i,j\}} \max\{t_k^{Box}(\delta/6), t_k^m(\delta/2)\}$ and otherwise solution of the optimization problem*

$$\textit{maximize } \frac{1}{2} \sum_{k\in\{i,j\}} N_{n,k} \log(1 + y_k)$$

such that $\forall k \in \{i,j\}, \quad y_k \geq 0,\ x_k y_k \leq \varepsilon_\mu(N_{n,k}, \delta/2),\ x_k \geq 1 - \varepsilon_{-,\sigma}(N_{n,k} - 1, \delta/2)$,

and $\frac{1}{2} \sum_{k \in \{i,j\}} N_{n,k} f(x_k, y_k) \leq c_{i,j}^{\mathrm{KL}}(N_n, \delta/2)$,

yields a δ-correct family of thresholds for the GLR stopping rule on instances $\nu \in \mathcal{D}_{\mathcal{N}}^K$ with mean $\mu \in \mathcal{S}$.

Proof. The proof is obtained by combining the concentration results used to prove Lemma 3.5 and Theorem 3.6, and maximizing the empirical transportation costs under the constraints imposed by concentration. ■

Since (3.4) is smaller than (3.9), the KL constraint is an upper bound on the BoB threshold. Compared to the box threshold, the maximization underlying the BoB threshold has an additional constraint. Therefore, we have $c_{i,j}^{\mathrm{BoB}}(N_n, \delta) \leq \min\{c_{i,j}^{\mathrm{Box}}(N_n, \delta/2), c_{i,j}^{\mathrm{KL}}(N_n, \delta/2)\}$. In particular, the BoB threshold combines the best of both thresholds for the (n, δ) dependencies.

3.3.3 Simulations

We perform numerical simulations to compare the family of thresholds introduced above. Taking $K = 2$, we consider the instance $\mu = (0, -0.2)$ and $\sigma^2 = (1, 0.5)$. Since we are not interested in observing the influence of the sampling rule, the stream of data is uniform between both arms. For the thresholds, we set the parameters to $s = 2$, $\gamma = 1.2$ and $\eta_0 = \eta_1 = \log(1/\delta)^{-1}$.

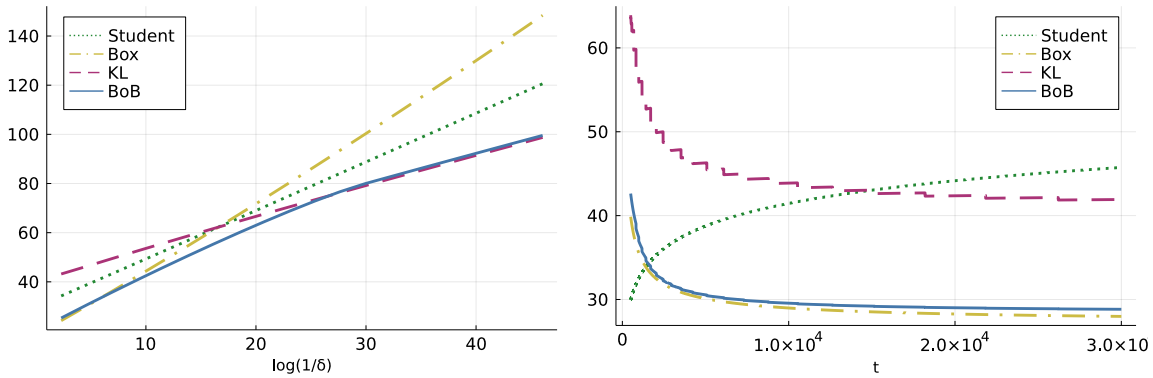

Figure 3.1 – Thresholds for (3.2) as a function of (a) $\log(1/\delta)$ for $n = 5000$ and (b) n for $\delta = 0.01$.

Figure 3.1 plots the dependency of the thresholds in $\log(1/\delta)$ and n . In Figure 3.1(a), we are interested by the slopes, and smaller slopes imply better dependency in $\log(1/\delta)$. As expected, Student thresholds have poor performance for both variables. While box thresholds

improve in n , they suffer from a worse dependency in $\log(1/\delta)$. KL thresholds circumvent this issue with the best dependency in $\log(1/\delta)$ so far. However, they incur a large constant cost making it worse than the box threshold. As hoped, BoB thresholds combine the good performance in n of the box threshold and the asymptotic dependency in $\log(1/\delta)$ of the KL threshold.

The improved theoretical dependency of the BoB threshold comes at the price of a higher computational cost: on average 400 , 600 and 800 times larger than for the KL threshold, the Box threshold and the Student threshold respectively. When the computational cost is a major concern, the Box threshold should be used since it has low computational cost and good empirical performance. Alternatively, we could use the BoB threshold and evaluate the stopping rule only on a geometric grid of times. This "lazy" stopping rule is still δ-correct.

3.4 Sampling Rule Wrappers

After calibrating the stopping threshold to ensure δ-correctness, we need to design a sampling rule which requires few samples before stopping. Given any BAI algorithm for Gaussian with known variances, we propose two wrappers that can adapt the algorithm to tackle unknown variances: plugging in the empirical variance or adapting the transportation cost.

When the variances are unknown, a natural idea is to plug in the empirical variances instead of using the true variances which are now unknown. We can apply this wrapper to any BAI algorithm. Section 3.2.1 discusses the differences and links between the transportation costs for known and unknown variances. Leveraging this interplay, we can adapt a BAI algorithm to use the transportation costs for unknown variances instead of the ones for known variances. We can apply this wrapper to any BAI algorithm relying on transportation costs. We illustrate how to instantiate each wrapper (Section 3.4.1), derive guarantees on their asymptotic expected sample complexity (Section 3.4.2), and assess their empirical performance (Section 3.4.3).

3.4.1 Instantiating the Wrappers

As initialization, we start by pulling each arm $n_0 \geq 2$ times.

Track-and-Stop The Track-and-Stop (TaS) algorithm [Garivier and Kaufmann, 2016] computes at each time $n > n_0 K$ the optimal allocation for the considered transportation costs, *i.e.* $w_n = w^\star(\nu_n; \sigma)$ for Gaussian with known variances. Given the vector w_n in the simplex, it uses a so-called tracking procedure to obtain an arm I_n to sample. We describe and use the one called C-tracking by Garivier and Kaufmann [2016]. On top of this tracking a forced exploration is used to enforce convergence towards the optimal allocation for the true unknown

parameters. Let $\varepsilon \in (0, 1/K]$ and $\Sigma_K^\varepsilon = \{w \in [\varepsilon, 1]^K \mid \sum_{i\in[K]} w_i = 1\}$. Defining w_n^ε the ℓ_∞ projection of w_n on Σ_K^ε , C-Tracking pulls $I_n \in \arg\max_{i\in[K]}\{\sum_{t=n_0K}^n w_{t,a}^\varepsilon - N_{n,i}\}$.

Plugging in the empirical variance yields the EV-TaS (Empirical Variance Track-and-Stop) algorithm which computes $w_n = w^\star(\mu, \sigma_n)$. Adapting the transportation cost yields the TaS algorithm which uses $w_n = w^\star(\nu_n)$. Computing $w^\star(\nu_n; \sigma_n)$ and $w^\star(\nu_n)$ can be done by solving an equivalent optimization problem with one bounded variable, which can itself be numerically approximated with binary search.

Top Two algorithm At each time $n > n_0K$, the Top Two algorithm β-EB-TCI [Jourdan et al., 2022] pulls the EB leader $B_n^{\text{EB}} = \hat{\imath}_n$ with probability β . If B_n^{EB} is not sampled, then it pulls the TCI challenger C_n^{TCI} , *i.e.* $C_n^{\text{TCI}} \in \arg\min_{i\neq B_n^{\text{EB}}} \{C(B_n^{\text{EB}}, i; \nu_n, N_n; \sigma) + \log N_{n,i}\}$ for Gaussian with known variance.

Plugging in the empirical variance yields the β-EB-EVTCI algorithm which uses

$$C_n^{\text{EVTCI}} \in \operatorname*{arg\,min}_{i\neq B_n^{\text{EB}}} \left\{ W_n^{\text{EV}}(B_n^{\text{EB}}, i) + \log N_{n,i} \right\} .$$

Adapting the transportation cost yields the β-EB-TCI algorithm which computes

$$C_n^{\text{TCI}} \in \operatorname*{arg\,min}_{i\neq B_n^{\text{EB}}} \left\{ W_n(B_n^{\text{EB}}, i) + \log N_{n,i} \right\} .$$

The two wrappers could be used similarly on other Top Two algorithms (see Section 2.2). It is straightforward to plug in the empirical variance, and also simple to adapt the transportation costs. For the Bayesian approach, a posterior distribution over Gaussian with unknown variance is required (see Honda and Takemura [2014], Cowan et al. [2017]), yet it is unknown if they satisfy the "good" properties to obtain asymptotic (β-)optimality. For the frequentist approach, it is simpler since we adapt the transportation costs. For example, the UCB indices can be written as $U_{n,i} = \mu_{n,i} + \sigma_{n,i}\sqrt{\exp(2g(n)/N_{n,i}) - 1}$. It is an open problem to study IDS for other distributions than Gaussian with known variance. For Gaussian with unknown variance, it is defined as $\beta_n(i, j) = 1/2$ when $\mu_{n,i} \leq \mu_{n,j}$, and

$$\beta_n(i, j) = \frac{N_{n,i}}{2W_n(i, j)} \log\left(1 + \frac{(\mu_{n,i} - u_{i,j}(\nu_n, N_n))^2}{\sigma_{n,i}^2}\right) \quad \text{otherwise} ,$$

with $u_{i,j}(\nu, w)$ defined in Lemma 2.4 as minimizer of the transportation cost.

3.4.2 Sample Complexity Upper Bound

Definition 3.8 introduces the notion of asymptotically tight family threshold [Jourdan et al., 2022], which corresponds informally to $c_{i,j}(N,\delta) \sim_{\delta\to 0} \log(1/\delta)$. As hinted in Figure 3.1(a), the KL and the BoB thresholds are asymptotically tight, but not the Student and Box thresholds.

Definition 3.8. *A family of thresholds $(c_{i,j})_{(i,j)\in[K]^2}$ is said to be asymptotically tight if there exists $\alpha \in [0,1)$, $\delta_0 \in (0,1]$, functions $f, \bar{T} : (0,1] \to \mathbb{R}_+$ and C independent of δ satisfying: (1) for all $(i,j) \in [K]^2$, $\delta \in (0,\delta_0]$ and $N \in \mathbb{N}^K$ such that $\|N\|_1 \geq \bar{T}(\delta)$, then $c_{i,j}(N,\delta) \leq f(\delta) + C\|N\|_1^\alpha$, (2) $\limsup_{\delta\to 0} f(\delta)/\log(1/\delta) \leq 1$ and $\limsup_{\delta\to 0} \bar{T}(\delta)/\log(1/\delta) = 0$.*

Combined with the GLR stopping rule using the KL or the BoB thresholds, Theorem 3.9 shows that TaS (resp. β-EB-TCI) is a δ-correct and asymptotically optimal (resp. β-optimal) algorithm.

Theorem 3.9. *Using the GLR stopping rule with an asymptotically tight family of thresholds, TaS (resp. β-EB-TCI with $n_0 \geq 4$) satisfies that, for all $\nu \in \mathcal{D}_{\mathcal{N}}^K$ with mean $\mu \in \mathbb{R}^K$ such that $|i^\star(\mu)| = 1$ (resp. $\min_{i\neq j} |\mu_i - \mu_j| > 0$),*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\log(1/\delta)} \leq T^\star(\nu) \qquad (\textit{resp. } T_\beta^\star(\nu)) \ .$$

Proof. For Gaussian with unknown variance, it is possible to show similar regularities of the characteristic times as for the case when the variance is known. In other words, we prove that a similar result as Lemma 2.11 holds for Gaussian with unknown variance. The main reason is that the $\mathcal{K}_{\inf}$ functions are closely connected, up to a function $x\log(1+x)$ which preserves the desired properties. Among other things, Lemma 2.11 in Chapter 2 also hold true for Gaussian with unknown variance.

As regards β-EB-TCI, the proof follows along the same lines as the one in Section 2.3 of Chapter 2, which was actually written in sufficient generality to cope for many distributions. Similarly as in Appendix B.12 (resp. Appendix B.13), for Gaussian with unknown variance, the EB leader (resp. the TCI challenger) answer satisfy the Properties 2.15 and 2.18 (resp. Properties 2.16 and 2.21).

As regards TaS, the proof uses similar arguments as the ones used in Garivier and Kaufmann [2016]. The sole technical modification lies in the definition of the concentration event which should also control the empirical variance. This can be done with the same concentrations results that yield the family of Box thresholds (Lemma 3.5).

In both cases, we refer the reader to Appendix H in Jourdan et al. [2023a] for more details. ■

Using exactly the same arguments as in the proof of Theorem 3.9, it is also possible to derive a similar result involving $T^\star(\nu;\sigma)$ (resp. $T^\star_\beta(\nu;\sigma)$) for EV-TaS (resp. β-EB-EVTCI with $n_0 \geq 6$) combined with the EV-GLR stopping rule using an asymptotically threshold. However, since $T^\star(\nu;\sigma) < T^\star(\nu)$ and $T^\star_\beta(\nu;\sigma) < T^\star_\beta(\nu)$, neither of these algorithms can be δ-correct. Otherwise it would yield a contradiction with the lower bound in Lemma 3.2. Moreover, as there exists instances for which the ratios $T^\star(\nu)/T^\star(\nu;\sigma)$ and $T^\star_\beta(\nu)/T^\star_\beta(\nu;\sigma)$ are arbitrarily large, multiplying the thresholds by a problem independent constant is not sufficient to obtain δ-correctness, as expressed in Theorem 3.10.

Theorem 3.10. *There exists a sampling rule such that: for all asymptotically tight family of thresholds $(c_{i,j})_{(i,j)\in[K]^2}$ and problem independent constant $\alpha_0 > 0$, combining this sampling rule with the EV-GLR stopping rule using $(\alpha_0 c_{i,j})_{(i,j)\in[K]^2}$ yields an algorithm which is not δ-correct.*

Inspired by Section 3.3, we could also propose families of thresholds (EV-Student, EV-Box and EV-BoB) which are δ-correct for the EV-GLR stopping rule but are not asymptotically tight (Theorem 3.10). Still, in our experiments the empirical proportion of error is lower than δ even when using a heuristic, asymptotically tight threshold.

Based on Theorems 3.9 and 3.10, algorithms obtained by adapting the transportation costs enjoy stronger theoretical guarantees than the ones plugging in the empirical variance.

3.4.3 Experiments

We compare the empirical performance of the two wrappers for different BAI algorithms in the moderate regime ($\delta = 0.01$). As benchmarks, we consider FHN2 (procedure 2 in Fan et al. [2016]), uniform sampling and "fixed" sampling which is an oracle playing with proportions $w^\star(\nu)$. FHN2 is an elimination strategy which repeatedly samples all arms until only one arm is left. Its elimination mechanism is calibrated by resorting to continuous-time approximations. Therefore, FHN2 is only asymptotically δ-correct and has no guaranties on the sample complexity. Based on Degenne et al. [2019] and Wang et al. [2021], plugging in the empirical variance yields EV-DKM and EV-FWS, while DKM and FWS refers to the algorithms using the transportation costs for unknown variances. Even though those instances are not analyzed, we believe that similar guarantees on the sample complexity can be shown.

Algorithms obtained by plugging in the empirical variance uses the EV-GLR stopping rule, while the GLR stopping rule is used by the ones with adapted transportation cost and the

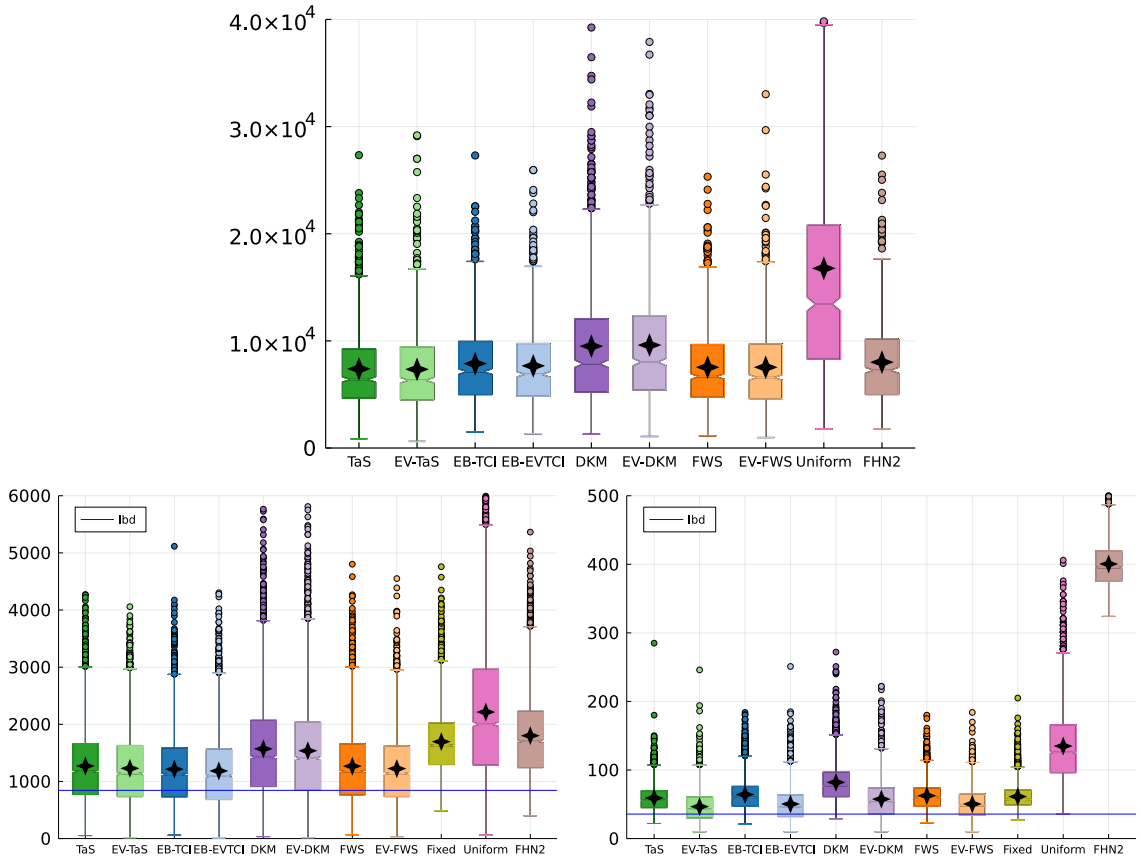

Figure 3.2 – Empirical stopping time on Gaussian (top) random instances with $K = 10$, (left) standard instance $(\mu, \sigma^2) = ((1.0, 0.85, 0.8, 0.7, 0.65)$, $(1.0, 0.6, 0.5, 0.4, 0.35))$ and (right) easy instance $(\mu, \sigma^2) = ((1.0, 0.2, 0.15, 0.1, 0.05)$, $(1.0, 0.05, 0.05, 0.05, 0.05))$. Lower bound is $T^\star(\nu) \log(1/\delta)$.

uniform sampling. We consider the stylized stopping threshold $c(n, \delta) = \log((1 + \log n)/\delta)$, which was proposed in Garivier and Kaufmann [2016]. While it doesn't ensure δ-correctness of the stopping threshold, it is asymptotically tight and yields an empirical error which is several order of magnitude lower than δ . Top Two algorithms use $\beta = 0.5$.

We assess the performance on 1000 random instances with $K = 10$ such that $(\mu_1, \sigma_1^2) = (0, 1)$. For $i \neq 1$, we set $(\mu_i, \sigma_i^2) = (-\Delta_i, r_i)$ where $\Delta_i \sim \mathcal{U}([0.2, 1.0])$ and $r_i \sim \mathcal{U}([0.1, 10])$. To illustrate the two regimes for $T^\star(\nu)/T^\star(\nu; \sigma)$, we consider a *standard* instance with $T^\star(\nu)/T^\star(\nu; \sigma) \approx 1.015$ and an *easy* instance with $T^\star(\nu)/T^\star(\nu; \sigma) \approx 1.384$. We average over 5000 runs.

In Figure 3.2, we observe that algorithms obtained by plugging in the empirical variance yield similar result as the ones using the adapted transportation cost, and slightly better performance on the easy instance. Moreover, those wrapped BAI algorithms outperforms uniform sampling and are on par with "fixed" sampling. On random instances FHN2 has

similar performance to the wrapped BAI algorithms, but it wastes precious samples on easy instances.

3.5 Discussion

In Chapter 3, we presented two approaches to deal with unknown variances, either by plugging in the empirical variance or by adapting the transportation costs. New time-uniform concentration results were derived to calibrate our two stopping rules. Then, we showed theoretical guarantees and competitive empirical performance of our two sampling rules wrappers on two existing algorithms.

While the literature abounds with designs of sampling rule, the optimal calibration of stopping rules is a most pressing issue as it leads to lower empirical stopping time. While calibrated thresholds have been derived with (near) optimal dependency in δ , those thresholds are known to be too conservative in the moderate confidence regime where their empirical error rate is orders of magnitude lower than δ . Avoiding this bottleneck on the expected sample complexity is an interesting open problem.

In the fixed-budget setting, characterizing the impact of not knowing the variances on the probability of misidentifying the best-arm is still an open problem. While similar approaches might be used to deal with the unknown variances, the resulting algorithms might not enjoy similar theoretical guarantees and empirical performance.

Finally, while this chapter goes beyond the class of one-parameter exponential families used in Chapter 2, it still considers a parametric class of distributions. In many applications, this is too restrictive and non-parametric classes of distributions should be considered. It is natural to wonder whether the approach of adapting the transportation costs can be extended. In Chapter 4, we will answer by the affirmative for the class of bounded distributions.

Chapter 4

Beyond Parametric Distributions

In Chapter 4, we study the vanilla BAI problem for the non-parametric class of bounded distributions, as studied in Chapters 2 and 3 for parametric distributions. The presented results were published in Jourdan et al. [2022].

Top Two algorithms arose as an adaptation of Thompson sampling to best arm identification in multi-armed bandit models [Russo, 2016] for parametric families of arms. Despite their good empirical performance, theoretical guarantees for fixed-confidence best arm identification have only been obtained for parametric distributions. In this chapter, we are interested in (near) optimal and computationally efficient strategies when the distributions belong to a non-parametric class of distributions. As an example motivated by applications in agriculture, we consider the set of bounded distributions. We instantiate several Top Two algorithms introduced in Chapter 2 for this class, and prove their asymptotic β-optimality.

Contents

4.1 Introduction

As detailed in Chapter 1, the motivation to study BAI for Gaussian distributions with non-parametric distributions comes from practical consideration. For applications to online marketing such as A/B testing [Kaufmann et al., 2014, Russac et al., 2021] assuming Bernoulli or Gaussian arms is fine, but more sophisticated distributions arise in other fields such as agriculture. In Section 4.4 we consider a crop-management problem: a group of farmers wants to identify the best planting date for a rainfed crop. The reward (crop yield) can be modeled as a complex distribution with multiple modes, but upper bounded by a known yield potential. Therefore, sequentially identifying the best planting date calls for efficient best arm identification algorithms for the class of bounded distributions with a known range.

We consider the set $\mathcal{D}_{[0,B]}$ of bounded distributions with support in $[0, B]$ with $B > 0$, hence $\mathcal{D}^K = \mathcal{D}^K_{[0,B]}$. Importantly, all distribution in $\kappa \in \mathcal{D}$ have a finite mean denoted by $m(\kappa)$. In the following, we assume that the set $\mathcal{I} = \{m(\kappa) \mid \kappa \in \mathcal{D}\}$ of possible means is such that $\mathcal{I} \subseteq (0, B)$, *i.e.* we exclude the Dirac distributions in $\{0\}$ and $\{B\}$. Let $\nu \in \mathcal{D}^K$ with mean $\mu = m(\nu)$ such that the set of arms with largest mean $i^\star(\mu) := \arg\max_{i\in[K]} \mu_i$ is reduced to a singleton denoted by $i^\star$ (or $i^\star(\mu)$ by abusing notation), *i.e.* $\mathcal{S} = \{\mu \in (0, B)^K \mid |i^\star(\mu)| = 1\}$.

A fixed-confidence algorithm is defined by a sampling rule, a recommendation rule and a stopping rule. At time n , we denote by $\hat{\imath}_n$ the candidate answer and by I_n the arm to pull. The stopping rule (and stopping time τ_δ) using a fixed confidence level $1 - \delta \in (0, 1)$ which ensures δ-correctness, *i.e.* $\mathbb{P}_\nu\left(\tau_\delta < +\infty,\ \hat{\imath}_{\tau_\delta} \neq i^\star(\mu)\right) \leq \delta$ for all instances $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathcal{S}$.

As discussed in Section 1.4.1, the δ-correctness requirement leads to a lower bound on the expected sample complexity on any instance.

Lemma 4.1 (Garivier and Kaufmann [2016] and Agrawal et al. [2020]). *An algorithm which is δ-correct on all problems in $\mathcal{D}^K_{[0,B]}$ satisfies that, for all $\nu \in \mathcal{D}^K_{[0,B]}$ with mean $\mu \in \mathcal{S}$, $\mathbb{E}_\nu[\tau_\delta] \geq T^\star(\nu) \log(1/(2.4\delta))$.*

For bounded distributions, Lemma 4.1 shows that $T^\star(\nu)$ is the asymptotic complexity of the BAI problem, where

$$T^\star(\nu)^{-1} = \sup_{w\in\Sigma_K} \min_{j\neq i} C(i, j; \nu, w) \quad \text{with}$$
$$C(i, j; \nu, w) = \mathbb{1}\left(\mu_i \geq \mu_j\right) \inf_{u\in(0,B)} \left\{w_i \mathcal{K}^-_{\inf}(\nu_i, u) + w_j \mathcal{K}^+_{\inf}(\nu_j, u)\right\} .$$

For one-parameter exponential families or Gaussian with unknown variance, the $\mathcal{K}_{\inf}$ functions have a closed form expression that can be computed efficiently. For bounded distributions, this

is not the case, yet the dual formulation obtained by Honda and Takemura [2010] offers a more explicit expression, *i.e.*

$$\mathcal{K}_{\inf}^{+}(\kappa, u) = \mathbb{1}\left(m(\kappa) \leq u\right) \sup_{\lambda \in [0,1]} \mathbb{E}_{X\sim\kappa}\left[\log\left(1 - \lambda\frac{X-u}{B-u}\right)\right] \quad \text{and} \tag{4.1}$$
$$\mathcal{K}_{\inf}^{-}(\kappa, u) = \mathbb{1}\left(m(\kappa) \geq u\right) \sup_{\lambda \in [0,1]} \mathbb{E}_{X\sim\kappa}\left[\log\left(1 + \lambda\frac{X-u}{u}\right)\right] ,$$

which are computationally expensive to evaluate when κ is a continuous distribution.

As in previous chapters, we say that an algorithm is asymptotically optimal (resp. β-optimal) on $\mathcal{D}^K$ if it is δ-correct and its sample complexity matches that lower bound, *i.e.* for all $\nu \in \mathcal{D}^K_{[0,B]}$ such that $\mu \in \mathcal{S}$, $\limsup_{\delta\to 0} \mathbb{E}_\nu[\tau_\delta]/\log(1/\delta) \leq T^\star(\nu)$ (resp. $T^\star_\beta(\nu)$). For $\beta \in (0,1)$, the definition of $T^\star_\beta(\nu)$ is the same as $T^\star(\nu)$ with the additional constraint on the outer maximization that $w_{i^\star} = \beta$, hence $T^\star(\nu) = \min_{\beta\in(0,1)} T^\star_\beta(\nu)$. For bounded distributions, we show that $T^\star_{1/2}(\nu) \leq 2T^\star(\nu)$ holds (Russo [2016] showed it only for one-parameter exponential families).

Contribution 4.2. *In Chapter 4, we propose a calibration of the GLR stopping rule and a concrete instantiation of the Top Two approach, based on a Dirichlet sampler for the randomized variants.*

- *We prove in Theorem 4.4 that those algorithms are asymptotically β-optimal. The proof follows the unified analysis provided in Chapter 2, hence showing that the sufficient properties on the leader and the challenger are satisfied. In particular, this requires to derive strong properties on the $\mathcal{K}_{\inf}$ functions for bounded distributions and on the Dirichlet sampler, which are of independent interest.*
- *We report results from numerical experiments on a challenging non-parametric task using real-world data from a crop-management problem for various members of the Top Two family of algorithms. Most of them perform significantly better than the baselines.*

4.1.1 GLR Stopping Rule

For an arm i , we denote its number of pulls by $N_{n,i} := \sum_{t\in[n-1]} \mathbb{1}\left(I_t = i\right)$, its empirical distribution by $\nu_{n,i} := N_{n,i}^{-1} \sum_{t\in[n-1]} \delta_{X_{t,I_t}} \mathbb{1}\left(I_t = i\right)$ and its empirical mean by $\mu_{n,i} := m(\nu_{n,i})$. Recall that $\mathcal{F}_n$ denotes the whole history before time n , which includes internal randomization of the algorithm at time n . For all $\mathcal{F}_n$-measurable sets A , let $\mathbb{P}_{|n}[A] := \mathbb{P}[A \mid \mathcal{F}_n]$ be its probability.

As candidate answer, we use $\hat{\imath}_n \in i^\star(\mu_n)$, *i.e.* the empirical best arm (EB). For the stopping rule, we use the GLR stopping rule (see Section 1.4.2 for more details). For bounded distri-

butions, the GLR can be written as $\min_{i \neq \hat{\imath}_n} W_n(\hat{\imath}_n, i)$, where the empirical transportation cost between arm i and arm j is defined as

$$W_n(i,j) = C(i,j;\nu_n,N_n) = \mathbb{1}\left(\mu_{n,i} \geq \mu_{n,j}\right) \inf_{u \in (0,B)} \left\{ N_{n,i} \mathcal{K}_{\inf}^{-}(\nu_{n,i},u) + N_{n,j} \mathcal{K}_{\inf}^{+}(\nu_{n,j},u) \right\} . \tag{4.2}$$

Given a threshold function $c(n,\delta)$, the GLR stopping rule is

$$\tau_\delta = \inf \left\{ n \in \mathbb{N} \mid \min_{j \neq \hat{\imath}_n} W_n(\hat{\imath}_n, j) > c(n-1,\delta) \right\} . \tag{4.3}$$

Lemma 4.3 gives a threshold ensuring that the GLR stopping rule is δ-correct for all $\delta \in (0,1)$, independently of the sampling rule. Its proof relies on an elegant martingale construction proposed by Agrawal et al. [2021b], and it is detailed in Appendix D.1.

Lemma 4.3. *Let $\delta \in (0,1)$. Given any sampling rule, using the threshold*

$$c(n,\delta) = \log\left(1/\delta\right) + 2\log\left(1+n/2\right) + 2 + \log(K-1) \tag{4.4}$$

with the stopping rule (4.3) yields a δ-correct algorithms for the set of bounded distributions with mean in $\mathcal{S}$.

Empirical transportation costs Evaluating the stopping rule requires the computation of the empirical transportation costs $W_n(i,j)$ defined in (4.2). As discussed above, the $\mathcal{K}_{\inf}$ functions for bounded distributions are more expensive to compute than their counterparts for one-parameter exponential families or Gaussian with unknown variance. Fortunately, we only evaluate those functions on empirical distributions, which are supported on a set points, *i.e.*

$$N_{n,j} \mathcal{K}_{\inf}^{+}(\nu_{n,j},u) = \mathbb{1}\left(\mu_{n,j} \leq u\right) \sup_{\lambda \in [0,1]} \sum_{t \in [n-1]} \mathbb{1}\left(I_t = j\right) \log\left(1 - \lambda \frac{X_{t,j} - u}{B - u}\right) \quad \text{and}$$

$$N_{n,i} \mathcal{K}_{\inf}^{-}(\nu_{n,i},u) = \mathbb{1}\left(\mu_{n,i} \geq u\right) \sup_{\lambda \in [0,1]} \sum_{t \in [n-1]} \mathbb{1}\left(I_t = i\right) \log\left(1 + \lambda \frac{X_{t,i} - u}{u}\right) .$$

The minimization in λ can be computed using a zero-order optimization algorithm (*e.g.* Brent's method Brent [2013]). The same optimizer can be used to compute the minimization in $u \in (0,B)$ featured in $W_n(i,j)$. By nesting those optimizations of univariate functions on a bounded interval, the computation of $W_n(i,j)$ in the stopping rule dominates the computational cost of our Top Tow algorithms (except the RS challenger, see below). Our experiments suggest that using (4.3) is twice as computationally expensive as the LUCB-based stopping rule, which is a mild price to pay for the improvement in terms of empirical stopping time. Algorithms

for non-parametric distributions are bound to be computationally more expensive than their parametric counterpart, where a vector of sufficient statistics can summarize the history $\mathcal{F}_n$.

4.2 Top Two Algorithms

As summarized in Algorithm 2.1, a Top Two sampling rule is defined by four components: leader answer, challenger answer, target allocation and mechanism to reach it. We refer the reader to Section 2.2 for more details on specific instances of those four choices, which were written in a generality that includes non-parametric classes of distributions such as bounded distributions.

Dirichlet sampler The TS leader and RS challenger require a sampler Π_n (see Section 2.2). Our proposed Dirichlet sampler for bounded distributions in $[0, B]$ has a product form: $\Pi_n = \Pi_{n,1} \times \cdots \times \Pi_{n,K}$ where $\Pi_{n,i}$ leverages $\mathcal{H}_{n,i} := (X_{1,i}, \dots, X_{N_{n,i},i})$, which is the history of samples from arm i collected in the first $n-1$ rounds. Let $\tilde{\nu}_{n,i}$ denote the empirical cdf of $\mathcal{H}_{n,i}$ augmented by the known bounds on the support, *i.e.* $\{0, B\}$. For each arm i , $\Pi_{n,i}$ outputs a random re-weighting of $\tilde{\nu}_{n,i}$. Concretely, letting $(w_1, \dots, w_{N_{n,i}+2})$ be drawn from a Dirichlet distribution $\mathrm{Dir}(\mathbf{1}_{N_{n,i}+2})$, a call to the sampler $\Pi_{n,i}$ returns

$$\sum_{t \in [N_{n,i}]} w_t X_{t,i} + B w_{N_{n,i}+1} \, .$$

This sampler is inspired by that used in the Non Parametric Thompson Sampling (NPTS) algorithm proposed by Riou and Honda [2020] for regret minimization in bounded bandits, with the notable difference that we have to add both 0 and B in the support, while NPTS only adds the upper bound B . Since adding 0 is only necessary to ensure that the re-sampling procedure stops (*i.e.* RS challenger), the TS leader could use a sampler based directly on $\mathcal{H}_{n,i}$.

Leader answer The EB leader selects $B_n^{\mathrm{EB}} = \hat{\imath}_n \in \arg\max_{i \in [K]} \mu_{n,i}$. The UCB leader uses $B_n^{\mathrm{UCB}} \in \arg\max_{i \in [K]} U_{n,i}$ with

$$U_{n,i} = \left\{ u \in [\mu_{n,i}, B] \mid N_{n,i} \mathcal{K}_{\inf}^{+}(\nu_{n,i}, u) \leq g(n) \right\} \, ,$$

where $g(n) = \Theta(\log n)$. For regret minimization, it was studied in Agrawal et al. [2021a]. The TS leader chooses $B_n^{\mathrm{TS}} \in i^\star(\theta_n)$ with $\theta_n \sim \Pi_n$.

Challenger answer The TC and TCI challenger consider

$$C_n^{\mathrm{TC}} \in \arg\min_{i \neq B_n} W_n(B_n, i) \quad \text{and} \quad C_n^{\mathrm{TCI}} \in \arg\min_{i \neq B_n} \{W_n(B_n, i) + \log N_{n,i}\} \ .$$

The RS challenger takes *i.e.* $C_n^{\mathrm{RS}} \in \arg\max_{i \in [K]} \theta_{n,i}$ with $\theta_n \sim \Pi_n$ until $B_n \notin i^\star(\theta_n)$.

Target allocation over arms In terms of fixed design, one can either fix the proportion to β or sample the least sampled arm. Note that

$$\arg\min_{i \in \{B_n, C_n\}} N_{n,i} \neq \arg\max_{i \in [K]} C(B_n, C_n; \nu_n, N_n + 1_i) \neq \arg\max_{i \in [K]} \frac{\partial C(B_n, C_n; \nu_n, N_n)}{\partial w_i} \ ,$$

for distributions other than Gaussian with unit variance, hence other fixed designs could be defined. It is an open problem to study IDS for other distributions than Gaussian with known variance. For bounded distributions, it is defined as $\beta_n(i,j) = 1/2$ when $\mu_{n,i} \leq \mu_{n,j}$, and

$$\beta_n(i,j) = \frac{N_{n,i} \mathcal{K}_{\inf}^-(\mu_{n,i}, u_{i,j}(\nu_n, N_n))}{W_n(i,j)} \quad \text{otherwise} \ ,$$

with $u_{i,j}(\nu, w)$ defined in Lemma 2.4 as minimizer of the transportation cost.

Mechanism to reach the target allocation Both randomization and tracking can be used since they yield similar empirical performance and theoretical guarantees.

4.3 Asymptotic Sample Complexity Upper Bound

In BAI for bounded distributions with known support $[0, B]$, Theorem 4.4 shows the asymptotic β-optimality of many Top Two sampling rule when combined with the GLR stopping rule.

Theorem 4.4. *Let $(\beta, \delta) \in (0,1)^2$. Combined with the GLR stopping rule (4.3) using the threshold (4.4), the Top Two sampling rule (Algorithm 2.1) using (i) any leader in $\{EB, UCB, TS\}$, (ii) any challenger in $\{TC, TCI, RS\}$, (iii) the fixed design β , and (iv) randomization (or tracking for a deterministic leader/challenger pair, see Section 2.2.4) yields an algorithm which is δ-correct and satisfies that, for all $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathbb{R}^K$ such that $\overline{\Delta}_{\min}(\mu) := \min_{i \neq j} |\mu_i - \mu_j| > 0$,*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\log(1/\delta)} \leq T_\beta^\star(\nu) \ .$$

Choosing a Top Two algorithm Choosing our favorite Top Two instances requires further empirical and computational considerations. Computing the EB leader has a constant computational cost, while the TS leader is computationally costly for large time n since it requires to sample from a Dirichlet distribution with $N_{n,i}+2$ parameters for each arm i . On the challenger side, the RS challenger is computationally very expensive for large time n as the sampler becomes concentrated around the true mean vector. On the contrary, by leveraging computations done in the stopping rule (4.3), the TC and TCI challengers can be computed in constant time. Based on these computational considerations, the most appealing Top Two algorithm for bounded distribution appears to be EB-TC-β . But experiments performed in Section 4.4 reveal its lack of robustness, and for bounded distributions the best trade-off between robustness and computational complexity is EB-TCI-β . More generally, TS-TC-β can also be a good choice provided that we have access to an efficient sampler.

Proof outline The proof of Theorem 4.4 follows from the same unified analysis as presented in Section 2.3 of Chapter 2. In Jourdan et al. [2022], we proposed the unified analysis for fixed design β precisely to tackle bounded distributions. We present important regularities properties (Section 4.3.1), mention why the analysis still works for EB-TC-β (Section 4.3.2), and explain why the Dirichlet sampler has the desired properties for bounded distributions (Section 4.3.3).

4.3.1 Regularity Properties

For detailed proof of the following properties for bounded distributions $\mathcal{D}_{[0,B]}$, we refer the reader to Appendix F in Jourdan et al. [2022].

The $\mathcal{K}_{\inf}$ functions When studying a class of non-parametric distributions (such as bounded distributions), we need to derive strong regularity properties on the $\mathcal{K}_{\inf}$ functions. This is a key challenge to tackle BAI with classes of non-parametric distributions. It is an open problem to clearly define what is the minimal set of regularity properties on the class of non-parametric distribution $\mathcal{D}$ to obtain sufficient regularity to study BAI algorithms, such as Track-and-Stop or Top Two algorithms.

Since $\mathcal{K}^-_{\inf}(\kappa,u) = \mathcal{K}^+_{\inf}(\kappa^{B-X}, B-u)$ where κ^{B-X} is the pushforward measure of κ through $x \to B-x$, we can restrict the analysis to studying $\mathcal{K}^+_{\inf}$. First, it is required to show the dual formulation of the $\mathcal{K}^+_{\inf}$ as in (4.1), which was done by Honda and Takemura [2010]. Leveraging this dual formulation, we obtain that $(\kappa,u) \to \mathcal{K}^+_{\inf}(\kappa,u)$ is continuous on $\mathcal{D}_{[0,B]} \times [0,B)$ (by adapting a method proposed by Agrawal et al. [2021b] in a slightly different setting) and convex on $\mathcal{D}_{[0,B]} \times [0,B]$.

Lemma 4.5 (Theorems 5 and 6 in Honda and Takemura [2010]). *Let $\kappa \in \mathcal{D}_{[0,B]}$. Then, $u \to \mathcal{K}_{\inf}^{+}(\kappa, u)$ is differentiable on $(m(\kappa), B]$ and*

$$\frac{\partial \mathcal{K}_{\inf}^{+}(\kappa, u)}{\partial u} = \lambda_{\star}^{+}(\kappa, u) \quad \textit{with} \quad \lambda_{\star}^{+}(\kappa, u) = \underset{\lambda \in [0,(B-u)^{-1}]}{\arg\max} \mathbb{E}_{X\sim\kappa}\left[\log\left(1 - \lambda(X-u)\right)\right] . \quad (4.5)$$

Let $u^{+}(\kappa) = B - \mathbb{E}_{X\sim\kappa}[(B-X)^{-1}]^{-1} \geq m(\kappa)$. Then, (i) $\lambda_{\star}^{+}(\kappa, u) = 0$ if and only if $u \leq m(\kappa)$, (ii) $u \in (m(\kappa), u^{+}(\kappa)]$ implies that $\mathbb{E}_{X\sim\kappa}[(1 - \lambda_{\star}^{+}(\kappa, u)(X-u))^{-1}] = 1$ and (iii) $\lambda_{\star}^{+}(\kappa, u) = (B-u)^{-1}$ if and only if $u \geq u^{+}(\kappa)$.

Our key novel result on the regularity of $u \to \mathcal{K}_{\inf}^{+}(\kappa, u)$ is its strict convexity on $(m(\kappa), B]$ (Lemma 4.6). The proof of Lemma 4.6 is detailed in Appendix D.2, it differs from the proof strategy proposed by Agrawal et al. [2021b] in a slightly different setting.

Lemma 4.6. *Let $\kappa \in \mathcal{D}_{[0,B]}$. The function $u \to \mathcal{K}_{\inf}^{+}(\kappa, u)$ is strictly convex and increasing on $(m(\kappa), B]$, and null on $[0, m(\kappa)]$.*

Characteristic times Leveraging the regularity properties on the $\mathcal{K}_{\inf}$ functions, we obtain regularity properties on the transportation costs, and on the characteristic times. In a nutshell, we proved similar properties for bounded distributions as the ones detailed for Gaussian with known variance in Lemma 2.11.

4.3.2 A Pedagogical Example: EB-TC

In Section 2.3.5, we presented the proof for the EB leader and the TC challenger. It is straightforward to see that the same arguments can be applied for the EB leader, to obtain both Properties 2.15 and 2.18.

For the TC challenger, the regularity properties of the transportation costs (see Section 4.3.1 for details) will allow to prove an equivalent of Lemma 2.23. To prove an equivalent of Lemma 2.24, we rely on similar arguments and use that $\mathcal{K}_{\inf}^{+}(\kappa, u) \leq -\log(1 - u/B)$ for all $(\kappa, u) \in \mathcal{D}_{[0,B]} \times [0, B)$ (Lemma 14 in Honda and Takemura [2010]). Combining those two lemmas on the rates of growth of the empirical transportation costs, we can show that Property 2.15 holds. Thanks to the equivalent of Lemma 2.11, the proof of Property 2.21 is similar. Let i such that $N_{n,i}/N_{n,i^\star} \geq \gamma + w_{\beta,i}^{\star}/\beta$, and $j \notin \{i^\star, i\}$ such that $N_{n,j}/N_{n,i^\star} \leq w_{\beta,j}^{\star}/\beta$ (which exists). Using that $w \to C(i^\star, i; \kappa, w)$ is increasing and the equality at equilibrium, we

obtain that

$$\frac{W_n(i^\star, i)}{W_n(i^\star, j)} \geq \frac{C(i^\star, i; \nu_n, w_\beta^\star + \gamma\beta 1_{\{i\}})}{C(i^\star, j; \nu_n, w_\beta^\star)} \underset{n\to+\infty}{\approx} \frac{C(i^\star, i; \nu, w_\beta^\star + \gamma\beta 1_{\{i\}})}{C(i^\star, j; \nu, w_\beta^\star)} = \frac{C(i^\star, i; \nu, w_\beta^\star + \gamma\beta 1_{\{i\}})}{C(i^\star, i; \nu, w_\beta^\star)} > 1 \,.$$

Since $W_n(i^\star, i) > W_n(i^\star, j)$, we conclude that $C_n \neq i$ (*i.e.* Property 2.21 holds true).

The generalization to the UCB leader or the TCI challenger is straightforward. We refer the reader to Appendix D.1 in Jourdan et al. [2022] for more details.

4.3.3 Dirichlet Sampler

As in Section 2.3.5, we prove the desired properties for the TS leader and RS challenger. Since

$$\max_{u\in\mathbb{R}}\{\mathbb{P}_{\Pi_n|n}(\theta_{n,i} \geq u) \prod_{j\neq i} \mathbb{P}_{\Pi_n|n}(\theta_{n,j} \leq u)\} \leq \mathbb{P}_{\Pi_n|n}(i \in i^\star(\theta_n)) \leq 1 - \max_{j\neq i} \mathbb{P}_{\Pi_n|n}(\theta_{n,i} < \theta_{n,j}) \,,$$

we need to control (i) the Boundary Crossing Probability (BCP) of an arm, *i.e.* $\mathbb{P}_{\Pi_n|n}(\theta_{n,j} \leq u)$ and $\mathbb{P}_{\Pi_n|n}(\theta_{n,i} \geq u)$ where u is a fixed threshold, and (ii) the probabilities that the empirical ordering is reversed, *i.e.* $\mathbb{P}_{\Pi_n|n}(\theta_{n,i} < \theta_{n,j})$ when $\mu_i > \mu_j$. Lemma 4.7 shows that we can obtain guarantees on $\mathbb{P}_{\Pi_n|n}(\theta_{n,i} < \theta_{n,j})$ by using BCP (see Lemma 64 in Jourdan et al. [2022]).

Lemma 4.7. *Let* $x \in \arg\max_{u\in\mathbb{R}} \mathbb{P}_{\kappa_2}(\theta_2 \geq u)\mathbb{P}_{\kappa_1}(\theta_1 \leq u)$ *and* $g(u) = u(1 - \log(u))$ *for all* $u \in [0, 1]$. *Then,* $\mathbb{P}_{\kappa_2}(\theta_2 \geq x)\mathbb{P}_{\kappa_1}(\theta_1 \leq x) \leq \mathbb{P}_{(\kappa_1,\kappa_2)}(\theta_2 \geq \theta_1) \leq g\left(\mathbb{P}_{\kappa_2}(\theta_2 \geq x)\mathbb{P}_{\kappa_1}(\theta_1 \leq x)\right)$.

Combining Lemma 4.7 with the proof technique of Lemma 15 in Riou and Honda [2020] to derive a upper bound on the BCP, we show the following concentration result

$$\mathbb{P}_{\theta_n\sim\Pi_n|n}(\theta_{n,j} \geq \theta_{n,i}) \leq f\left(C(i, j; \nu_n, N_n + 2 \cdot 1_{[K]})\right) \,, \tag{4.6}$$

where $f(x) = (1 + x)\exp(-x)$. Using (4.6), we can show that Properties 2.15 and 2.18 hold for the TS leader answer. Property 2.16 is proven by using (4.6) and the coarse anti-concentration result, *i.e.* $\mathbb{P}_{\theta_n\sim\Pi_n|n}(\theta_{n,i} \geq u) \geq (1 - u/B)^{N_{n,i}+1}$. Property 2.21 is proven by using (4.6) and a tight anti-concentration result, *i.e.* $\mathbb{P}_{\theta_n\sim\Pi_n|n}(\theta_{n,i} \geq \theta_{n,i^\star}) \gtrapprox \exp(-W_n(i^\star, i))$. We refer the reader to Appendices D.2 and G in Jourdan et al. [2022] for more details.

4.4 Experiments

We assess the empirical performance of our Top Two algorithms on the the DSSAT simulator[1] [Hoogenboom et al., 2019] and on Bernoulli instances in the moderate regime ($\delta = 0.01$). The set $\mathcal{D}_{\mathcal{B}}$ of Bernoulli distributions is an example of one-parameter exponential family which is contained in $\mathcal{D}_{[0,1]}$. The stopping rule (4.3) is used with the threshold $c(n, \delta)$ defined in (4.4)). As Top Two sampling rules, we present results for EB-TC-$1/2$, EB-TCI-$1/2$, TS-TC-$1/2$ and TS-TCI-$1/2$, hence we will omit the target $\beta = 1/2$ in the names.

As benchmarks for the sampling rule, we use KL-LUCB with Bernoulli divergence [Kaufmann and Kalyanakrishnan, 2013] (whose theoretical guarantees extend to any distribution bounded in $[0, 1]$), "fixed" sampling which is an oracle playing with proportions $w^\star(\nu)$ and uniform sampling. We also propose a heuristic adaptation of the DKM algorithm [Degenne et al., 2019] (which is asymptotically optimal for one-parameter exponential families) to tackle bounded distributions, which we denote by $\mathcal{K}_{\inf}$-DKM, and uses forced exploration instead of optimism. Inspired by the regret minimization algorithm $\mathcal{K}_{\inf}$-UCB [Agrawal et al., 2021a], we propose its LUCB variant Kalyanakrishnan et al. [2012], named $\mathcal{K}_{\inf}$-LUCB. The upper/lower confidence indices are obtained by inverting of $\mathcal{K}_{\inf}^{\pm}$, *i.e.*

$$\forall i \neq \hat{\imath}_n, \quad U_{n,\delta,i} = \max\left\{u \in [\mu_{n,i}, B] \mid N_{n,i}\mathcal{K}_{\inf}^{+}(F_{n,i}, u) \leq c(n-1,\delta)\right\} ,$$
$$L_{n,\delta,\hat{\imath}_n} = \min\left\{u \in [0, \mu_{n,\hat{\imath}_n}] \mid N_{n,\hat{\imath}_n}\mathcal{K}_{\inf}^{-}(F_{n,\hat{\imath}_n}, u) \leq c(n-1,\delta)\right\} .$$

LUCB-based algorithms [Kalyanakrishnan et al., 2012] use their own stopping rule, namely they stop when $L_{n,\delta,\hat{\imath}_n} \geq \max_{j \neq \hat{\imath}_n} U_{n,\delta,j}$. For Bernoulli distributions, $\mathcal{K}_{\inf}$-LUCB recovers KL-LUCB. While being asymptotically optimal for heavy-tailed distributions Agrawal et al. [2020] with an adequate stopping threshold, the Track-and-Stop algorithm is computationally intractable for bounded distributions as it requires to compute $w^\star(\nu_n)$ at each time n (or on a geometric grid). We hence omit it from our experiments.

Crop-management task In DSSAT, each arm corresponds to a choice of planting date and fixed soil conditions. To illustrate the problem's difficulty we represent an empirical estimate (independent of the runs of our algorithms) of the yield distributions in Figure 4.1(b). Since the gaps between means are small, the identification problem is hard. Moreover, $\mathcal{K}_{\inf}$ computations for non-parametric distributions are costlier than Bernoulli ones, so we only present the results for 100 runs.

In Figure 4.1, EB-TCI, TS-TC and TS-TCI slightly outperform $\mathcal{K}_{\inf}$-DKM and the fixed (oracle) sampling rule. Moreover, $\mathcal{K}_{\inf}$-LUCB performs significantly worse than uniform sampling. Due

[1] DSSAT is an Open-Source project maintained by the DSSAT Foundation, see https://dssat.net.

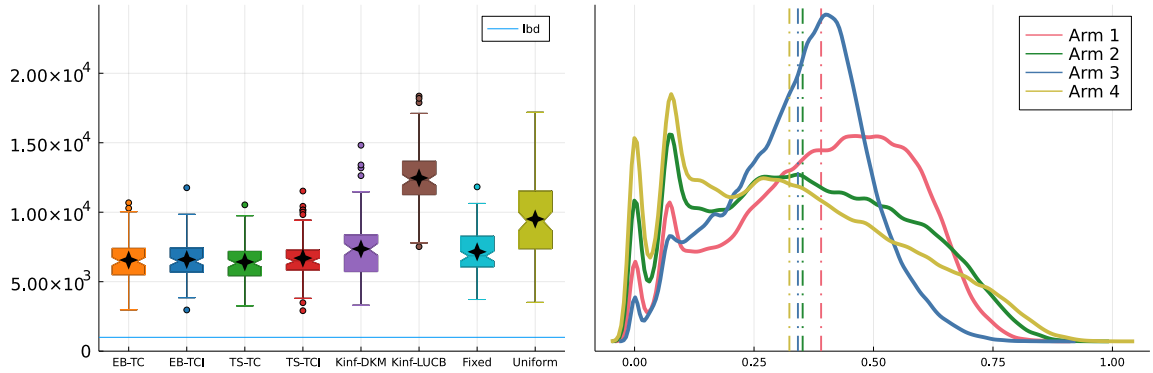

Figure 4.1 – Empirical stopping time (a) on scaled DSSAT instances with their density and mean (b). Lower bound is $T^\star(\nu)\log(1/\delta)$. "stars" equal means.

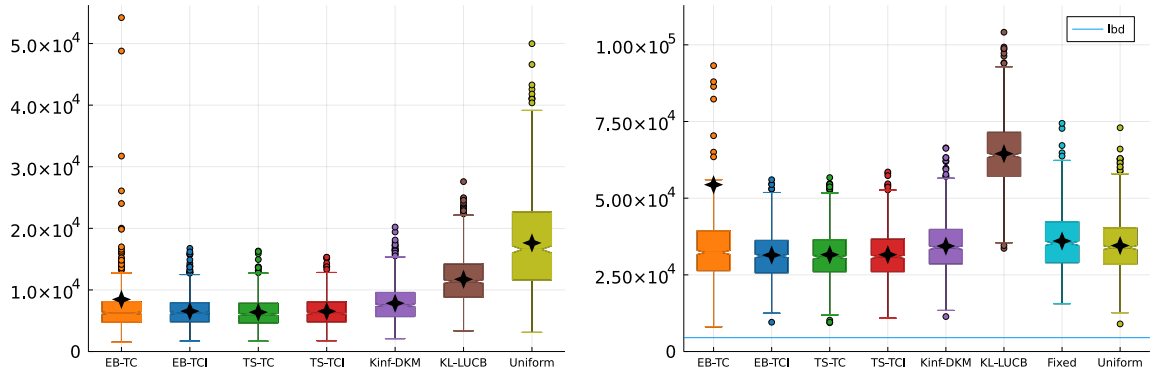

Figure 4.2 – Empirical stopping time on Bernoulli (a) random instances with $K = 10$ and (b) instance $\mu = (0.5, 0.45, 0.45)$.

to the small number of runs, we don't observe large outliers for β-EB-TC. KL-LUCB performs ten times worse than $\mathcal{K}_{\inf}$-LUCB, hence we omit it from Figure 4.1.

Bernoulli instances Next we assess the performance on 1000 random Bernoulli instances with $K = 10$ such that $\mu_1 = 0.6$ and $\mu_i \sim \mathcal{U}([0.2, 0.5])$ for all $i \neq 1$, where we enforce that $\Delta_{\min} \geq 0.01$. We also study the instance $\mu = (0.5, 0.45, 0.45)$, in which $\Delta_{\min} = 0$, and perform 1000 runs.

In Figure 4.2(a), we see that EB-TCI, TS-TC and TS-TCI outperform other algorithms. While this gain is slim compared to $\mathcal{K}_{\inf}$-DKM, the empirical stopping time is twice (resp. three times) as large for KL-LUCB (resp. uniform sampling). Even when $\Delta_{\min} = 0$, Figure 4.2(b) hints that their empirical performance might be preserved. Figure 4.2 confirms the lack of robustness of EB-TC, which is prone to large outliers. For the symmetric instance in Figure 4.2(b), uniform sampling outperforms KL-LUCB and perform on par with the "fixed" sampling.

4.5 Discussion

In Chapter 4, we instantiated several Top Two algorithms introduced in Chapter 2 for the class of bounded distributions, and proved their asymptotic β-optimality. On experiments on distributions coming from a real world application, several Top Two variants (in particular TS-TC-β and EB-TCI-β) proved to be more effective than all baselines. Furthermore, EB-TCI-β is computationally not costlier than computing the stopping rule.

This chapter goes beyond the class of parametric distributions used in Chapters 2 and 3, yet it considered a specific non-parametric class of distributions. An interesting direction of research lies in extending the analysis of Top Two algorithms to other class of non-parametric distributions by extracting sufficient (and minimal) conditions on the class $\mathcal{D}$. As we discussed in this chapter, those conditions should yield enough regularity on the $\mathcal{K}_{\inf}$ functions and on the characteristic times.

Finally, Top Two algorithms are promising algorithms to tackle the setting of fixed-budget identification, in which the algorithms have to stop at a given time and should then make as few mistakes as possible. As their sampling rule is anytime (*i.e.* independent of δ), Top Two algorithms might also have theoretical guarantees for BAI in the fixed-budget setting or even the anytime one, in which guarantees on the error probability should be given at all time. In Chapter 5, we will prove such guarantees in vanilla ε-BAI for Gaussian distributions with known variances.

Part II

Relaxed Identification in the Anytime Setting

Chapter 5

Epsilon Best Arm Identification

In Chapter 5, we study the ε-BAI problem for vanilla bandits in the anytime setting, as described in Chapter 1. The presented results were published in Jourdan et al. [2023b].

We propose EB-TC$_\varepsilon$, a novel sampling rule for ε-best arm identification in stochastic bandits. It is the first instance of Top Two algorithm analyzed for approximate best arm identification. EB-TC$_\varepsilon$ is an *anytime* sampling rule that can therefore be employed without modification for fixed confidence or fixed budget identification (without prior knowledge of the budget). We provide three types of theoretical guarantees for EB-TC$_\varepsilon$. First, we prove bounds on its expected sample complexity in the fixed confidence setting, notably showing its asymptotic optimality in combination with an adaptive tuning of its exploration parameter. We complement these findings with upper bounds on its probability of error at any time and for any error parameter, which further yield upper bounds on its simple regret at any time. Finally, we show through numerical simulations that EB-TC$_\varepsilon$ performs favorably compared to existing algorithms, in different settings.

Contents

5.1 Introduction

As detailed in Section 1.5, the motivation for the anytime setting comes from practical considerations. Practitioners might have different pre-defined constraints, *e.g.* the maximal budget might be fixed in advance or the error made should be smaller than a fixed admissible error. However, in many cases, fixing such constraints in advance can be challenging since a "good" choice typically depends on unknown quantities. Moreover, while the budget is limited in clinical trials, it is often not fixed beforehand. The physicians can decide to stop earlier or might obtain additional funding for their experiments. In light of those real-world constraints, regardless of its primal objective any strategy for choosing the next treatment should ideally come with guarantees on its current candidate answer that hold at any time.

As described in Chapter 1, the motivation of the ε-BAI problem stems from the sampling cost of the BAI problem. If several arms have means very close to the maximum, finding the one with the highest mean might be difficult. In practice we are often satisfied by any good enough arm, in the sense that its mean is greater than $\mu_\star - \varepsilon$, where $\mu_\star = \max_{i\in[K]} \mu_i$ and $\varepsilon \geq 0$. This is the ε-BAI task, in which $\mathcal{S} = \mathbb{R}^K$ when $\varepsilon > 0$. Let $\mathcal{I}_\varepsilon(\mu) = \{i \in [K] \mid \mu_i \geq \mu_\star - \varepsilon\}$ be the set of *ε-good* (or *ε-close*) arms, which are the multiple correct answers. Our results can also be adapted to multiplicative ε-BAI for $\varepsilon \in [0, 1)$, in which all means are non-negative, *i.e.* $\mathcal{S} = \{\mu \in \mathbb{R}^K \mid \min_{i\in[K]} \mu_i \geq 0\}$, and we want to find an arm with mean $\mu_i \geq (1-\varepsilon)\mu_\star$.

We consider the set $\mathcal{D}_\sigma$ of σ-sub-Gaussian distributions and assume that $\sigma_i = 1$ for all $i \in [K]$ by scaling, hence $\mathcal{D}^K = \mathcal{D}_1^K$. Let $\nu \in \mathcal{D}^K$ with mean vector $\mu \in \mathbb{R}^K$. Let $i^\star(\mu) := \arg\max_{i\in[K]} \mu_i$ be the set of arms with largest mean (*i.e.* $i^\star(\mu) = \mathcal{I}_0(\mu)$). Let $\Delta_i := \mu_\star - \mu_i$ denote the sub-optimality gap of arm i .

Performance criteria An anytime algorithm is defined of by a sampling rule and a recommendation rule. At time n , we denote by $\hat{\imath}_n$ the candidate answer and by I_n the arm to pull. Let $\mathbb{P}_\nu(\hat{\imath}_n \notin \mathcal{I}_\varepsilon(\mu))$ be the *probability of ε-error* of the recommendation at n , $\mathbb{E}_\nu[\mu_\star - \mu_{\hat{\imath}_n}]$ be its the expected *simple regret* which is independent of any parameter ε . As detailed in Section 1.5.2, there are several ways to evaluate the performance of an anytime algorithm for ε-BAI.

- Fixed confidence: given a known parameter $\delta \in (0, 1)$, we augment the algorithm with a *stopping rule* (hence a stopping time $\tau_{\varepsilon,\delta}$) which ensures that the algorithm is (ε, δ)-PAC, *i.e.* $\mathbb{P}_\nu(\tau_{\varepsilon,\delta} < +\infty, \hat{\imath}_{\tau_{\varepsilon,\delta}} \notin \mathcal{I}_\varepsilon(\mu)) \leq \delta$ for all $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathcal{S}$. The goal is to minimize the expected sample complexity $\mathbb{E}[\tau_{\varepsilon,\delta}]$, defined as the expected number of samples it needs to collect before it can stop and return a good arm with confidence $1 - \delta$.

- Fixed budget: given a known budget $T \in \mathbb{N}$, the algorithm is run until a predefined time T and we evaluate it based on the probability of error at T . This setting has been mostly

studied for $\varepsilon = 0$ [Audibert et al., 2010, Karnin et al., 2013], but Zhao et al. [2023] present the first bounds for $\varepsilon > 0$ for an algorithm that is actually agnostic to this value.

- Simple regret minimization: we evaluate the expected simple regret at a known budget $T \in \mathbb{N}$ [Bubeck et al., 2011, Zhao et al., 2023].

The anytime algorithm that we propose in this chapter is motivated by the fixed-confidence ε-BAI problem. We shall analyze its sample complexity in the fixed confidence setting but thanks to the anytime property we will also be able to prove guarantees on its probability of ε-error for every $\varepsilon \geq 0$ and its simple regret at any time.

Fixed-confidence ε-BAI Let $\varepsilon \geq 0$ and $\delta \in (0,1)$ be fixed error and confidence parameters. The (ε, δ)-PAC requirement leads to an asymptotic lower bound on the expected sample complexity.

Lemma 5.1 (Theorem 1 in Degenne and Koolen [2019]). *Let $\delta \in (0,1)$ and $\varepsilon \geq 0$. For all (ε, δ)-PAC algorithms and all instances $\nu = \mathcal{N}(\mu, 1_K)$ with $\mu \in \mathbb{R}^K$, $\liminf_{\delta \to 0} \mathbb{E}_\nu[\tau_{\varepsilon,\delta}] / \log(1/\delta) \geq T_\varepsilon(\nu)$ where $T_\varepsilon(\nu) = \min_{i \in \mathcal{I}_\varepsilon(\mu)} \min_{\beta \in (0,1)} T_{\varepsilon,\beta}(\nu, i)$ with*

$$T_{\varepsilon,\beta}(\nu, i)^{-1} = \max_{w \in \Sigma_K, w_i = \beta} \min_{j \neq i} \frac{1}{2} \frac{(\mu_i - \mu_j + \varepsilon)^2}{1/\beta + 1/w_j} . \tag{5.1}$$

An algorithm is asymptotically optimal (resp. β-optimal) if its sample complexity matches that lower bound, that is if $\limsup_{\delta \to 0} \mathbb{E}_\nu[\tau_{\varepsilon,\delta}] / \log(1/\delta) \leq T_\varepsilon(\nu)$ for all $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathbb{R}^K$ (resp. $T_{\varepsilon,\beta}(\nu) = \min_{i \in \mathcal{I}_\varepsilon(\mu)} T_{\varepsilon,\beta}(\nu, i)$). As in BAI, we can show that $T_{\varepsilon,1/2}(\nu) \leq 2 T_\varepsilon(\nu)$ [Russo, 2016]. The asymptotic characteristic time $T_\varepsilon(\nu)$ is of order $\sum_{i=1}^K \min\{\varepsilon^{-2}, \Delta_i^{-2}\}$. It is computed as a minimum over all ε-good arms $i \in \mathcal{I}_\varepsilon(\mu)$ of an arm-specific characteristic time, which can be interpreted as the time required to verify that i is a correct answer (*i.e.* ε-good). Each of the times $\min_{\beta \in (0,1)} T_{\varepsilon,\beta}(\nu, i)$ correspond to the complexity of a BAI instance (*i.e.* ε-BAI with $\varepsilon = 0$) in which the mean of arm i is increased by ε . Let $w_{\varepsilon,\beta}(\nu, i)$ be the maximizer of (5.1). In Garivier and Kaufmann [2021], they show that $T_\varepsilon(\nu) = T_{\varepsilon,\beta^\star(i^\star)}(\nu, i^\star)$ and $T_{\varepsilon,\beta}(\nu) = T_{\varepsilon,\beta}(\nu, i^\star)$, where $i^\star \in i^\star(\mu)$ and $\beta^\star(i^\star) = \arg\min_{\beta \in (0,1)} T_{\varepsilon,\beta}(\nu, i^\star)$. Note that the characteristic time for σ-sub-Gaussian distributions (which does not have a form as "explicit" as (5.1)) is always smaller than the ones for Gaussian having the same means and variance σ^2 . Lemma E.1 in Appendix E.1 gives a reduction of a ε-BAI problem to a BAI one on a modified instance.

Any time and uniform ε-error bound In addition to the fixed-confidence guarantees, we will prove a bound on the probability of error for any deterministic time n and any error ε , similarly

to the results of Zhao et al. [2023]. That is, we bound $\mathbb{P}_\nu(\hat{\imath}_n \notin \mathcal{I}_\varepsilon(\mu))$ for any deterministic n and any ε . This gives a bound on the probability of error in ε-BAI, and a bound on the simple regret of the sampling rule by integrating: $\mathbb{E}_\nu[\mu_\star - \mu_{\hat{\imath}_n}] = \int \mathbb{P}_\nu(\hat{\imath}_n \notin \mathcal{I}_\varepsilon(\mu))\mathrm{d}\,\varepsilon$.

Contribution 5.2. *In Chapter 5, we propose the EB-TC$_{\varepsilon_0}$ algorithm, with a slack parameter $\varepsilon_0 > 0$, originally motivated by ε_0-BAI. It is the first Top Two algorithm for the ε-BAI problem when $\varepsilon > 0$. We study its combination with a stopping rule for fixed confidence ε-BAI (possibly with $\varepsilon_0 \neq \varepsilon$) and also its probability of error and simple regret at any time.*

- *EB-TC$_{\varepsilon_0}$ performs well empirically compared to existing methods, both for the expected sample complexity criterion in fixed confidence ε-BAI and for the anytime simple regret criterion. It is in addition easy to implement and computationally inexpensive in our regime.*
- *We prove upper bounds on the sample complexity of EB-TC$_{\varepsilon_0}$ in fixed confidence ε-BAI with 1-sub-Gaussian distributions, both asymptotically (Theorem 5.5) as $\delta \to 0$ and for any δ (Theorem 5.6). In particular, EB-TC$_\varepsilon$ with $\varepsilon > 0$ is asymptotically optimal for ε-BAI with Gaussian distributions.*
- *We prove a uniform ε-error bound valid for any time for EB-TC$_{\varepsilon_0}$. This gives in particular a fixed budget error bound and a control of the expected simple regret of the algorithm at any deterministic time n (Theorem 5.10 and Corollary 5.11).*

5.2 Anytime Top Two Sampling Rule

We propose an anytime Top Two algorithm, named EB-TC$_{\varepsilon_0}$ and summarized in Figure 5.1. This is an instance of Top Two algorithm, see Chapter 2.2 for more details on those methods.

We start by sampling each arm once. Let $N_{n,i}$ and $\mu_{n,i}$ be the empirical count and empirical mean of arm $i \in [K]$ before time n . At time $n > K$, we recommend the Empirical Best (EB) arm $\hat{\imath}_n \in i^\star(\mu_n)$ (where ties are broken arbitrarily). At time $n > K$, a Top Two sampling rule defines a leader $B_n \in [K]$ and a challenger $C_n \neq B_n$. It then chooses the arm to pull among them. For the leader/challenger pair, we consider the Empirical Best (EB) leader $B_n^{\mathrm{EB}} = \hat{\imath}_n$ and, given a slack $\varepsilon_0 > 0$, the Transportation Cost (TC$_{\varepsilon_0}$) challenger

$$C_n^{\mathrm{TC}\varepsilon_0} \in \operatorname*{arg\,min}_{i \neq B_n^{\mathrm{EB}}} \frac{\mu_{n,B_n^{\mathrm{EB}}} - \mu_{n,i} + \varepsilon_0}{\sqrt{1/N_{n,B_n^{\mathrm{EB}}} + 1/N_{n,i}}} \,. \tag{5.2}$$

The TC$_{\varepsilon_0}$ challenger seeks to minimize an empirical version of a quantity that appears in the lower bound for ε_0-BAI (Lemma 5.1), hence it is a natural extension of the TC challenger used in the T3C algorithm [Shang et al., 2020] for $\varepsilon_0 = 0$.

Input: Slack $\varepsilon > 0$, proportion $\beta \in (0,1)$ (only for fixed proportions).
Output: Next arm to sample I_n and next recommendation $\hat{\imath}_n$.
Set $\hat{\imath}_n \in \arg\max_{i \in [K]} \mu_{n,i}$; `// Candidate answer`
Set $B_n = \hat{\imath}_n$ and $C_n \in \arg\min_{i \neq B_n} \frac{\mu_{n,B_n} - \mu_{n,i} + \varepsilon_0}{\sqrt{1/N_{n,B_n} + 1/N_{n,i}}}$; `// Leader and challenger`
Set [**fixed**] $\beta_n(B_n, C_n) = \beta$ or [**IDS**] $\beta_n(B_n, C_n) = N_{n,C_n}/(N_{n,B_n} + N_{n,C_n})$, then update $\bar{\beta}_{n+1}(B_n, C_n)$ and $T_{n+1}(B_n, C_n)$; `// Target allocation`
Set $I_n = \begin{cases} C_n & \text{if } N^{B_n}_{n,C_n} \leq (1 - \bar{\beta}_{n+1}(B_n, C_n)) T_{n+1}(B_n, C_n) \\ B_n & \text{otherwise} \end{cases}$; `// Tracking`

Algorithm 5.1: EB-TC$_{\varepsilon_0}$ algorithm with **fixed** or **IDS** proportions.

We select $I_n \in \{B_n, C_n\}$ thanks to the tracking procedure associated to the leader/challenger pair (B_n, C_n) , hence we have $K(K-1)$ independent tracking procedures. The tracking procedure ensures that the proportion of times the algorithm pulled the leader i remains close to a target average proportion $\bar{\beta}_n(i,j) \in (0,1)$. We define two variants of the algorithm that differ in the way they set the proportions $\bar{\beta}_n(i,j)$. *Fixed* proportions set $\bar{\beta}_n(i,j) = \beta$ for all $(n,i,j) \in \mathbb{N} \times [K]^2$, where $\beta \in (0,1)$ is fixed beforehand. Information-Directed Selection (*IDS*) [You et al., 2023] defines $\beta_n(i,j) = N_{n,j}/(N_{n,i} + N_{n,j})$ and sets $\bar{\beta}_n(i,j) := T_n(i,j)^{-1} \sum_{t \in [n-1]} \mathbb{1}\left((B_t, C_t) = (i,j)\right) \beta_t(i,j)$ where $T_n(i,j) := \sum_{t \in [n-1]} \mathbb{1}\left((B_t, C_t) = (i,j)\right)$ is the selection count of arms (i,j) as leader/challenger. Importantly, the IDS proportions are independent of the slack ε for Gaussian with known variance. Empirically, we observe slightly better performances when using IDS.

Let $N^i_{n,j} := \sum_{t \in [n-1]} \mathbb{1}\left((B_t, C_t) = (i,j),\ I_t = j\right)$ be the number of pulls of arm j at rounds in which i was the leader. We set $I_n = C_n$ if $N^{B_n}_{n,C_n} \leq (1 - \bar{\beta}_{n+1}(B_n, C_n)) T_{n+1}(B_n, C_n)$ and $I_n = B_n$ otherwise. Using Theorem 6 in Degenne et al. [2020b] for each tracking procedure (*i.e.* each pair (i,j)) yields Lemma 5.3.

Lemma 5.3. *For all $n > K$, $i \in [K]$, $j \neq i$, we have $-1/2 \leq N^i_{n,j} - (1 - \bar{\beta}_n(i,j)) T_n(i,j) \leq 1$.*

Choosing ε_0 Jourdan et al. [2022] shows that EB-TC (*i.e.* EB-TC$_{\varepsilon_0}$ with slack $\varepsilon_0 = 0$) suffers from poor empirical performance for moderate δ in BAI (see Appendix D.3 in Jourdan et al. [2022] for a detailed discussion). Therefore, the choice of the slack $\varepsilon_0 > 0$ is critical since it acts as a regularizer which naturally induces sufficient exploration. By setting ε_0 too small, the EB-TC$_{\varepsilon_0}$ algorithm will become as greedy as EB-TC and perform poorly. Having ε_0 too large will flatten differences between sub-optimal arms, hence it will behave more uniformly. We observe from the theoretical guarantees and from our experiments that it is best to take $\varepsilon_0 = \varepsilon$ for ε-BAI, but the empirical performance is only degrading slowly for $\varepsilon_0 > \varepsilon$. Taking $\varepsilon_0 < \varepsilon$

leads to very poor performance. When tackling BAI, the limitation of EB-TC$_0$ can be solved by adding an implicit exploration mechanism in the choice of the leader/challenger pair.

Anytime sampling rule EB-TC$_{\varepsilon_0}$ is independent of a budget of samples T or a confidence parameter δ. This anytime sampling rule can be regarded as a stream of empirical means/counts $(\mu_n, N_n)_{n>K}$ (which could trigger stopping) and a stream of recommendations $\hat{\imath}_n = i^\star(\mu_n)$. These streams can be used by agents with different kinds of objectives. The fixed-confidence setting couples it with a stopping rule to be (ε, δ)-PAC. It can also be used to get an ε-good recommendation with large probability at any given time n.

5.2.1 Stopping Rule for Fixed-confidence ε-Best-arm Identification

In addition to the sampling and recommendation rules, the fixed-confidence setting requires a stopping rule. Given an error/confidence pair, the GLR$_\varepsilon$ stopping rule [Garivier and Kaufmann, 2016] prescribes to stop at the time

$$\tau_{\varepsilon,\delta} = \inf \left\{ n > K \mid \min_{i \neq \hat{\imath}_n} \frac{\mu_{n,\hat{\imath}_n} - \mu_{n,i} + \varepsilon}{\sqrt{1/N_{n,\hat{\imath}_n} + 1/N_{n,i}}} > \sqrt{2c(n-1,\delta)} \right\} \quad \text{with} \quad \hat{\imath}_n \in i^\star(\mu_n)\,, \tag{5.3}$$

where $c : \mathbb{N} \times (0,1) \to \mathbb{R}_+$ is a threshold function. Lemma 5.4 gives a threshold ensuring that the GLR$_\varepsilon$ stopping rule is (ε, δ)-PAC for all $\varepsilon \geq 0$ and $\delta \in (0,1)$, regardless of the sampling rule.

Lemma 5.4 ([Kaufmann and Koolen, 2021])**.** *Let $\varepsilon \geq 0$ and $\delta \in (0,1)$. Given any sampling rule, using the threshold*

$$c(n,\delta) = 2\mathcal{C}_G(\log((K-1)/\delta)/2) + 4\log(4 + \log(n/2)) \tag{5.4}$$

with the stopping rule (5.3) with error/confidence pair (ε, δ) yields a (ε, δ)-PAC algorithm for 1-sub-Gaussian distributions with mean in $\mathbb{R}^K$. The function $\mathcal{C}_G$ is defined in (B.1). It satisfies $\mathcal{C}_G(x) \approx x + \log(x)$.

5.3 Fixed-confidence Guarantees

To study ε-BAI in the fixed-confidence setting, we couple EB-TC$_{\varepsilon_0}$ with the GLR$_\varepsilon$ stopping rule (5.3) using error $\varepsilon \geq 0$, confidence $\delta \in (0,1)$ and threshold (5.4), hence this algorithm is (ε, δ)-PAC. We derive upper bounds on the expected sample complexity $\mathbb{E}_\nu[\tau_{\varepsilon,\delta}]$ both in the asymptotic regime of $\delta \to 0$ (Theorem 5.5) and for finite confidence when $\varepsilon = \varepsilon_0$ (Theorem 5.6).

Theorem 5.5. *Let $\varepsilon \geq 0$, $\varepsilon_0 > 0$ and $(\beta, \delta) \in (0,1)^2$. Combined with GLR$_\varepsilon$ stopping (5.3), the EB-TC$_{\varepsilon_0}$ algorithm is (ε, δ)-PAC and it satisfies that, for all $\nu \in \mathcal{D}^K$ with mean μ such that $|i^\star(\mu)| = 1$,*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau_{\varepsilon,\delta}]}{\log(1/\delta)} \leq D_{\varepsilon,\varepsilon_0}(\mu) \begin{cases} T_{\varepsilon_0}(\nu) & \textbf{[IDS]} \\ T_{\varepsilon_0,\beta}(\nu) & \textbf{[fixed } \beta\textbf{]} \end{cases},$$

where $D_{\varepsilon,\varepsilon_0}(\mu) = (1 + \max_{i \neq i^\star}(\varepsilon_0 - \varepsilon)/(\mu_\star - \mu_i + \varepsilon))^2$.

Proof. The proof follows along the same lines as the one in Section 2.3 of Chapter 2, which was actually written in sufficient generality to also cope for ε-BAI. We have shown that the EB leader satisfies the Properties 2.15 and 2.18 in Appendix B.12. As for the TC challenger studied in Appendix B.13, the TC$_{\varepsilon_0}$ challenger satisfies the Properties 2.16 and 2.21. Interestingly, the proof also holds true when $\varepsilon \neq \varepsilon_0$ at an extra cost of $D_{\varepsilon,\varepsilon_0}(\mu)$ since we approximate empirically (β-)optimal allocation for ε_0-BAI while stopping for ε-BAI. We refer the reader to Appendix D in Jourdan et al. [2023b] for more details. We also note that Theorem 5.5 is not conflicting with the lower bound of Lemma 5.1, as shown by Lemma E.2 in Appendix E.1. ■

While Theorem 5.5 holds for all 1-sub-Gaussian distributions, it is particularly interesting for Gaussian ones, in light of Lemma 5.1. When choosing $\varepsilon = \varepsilon_0$ (*i.e.* $D_{\varepsilon_0,\varepsilon_0}(\mu) = 1$), Theorem 5.5 shows that EB-TC$_{\varepsilon_0}$ is asymptotically optimal for Gaussian bandits when using IDS proportions and asymptotically β-optimal when using fixed proportions β . By direct computation, one can show that Theorem 5.5 is not conflicting with the lower bound of Lemma 5.1. Empirically we observe that the empirical stopping time can be drastically worse when taking $\varepsilon_0 < \varepsilon$, and close to the optimal one when $\varepsilon_0 > \varepsilon$.

Until recently [You et al., 2023], proving asymptotic optimality of Top Two algorithms with adaptive choice β was an open problem in BAI. In this work, we prove that their choice of IDS proportions also yield asymptotically optimal algorithms for ε-BAI. While the proof of Theorem 5.5 assumes the existence of a unique best arm, it holds for instances having sub-optimal arms with the same mean. This is an improvement compared to existing asymptotic guarantees on Top Two algorithms which rely on the assumption that the means of all arms are different [Qin et al., 2017, Shang et al., 2020, Jourdan et al., 2022]. The improvement is possible thanks to the regularization induced by the slack $\varepsilon_0 > 0$.

While asymptotic optimality in the ε-BAI setting was already achieved for various algorithms (*e.g.* ε-Track-and-Stop (TaS) [Garivier and Kaufmann, 2021], Sticky TaS [Degenne and Koolen, 2019] or LεBAI [Jourdan and Degenne, 2022]), none of them obtained non-asymptotic guarantees. Despite their theoretical interest, asymptotic results provide no guarantee on the

performance for moderate δ . Furthermore, asymptotic results on Top Two algorithms require a unique best arm regardless of the considered error ε: reaching asymptotic (β-)optimality implicitly means that the algorithm eventually allocates samples in an optimal way that depends on the identity of the unique best arm, and that requires the unique best arm to be identified. As our focus is ε-BAI, our guarantees should only require that one of the ε-good arms is identified and should hold for instances having multiple best arms. The upper bound should scale with ε_0^{-2} instead of $\Delta_{\min}^{-2}$ when $\Delta_{\min}$ is small. Theorem 5.6 satisfies these requirements.

Theorem 5.6. *Let $\delta \in (0,1)$ and $\varepsilon_0 > 0$. Combined with $\mathrm{GLR}_{\varepsilon_0}$ stopping (5.3), the EB-TC$_{\varepsilon_0}$-1/2 algorithm is (ε_0, δ)-PAC and satisfies that, for all $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathbb{R}^K$,*

$$\mathbb{E}_\nu[\tau_{\varepsilon_0,\delta}] \leq \inf_{\tilde{\varepsilon}\in[0,\varepsilon_0]} \max\left\{T_{\mu,\varepsilon_0}(\delta,\tilde{\varepsilon}) + 1,\ S_{\mu,\varepsilon_0}(\tilde{\varepsilon})\right\} + 2K^2 \,,$$

where $T_{\mu,\varepsilon_0}(\delta,\tilde{\varepsilon})$ and $S_{\mu,\varepsilon_0}(\tilde{\varepsilon})$ are defined by

$$T_{\mu,\varepsilon_0}(\delta,\tilde{\varepsilon}) = \sup\{n \mid n-1 \leq 2(1+\gamma)^2 \sum_{i\in\mathcal{I}_{\tilde{\varepsilon}}(\mu)} T_{\varepsilon_0,1/2}(\nu,i)(\sqrt{c(n-1,\delta)} + \sqrt{4\log n})^2\} \,,$$

$$S_{\mu,\varepsilon_0}(\tilde{\varepsilon}) = h_1\left(\frac{16(1+\gamma^{-1})}{a_{\mu,\varepsilon_0}(\tilde{\varepsilon})} H_{\mu,\varepsilon_0}(\tilde{\varepsilon}),\ \frac{(1+\gamma^{-1})K^2}{a_{\mu,\varepsilon_0}(\tilde{\varepsilon})} + 1\right) \,,$$

$$a_{\mu,\varepsilon_0}(\tilde{\varepsilon}) = \frac{\min_{i\in\mathcal{I}_{\tilde{\varepsilon}}(\mu)} T_{\varepsilon_0,1/2}(\nu,i)}{\sum_{i\in\mathcal{I}_{\tilde{\varepsilon}}(\mu)} T_{\varepsilon_0,1/2}(\nu,i)} \min_{i\in\mathcal{I}_{\tilde{\varepsilon}}(\mu), j\neq i} w_{\varepsilon_0,1/2}(\nu,i)_j \,,$$

where $\gamma \in (0, 1/2]$ is an analysis parameter and $h_1(y,z) \approx z + y\log(z + y\log(y))$. $T_{\varepsilon_0,1/2}(\mu,i)$ and $w_{\varepsilon_0,1/2}(\nu,i)$ are defined in (5.1) and

$$H_{\mu,\varepsilon_0}(\tilde{\varepsilon}) := \frac{2|i^\star(\mu)|}{\Delta_\mu(\tilde{\varepsilon})^2} + (|\mathcal{I}_{\tilde{\varepsilon}}(\mu) \setminus i^\star(\mu)|) C_{\mu,\varepsilon_0}(\tilde{\varepsilon})^2 + \sum_{i\notin\mathcal{I}_{\tilde{\varepsilon}}(\mu)} \max\{C_{\mu,\varepsilon_0}(\tilde{\varepsilon}), \sqrt{2}\Delta_i^{-1}\}^2 \,, \quad (5.5)$$

with $\Delta_\mu(\tilde{\varepsilon}) = \min_{k\notin\mathcal{I}_{\tilde{\varepsilon}}(\mu)} \Delta_k$ and $C_{\mu,\varepsilon_0}(\tilde{\varepsilon}) = \max\{2\Delta_\mu(\tilde{\varepsilon})^{-1} - \varepsilon_0^{-1}, \varepsilon_0^{-1}\}$.

Proof. The proof follows along the same lines as the one in Section 2.4 of Chapter 2, which was actually written in sufficient generality to also cope for ε-BAI. While we detail some key lemmas below, we refer the reader to Appendix E in Jourdan et al. [2023b] for more details. ■

The upper bound on $\mathbb{E}_\nu[\tau_{\varepsilon_0,\delta}]$ involves a δ-dependent term $T_{\mu,\varepsilon_0}(\delta,\tilde{\varepsilon})$ and a δ-independent term $S_{\mu,\varepsilon_0}(\tilde{\varepsilon})$. The choice of $\tilde{\varepsilon}$ influences the compromise between those, and the infimum over $\tilde{\varepsilon}$ means that our algorithm benefits from the best possible trade-off. In the asymptotic regime, we take $\tilde{\varepsilon} = 0$ and $\gamma \to 0$ and we obtain $\lim_{\delta\to 0} \mathbb{E}_\nu[\tau_{\varepsilon_0,\delta}]/\log(1/\delta) \leq 2|i^\star(\mu)| T_{\varepsilon_0,1/2}(\nu)$. When $|i^\star(\mu)| = 1$, we recover the asymptotic result of Theorem 5.5 up to a multiplicative factor 2 .

For multiple best arms, the asymptotic sample complexity is at most a factor $2|i^\star(\mu)|$ from the β-optimal one.

Given a finite confidence, the dominant term will be $S_{\mu,\varepsilon_0}(\tilde{\varepsilon})$. For $\tilde{\varepsilon} = 0$, we show that $H_{\mu,\varepsilon_0}(0) = \mathcal{O}(K \min\{\Delta_{\min}, \varepsilon_0\}^{-2})$, hence we should consider $\tilde{\varepsilon} > 0$ to avoid the dependency in $\Delta_{\min}$. For $\tilde{\varepsilon} = \varepsilon_0$, there exists instances such that $\max_{i \in \mathcal{I}_{\varepsilon_0}(\mu)} T_{\varepsilon_0,1/2}(\nu, i)$ is arbitrarily large, hence $S_{\mu,\varepsilon_0}(\varepsilon_0)$ will be very large as well. The best trade-off is attained in the interior of the interval $(0, \varepsilon_0)$. Proven in Appendix E.2, Lemma 5.7 yields that, for $\tilde{\varepsilon} = \varepsilon_0/2$, we have $T_{\varepsilon_0,1/2}(\nu, i) = \mathcal{O}(K/\varepsilon_0^2)$ for all $i \in \mathcal{I}_{\varepsilon_0/2}(\mu)$ and $H_{\mu,\varepsilon_0}(\varepsilon_0/2) = \mathcal{O}(K/\varepsilon_0^2)$. Therefore, we obtain an upper bound $\mathcal{O}(|\mathcal{I}_{\varepsilon_0/2}(\mu)| K \varepsilon_0^{-2} \log \varepsilon_0^{-1})$.

Lemma 5.7. *Let $\varepsilon > 0$ and $\mu \in \mathbb{R}^K$. Then, for all $i \in \mathcal{I}_{\varepsilon/2}(\mu)$, we have $T_{\varepsilon,1/2}(\nu, i) \leq 32K/\varepsilon^2$ and $\min_{j \neq i} w_{\varepsilon,1/2}(\nu, i)_j \geq (16(K-2)+2)^{-1}$.*

While the dependency in $a_{\mu,\varepsilon_0}(\varepsilon_0/2)$ is milder in ε-BAI than in BAI (as it is bounded away from 0), we can improve it by using a refined analysis. Introduced in Jourdan and Degenne [2023], this method allows to clip $\min_{j \neq i} w_{\varepsilon_0,1/2}(\nu, i)_j$ by a fixed value $x \in [0, (K-1)^{-1}]$ for all $i \in \mathcal{I}_{\tilde{\varepsilon}}(\mu)$.

Comparison with existing upper bounds The LUCB algorithm [Kalyanakrishnan et al., 2012] has a structure similar to a Top Two algorithm, with the differences that LUCB samples both the leader and the challenger and that it stops when the gap between the UCB and LCB indices is smaller than ε_0 . As LUCB satisfies $\mathbb{E}_\nu[\tau_{\varepsilon_0,\delta}] \leq 292 H_{\varepsilon_0}(\mu) \log(H_{\varepsilon_0}(\mu)/\delta) + 16$ where $H_{\varepsilon_0}(\mu) = \sum_i (\max\{\Delta_i, \varepsilon_0/2\})^{-2}$, it enjoys better scaling than EB-TC$_{\varepsilon_0}$-1/2 for finite confidence. However, since the empirical allocation of LUCB is not converging towards $w_{\varepsilon_0,1/2}(\mu)$, it is not asymptotically $1/2$-optimal. While LUCB has better moderate confidence guarantees, there is no hope to prove anytime performance bounds since LUCB indices depends on δ . In contrast, EB-TC$_{\varepsilon_0}$-1/2 enjoys such guarantees (see Section 5.4).

Key technical tool for the non-asymptotic analysis We want to ensure that EB-TC$_{\varepsilon_0}$-1/2 eventually selects only ε-good arms as leader, for any error $\varepsilon \geq 0$. Our proof strategy is to show that if the leader is not an ε-good arm and empirical means do not deviate too much from the true means, then either the current leader or the current challenger was selected as leader or challenger less than a given quantity. We obtain a bound on the total number of times that can happen. The proof of Lemma 5.8 is detailed in Appendix E.3.

Lemma 5.8. *Let $\delta \in (0,1]$ and $n > K$. Let $T_n(i) := \sum_{j \neq i}(T_n(i,j) + T_n(j,i))$ be the number of times arm i was selected in the leader/challenger pair. Assume there exists a sequence of events $(A_t(n,\delta))_{K<t\leq n}$ and positive reals $(D_i(n,\delta))_{i\in[K]}$ such that, for all $t \in \{K+1,\dots,n\}$, under the event $A_t(n,\delta)$,*

$$\exists i_t \in [K], \quad T_t(i_t) \leq D_{i_t}(n,\delta) \quad \text{and} \quad T_{t+1}(i_t) = T_t(i_t) + 1 \,. \tag{5.6}$$

Then, we have $\sum_{t=K+1}^{n} \mathbb{1}\left(A_t(n,\delta)\right) \leq \sum_{i\in[K]} D_i(n,\delta)$.

To control the deviation of the empirical means and empirical gaps to their true value, we use a sequence of concentration events $(\mathcal{E}_{n,\delta})_{n>T}$ defined as $\mathcal{E}_{n,\delta} = \mathcal{E}^1_{n,\delta} \cap \mathcal{E}^2_{n,\delta}$ with

$$\mathcal{E}^1_{n,\delta} := \left\{ \forall k \in [K], \forall t \leq n,\ |\mu_{t,k} - \mu_k| < \sqrt{\frac{2f_1(n,\delta)}{N_{t,k}}} \right\} ,$$
$$\mathcal{E}^2_{n,\delta} := \left\{ \forall (i,k) \in [K]^2 \text{ s.t. } i \neq k,\ \forall t \leq n,\ \frac{|(\mu_{t,i} - \mu_{t,k}) - (\mu_i - \mu_k)|}{\sqrt{1/N_{t,i} + 1/N_{t,k}}} < \sqrt{2f_2(n,\delta)} \right\} ,$$

where $f_1(x,\delta) = \log(1/\delta) + (1+s)\log x$ and $f_2(x,\delta) = \log(1/\delta) + (2+s)\log(x)$. It satisfies that $\mathbb{P}_\nu(\mathcal{E}^{\complement}_{n,\delta}) \leq K^2\delta n^{-s}$ where $s \geq 0$ and $\delta \in (0,1]$. For the EB-TC$_{\varepsilon_0}$-1/2 algorithm with fixed $\beta = 1/2$, we prove that, under $\mathcal{E}_{n,\delta}$, $\{B^{\mathrm{EB}}_t \notin \mathcal{I}_\varepsilon(\mu)\}$ is a "bad" event satisfying the assumption of Lemma 5.8. This yields Lemma 5.9, which essentially says that the leader is an ε-good arm except for a logarithmic number of rounds. The proof of Lemma 5.9 is detailed in Appendix E.4.

Lemma 5.9. *Let $\delta \in (0,1]$, $n > K$ and $\varepsilon \geq 0$. Under the event $\mathcal{E}_{n,\delta}$, we have*

$$\sum_{i\in\mathcal{I}_\varepsilon(\mu)} \sum_j T_n(i,j) \geq n - 1 - 8H_{\mu,\varepsilon_0}(\varepsilon) f_2(n,\delta) - 3K^2 \,,$$

where $f_2(n,\delta) = \log(1/\delta) + (2+s)\log n$ and $H_{\mu,\varepsilon_0}(\varepsilon)$ is defined in (5.5).

Noticeably, Lemma 5.9 holds for any $\varepsilon \geq 0$ even when there are multiple best arms. As expected the number of times the leader is not among the ε_0-good arms depends on $H_{\mu,\varepsilon_0}(\varepsilon_0) = \mathcal{O}(K/\varepsilon_0^2)$. The number of times the leader is not among the best arms depends on $H_{\mu,\varepsilon_0}(0) = \mathcal{O}(K(\min\{\Delta_{\min},\varepsilon_0\})^{-2})$.

Time-varying slack Theorem 5.5 shows the asymptotic optimality of the EB-TC$_{\varepsilon_0}$-IDS algorithm for ε_0-BAI (where $\varepsilon_0 > 0$). To obtain optimality for BAI, we consider time-varying slacks $(\varepsilon_n)_n$, where $(\varepsilon_n)_n$ is decreasing, $\varepsilon_n > 0$ and $\varepsilon_n \to_{+\infty} 0$. Thanks to a direct adaptation of our

asymptotic analysis on $\mathbb{E}_\nu[\tau_{0,\delta}]$, regardless of the choice of $(\varepsilon_n)_n$, one can show that using GLR_0 stopping, the EB-TC$_{(\varepsilon_n)_n}$-IDS algorithm with IDS is $(0,\delta)$-PAC and is asymptotically optimal in BAI. Its empirical performance is however very sensitive to the choice of $(\varepsilon_n)_n$.

5.4 Anytime Guarantees on the Probability of Error

Could an algorithm analyzed in the fixed-confidence setting be used for the fixed-budget or even anytime setting? This question is especially natural for EB-TC$_{\varepsilon_0}$, which does not depend on the confidence parameter δ . Yet its answer is not obvious, as it is known that algorithms that have *optimal* asymptotic guarantees in the fixed-confidence setting can be sub-optimal in terms of error probability. Indeed Komiyama et al. [2022] prove in their Appendix C that for any asymptotically optimal (exact) BAI algorithm, there exists instances in which the error probability cannot decay exponentially with the horizon, which makes them worse than the (minimax optimal) uniform sampling strategy [Bubeck et al., 2011].

Their argument also applies to β-optimal algorithms, hence to EB-TC$_0$ with $\beta = 1/2$. Since their argument relies on the sparsity of the optimal allocation when considering the limit of $\Delta_{\min} \to 0$, it will not apply to ε-BAI as the optimal allocation are not asymptotically sparse (Lemma 5.7). However, whenever ε_0 is positive, Theorem 5.10 reveals that the error probability of EB-TC$_{\varepsilon_0}$-1/2 always decays exponentially, which redeems the use of optimal fixed-confidence algorithms for a relaxed BAI problem in the anytime setting. Going further, this result provides an anytime bound on the probability to recommend an arm that is not ε-optimal, for any error $\varepsilon \geq 0$. This bound involves instance-dependent complexities depending solely on the gaps in μ . To state it, we define $C_\mu := |\{\mu_i \mid i \in [K]\}|$ as the number of distinct arm means in μ and let $\mathcal{C}_\mu(i) := \{k \in [K] \mid \mu_\star - \mu_k = \Delta_i\}$ be the set of arms having mean gap Δ_i where the gaps are sorted by increasing order $0 = \Delta_1 < \Delta_2 < \cdots < \Delta_{C_\mu}$. For all $\varepsilon \geq 0$, let $i_\mu(\varepsilon) = i$ if $\varepsilon \in [\Delta_i, \Delta_{i+1})$ (with the convention $\Delta_{C_\mu+1} = +\infty$). Theorem 5.10 shows that the exponential decrease of $\mathbb{P}_\nu\left(\hat{\imath}_n \notin \mathcal{I}_\varepsilon(\mu)\right)$ is linear. The proof of Theorem 5.10 is detailed in Section 5.4.1 (see also Appendix F in Jourdan et al. [2023b]).

Theorem 5.10. *Let $\varepsilon_0 > 0$. The EB-TC$_{\varepsilon_0}$-1/2 algorithm satisfies that, for all $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathbb{R}^K$, for all $\varepsilon \geq 0$, for all $n > 5K^2/2$,*

$$\mathbb{P}_\nu\left(\hat{\imath}_n \notin \mathcal{I}_\varepsilon(\mu)\right) \leq K^2 e^2 (2+\log n)^2 \exp\left(-p\left(\frac{n - 5K^2/2}{8H_{i_\mu(\varepsilon)}(\mu,\varepsilon_0)}\right)\right) .$$

where $p(x) = x - \log x$ and $(H_i(\mu,\varepsilon_0))_{i\in[C_\mu-1]}$ are such that $H_1(\mu,\varepsilon_0) = K(2\Delta_{\min}^{-1} + 3\varepsilon_0^{-1})^2$ and $K/\Delta_{i+1}^{-2} \leq H_i(\mu,\varepsilon_0) \leq K \min_{j\in[i]} \max\{2\Delta_{j+1}^{-1},\ 2\frac{\Delta_j/\varepsilon_0+1}{\Delta_{i+1}-\Delta_j} + 3\varepsilon_0^{-1}\}^2$ for all $i > 1$.

This bound can be compared with the following uniform ε-error bound of the strategy using uniform sampling and recommending the empirical best arm:

$$\mathbb{P}_\nu\left(\hat{\imath}_n^{\mathrm{U}} \notin \mathcal{I}_\varepsilon(\mu)\right) \leq \sum_{i \notin \mathcal{I}_\varepsilon(\mu)} \exp\left(-\frac{\Delta_i^2 \lfloor n/K \rfloor}{4}\right) \leq K \exp\left(-\frac{n-K}{4K\Delta_{i_\mu(\varepsilon)+1}^{-2}}\right) .$$

Recalling that the quantity $H_i(\mu, \varepsilon_0)$ in Theorem 5.10 is always bounded from below by $2K\Delta_{i+1}^{-1}$, we get that our upper bound is larger than the probability of error of the uniform strategy, but the two should be quite close. For example for $\varepsilon = 0$, we have

$$\mathbb{P}_\nu\left(\hat{\imath}_n \notin i^\star(\mu)\right) \leq \exp\left(-\Theta\left(\frac{n}{K(\Delta_{\min}^{-1} + \varepsilon_0^{-1})^2}\right)\right) , \quad \mathbb{P}_\nu\left(\hat{\imath}_n^{\mathrm{U}} \notin i^\star(\mu)\right) \leq \exp\left(-\Theta\left(\frac{n}{K\Delta_{\min}^{-2}}\right)\right) .$$

Even if they provide a nice sanity-check for the use of a sampling rule with optimal fixed-confidence guarantees for ε_0-BAI in the anytime regime, we acknowledge that these guarantees are far from optimal. Indeed, the work of Zhao et al. [2023] provides tighter anytime uniform ε-error probability bounds for two algorithms: an anytime version of Sequential Halving [Karnin et al., 2013] using a doubling trick (called DSH), and an algorithm called Bracketting Sequential Halving, that is designed to tackle a very large number of arms. Their upper bounds are of the form $\mathbb{P}_\nu\left(\hat{\imath}_n \notin \mathcal{I}_\varepsilon(\mu)\right) \leq \exp\left(-\Theta\left(n/H(\varepsilon)\right)\right)$ with $H(\varepsilon) = \frac{1}{g(\varepsilon/2)} \max_{i \geq g(\varepsilon)+1} \frac{i}{\Delta_i^2}$ where $g(\varepsilon) = |\{i \in [K] \mid \mu_i \geq \mu^\star - \varepsilon\}|$. Therefore, they can be much smaller than $K\Delta_{i_\mu(\varepsilon)+1}^{-2}$.

The BUCB algorithm of Katz-Samuels and Jamieson [2020] is also analyzed for any level of error ε , but in a different fashion. The authors provide bounds on its (ε, δ)-*unverifiable sample complexity*, defined as the expectation of the smallest stopping time $\tilde{\tau}$ satisfying $\mathbb{P}(\forall t \geq \tilde{\tau}, \hat{\imath}_n \in \mathcal{I}_\varepsilon(\mu)) \geq 1-\delta$. This notion is different from the sample complexity we use in this paper, which is sometimes called *verifiable* since it is the time at which the algorithm can guarantee that its error probability is less than δ . Interestingly, to prove Theorem 5.10 we first prove a bound on the unverifiable sample complexity of EB-TC$_{\varepsilon_0}$-1/2 which is valid for all (ε, δ) , neither of which are parameters of the algorithm. More precisely, we prove that $\mathbb{P}_\nu\left(\forall n > U_{i_\mu(\varepsilon),\delta}(\mu, \varepsilon_0),\ \hat{\imath}_n \in \mathcal{I}_\varepsilon(\mu)\right) \geq 1 - \delta$ for $U_{i,\delta}(\mu, \varepsilon_0) =_{\delta \to 0} 8H_i(\mu, \varepsilon_0)\log(1/\delta) + \mathcal{O}(\log\log(1/\delta))$. As this statement is valid for all $\delta \in (0, 1)$, applying it for each n to δ_n such that $U_{i_\mu(\varepsilon),\delta_n}(\mu, \varepsilon_0) = n$, we obtain Theorem 5.10. We remark that such a trick cannot be applied to BUCB to get uniform ε-error bounds for any time, as the algorithm does depend on δ .

Simple regret As already noted by Zhao et al. [2023], uniform ε-error bounds easily yield simple regret bounds. We state in Corollary 5.11 the one obtained for EB-TC$_{\varepsilon_0}$. As a motivation to derive simple regret bounds, we observe that they readily translate to bounds on the cumulative regret for an agent observing the stream of recommendations $(\hat{\imath}_n)$ and playing

arm $\hat{\imath}_n$. An exponentially decaying simple regret leads to a constant cumulative regret in this decoupled exploration/exploitation setting [Avner et al., 2012, Rouyer and Seldin, 2020].

Corollary 5.11. *Let $\varepsilon_0 > 0$. Let $p(x)$ and $(H_i(\mu, \varepsilon_0))_{i\in[C_\mu - 1]}$ be defined as in Theorem 5.10. The EB-TC$_{\varepsilon_0}$-1/2 algorithm satisfies that, for all $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathbb{R}^K$, for all $n > 5K^2/2$,*

$$\mathbb{E}_\nu[\mu_\star - \mu_{\hat{\imath}_n}] \leq K^2 e^2 (2 + \log n)^2 \sum_{i\in[C_\mu - 1]} (\Delta_{i+1} - \Delta_i) \exp\left(-p\left(\frac{n - 5K^2/2}{8H_i(\mu, \varepsilon_0)}\right)\right) .$$

Proof. Using $\mathbb{E}_\nu[\mu_\star - \mu_{\hat{\imath}_n}] = \int \mathbb{P}_\nu(\hat{\imath}_n \notin \mathcal{I}_\varepsilon(\mu))\mathrm{d}\,\varepsilon$, it is a direct corollary of Theorem 5.10. ■

Following the discussion above, this bound is not expected to be state-of-the-art, it rather justifies that EB-TC$_{\varepsilon_0}$ with $\varepsilon_0 > 0$ is not too much worse than the uniform sampling strategy. Yet, as we shall see in our experiments, the practical story is different. In Section 5.5, we compare the simple regret of EB-TC$_{\varepsilon_0}$-1/2 to that of DSH in synthetic experiments with a moderate number of arms, revealing the superior performance of EB-TC$_{\varepsilon_0}$-1/2.

A versatile algorithm In Section 2.2.6, we explained why Top Two algorithms are simple, interpretable and generalizable. According to the requirements we set to ourselves in Section 1.1, versatility is the last characteristic of Top Two algorithms which needs to be illustrated. Based on the previous results, EB-TC$_{\varepsilon_0}$-1/2 is truly a versatile algorithm which enjoys guarantees that hold at any deterministic time. Moreover, EB-TC$_{\varepsilon_0}$-1/2 can be used for different downstream tasks, and we refer the reader to Section 1.5.2 for more details on this aspect.

5.4.1 Proof of Theorem 5.10

Let $\Delta_{\max} = \max_{i\in[K]} \mu_\star - \mu_i$. Let (ε_1, δ) be analysis parameters, *i.e.* not the error/confidence pairs from the stopping rule (5.3). When $\varepsilon_1 \geq \Delta_{\max}$, we have $\mathcal{I}_{\varepsilon_1}(\mu)^\complement = \emptyset$, hence $\mathbb{P}_\nu(\hat{\imath}_n \notin \mathcal{I}_{\varepsilon_1}(\mu)) = 0$. In the following, we consider $\varepsilon_1 \in [0, \Delta_{\max})$.

We first define concentration events to control the deviations of the random variables used in the EB leader and the TC$_{\varepsilon_0}$ challenger. For all $n > K$ and $\delta \in (0, 1)$, let $\tilde{\mathcal{E}}_{n,\delta} = \tilde{\mathcal{E}}^1_{n,\delta} \cap \tilde{\mathcal{E}}^2_{n,\delta}$ with

$$\tilde{\mathcal{E}}^1_{n,\delta} = \left\{\forall i \in [K], \forall t \leq n,\ |\mu_{t,i} - \mu_i| < \sqrt{2\tilde{f}_1(n,\delta)/N_{t,i}}\right\} ,$$

$$\tilde{\mathcal{E}}^2_{n,\delta} = \left\{\forall (i,j) \in [K]^2 \text{ s.t. } i \neq j,\ \forall t \leq n,\ \frac{|(\mu_{t,i} - \mu_{t,j}) - (\mu_i - \mu_j)|}{\sqrt{1/N_{t,i} + 1/N_{t,j}}} < \sqrt{2\tilde{f}_2(n,\delta)}\right\} ,$$

where $\tilde{f}_1(n,\delta) = \frac{1}{2}\overline{W}_{-1}(2\log(1/\delta) + 2\log(2+\log n) + 2)$ with $\overline{W}_{-1}(x) = -W_{-1}(-e^{-x})$ for all $x \geq 1$, and $\tilde{f}_2(n,\delta) = \overline{W}_{-1}\left(\log(1/\delta) + 2\log(2+\log n) + 2\right)$. We recall that $\overline{W}_{-1}(x) \approx x + \log(x)$ (see Appendix A), and one can show that $\tilde{f}_1(n,\delta) \leq \tilde{f}_2(n,\delta)$. Using concentration arguments, it is straightforward to show that $\mathbb{P}_\nu(\tilde{\mathcal{E}}_{n,\delta}^{\complement}) \leq K^2\delta$.

Using Lemma 5.12, the proof boils down to constructing an infinite number of times $\{T_{\varepsilon_1}(\delta,\varepsilon)\}_{\varepsilon\in[0,\varepsilon_1]}$ such that $\tilde{\mathcal{E}}_{n,\delta} \subseteq \{\hat{\imath}_n \in \mathcal{I}_{\varepsilon_1}(\mu)\}$ for $n > \inf_{\varepsilon\in[0,\varepsilon_1]} T_{\varepsilon_1}(\delta,\varepsilon)$ since it yields that

$$\mathbb{P}_\nu(\hat{\imath}_n \notin \mathcal{I}_{\varepsilon_1}(\mu)) \leq K^2 \inf\{\delta \mid n > \inf_{\varepsilon\in[0,\varepsilon_1]} T_{\varepsilon_1}(\delta,\varepsilon)\} .$$

Lemma 5.12. *Let $\varepsilon \geq 0$ and $(\mathcal{E}_{n,\delta})_{n>K,\delta\in(0,1)}$ such that $\mathbb{P}_\nu(\mathcal{E}_{n,\delta}^{\complement}) \leq C\delta$ with $C > 0$. Let $T_\varepsilon(\delta) > K$ such that $\mathcal{E}_{n,\delta} \subseteq \{\hat{\imath}_n \in \mathcal{I}_\varepsilon(\mu)\}$ for $n > T_\varepsilon(\delta)$. Then, $\mathbb{P}_\nu(\hat{\imath}_n \notin \mathcal{I}_\varepsilon(\mu)) \leq C \inf\{\delta \mid n > T_\varepsilon(\delta)\}$.*

Proof. Using $\mathbb{P}_\nu(\hat{\imath}_n \notin \mathcal{I}_\varepsilon(\mu)) \leq \mathbb{P}_\nu(\mathcal{E}_{n,\delta}^{\complement}) \leq C\delta$ for all $n > T_\varepsilon(\delta)$ and taking the infimum. ■

Let $\varepsilon \in [0,\varepsilon_1]$. Since $\hat{\imath}_n = B_n^{\mathrm{EB}}$, we will construct a time $T_{\varepsilon_1}(\delta,\varepsilon)$ such that $\tilde{\mathcal{E}}_{n,\delta} \cap \{B_n^{\mathrm{EB}} \notin \mathcal{I}_{\varepsilon_1}(\mu)\} = \emptyset$ for all $n > T_{\varepsilon_1}(\delta,\varepsilon)$. Let us denote by $U_{\varepsilon,\varepsilon_1,t}(n,\delta)$ the set of undersampled arms which are not ε_1-good, *i.e.*

$$U_{\varepsilon,\varepsilon_1,t}(n,\delta) := \mathcal{I}_{\varepsilon_1}(\mu)^{\complement} \cap \left\{i \mid N_{t,i} \leq 4C_{\varepsilon,i}\tilde{f}_2(n,\delta)\right\} \quad \text{with} \quad C_{\varepsilon,i} := \frac{2}{\min_{k\in\mathcal{I}_\varepsilon(\mu)}(\Delta_i - \Delta_k)^2} .$$

Lemma 5.13 shows that, for n large enough, a necessary condition for an error to occur is to have an undersampled leader. The proof of Lemma 5.13 is detailed in Appendix E.5.

Lemma 5.13. *Let $\varepsilon_1 > 0$ and $\varepsilon \in [0,\varepsilon_1]$. Let $H_{\mu,\varepsilon_0}(\varepsilon)$ as in Lemma 5.9 and define*

$$\tilde{H}_{\varepsilon,\varepsilon_1}(\mu,\varepsilon_0) := H_{\mu,\varepsilon_0}(\varepsilon) + \frac{2|\mathcal{I}_\varepsilon(\mu)|}{\min_{(i,j)\in\mathcal{I}_\varepsilon(\mu)\times\mathcal{I}_{\varepsilon_1}(\mu)^{\complement}}(\Delta_j - \Delta_i)^2} . \tag{5.7}$$

Let us define $S_{\varepsilon,\varepsilon_1,\varepsilon_0,\mu}(\delta) = \sup\left\{n \mid n \leq \frac{4\tilde{H}_{\varepsilon,\varepsilon_1}(\mu,\varepsilon_0)}{\min\{\beta,1-\beta\}}\tilde{f}_2(n,\delta) + (3/2+1/\beta)K^2\right\}$. For all $n > S_{\varepsilon,\varepsilon_1,\varepsilon_0,\mu}(\delta)$, under the event $\tilde{\mathcal{E}}_{n,\delta}$, we have $B_n^{EB} \notin \mathcal{I}_{\varepsilon_1}(\mu)$ implies that $B_n^{EB} \in U_{\varepsilon,\varepsilon_1,n}(n,\delta)$.

Lemma 5.14 shows that, if there are still undersampled arms which are not ε_1-good, then either the leader or the challenger was not often selected as leader or challenger. The proof of Lemma 5.14 is detailed in Appendix E.6.

Lemma 5.14. *Let* $\varepsilon_1 \geq 0$ *and* $\varepsilon \in [0, \varepsilon_1]$ *. Let* $\Delta_\mu(\varepsilon) = \min_{k \notin \mathcal{I}_\varepsilon(\mu)} \Delta_k$ *,* $C_{\mu,\varepsilon_0}(\varepsilon) = 2/\Delta_\mu(\varepsilon) - \varepsilon_0^{-1}$ *,* $C_{\mu,\varepsilon_0}(\varepsilon, \varepsilon_1) = 2 \max_{j \notin \mathcal{I}_{\varepsilon_1}(\mu)} \frac{\Delta_j/\varepsilon_0 + 1}{\min_{k \in \mathcal{I}_\varepsilon(\mu)}(\Delta_j - \Delta_k)} + \varepsilon_0^{-1}$ *and* $A_{\varepsilon,\varepsilon_1,\varepsilon_0,i} = \max\{2/\Delta_\mu(\varepsilon)^2, C_{\mu,\varepsilon_0}(\varepsilon,\varepsilon_1)^2\}$ *for all* $i \in i^\star(\mu)$ *,* $A_{\varepsilon,\varepsilon_1,\varepsilon_0,i} = \max\{C_{\mu,\varepsilon_0}(\varepsilon)^2, C_{\mu,\varepsilon_0}(\varepsilon,\varepsilon_1)^2\}$ *for all* $i \in (\mathcal{I}_\varepsilon(\mu) \setminus i^\star(\mu)) \cup ([K] \setminus \mathcal{I}_{\varepsilon_1}(\mu))$ *, and otherwise* $A_{\varepsilon,\varepsilon_1,\varepsilon_0,i} = \max\{C_{\mu,\varepsilon_0}(\varepsilon)^2, C_{\mu,\varepsilon_0}(\varepsilon,\varepsilon_1)^2, 2/\Delta_i^2\}$ *for all* $i \in \mathcal{I}_{\varepsilon_1}(\mu) \setminus \mathcal{I}_\varepsilon(\mu)$ *. For all* $n > K$ *, under event* $\tilde{\mathcal{E}}_{n,\delta}$ *, for all* $t \in [n] \setminus [K]$ *such that* $U_{\varepsilon,\varepsilon_1,t}(n,\delta) \neq \emptyset$ *, there exists* $i_t \in [K]$ *such that*

$$T_t(i_t) \leq \frac{4\tilde{f}_2(n,\delta)}{\min\{\beta, 1-\beta\}} A_{\varepsilon,\varepsilon_1,\varepsilon_0,i_t} + 3(K-1)/2 \quad \text{and} \quad T_{t+1}(i_t) = T_t(i_t) + 1 \,.$$

Combining Lemma 5.8 and 5.14, Lemma 5.15 shows that, for n large enough, there is no undersampled arms which are not ε_1-good. The proof of Lemma 5.15 is detailed in Appendix E.7.

Lemma 5.15. *Let* $\varepsilon_1 \geq 0$ *and* $\varepsilon \in [0, \varepsilon_1]$ *. Let* $\Delta_\mu(\varepsilon)$ *,* $C_{\mu,\varepsilon_0}(\varepsilon,\varepsilon_1)$ *and* $C_{\mu,\varepsilon_0}(\varepsilon)$ *as in Lemma 5.14 and*

$$\begin{aligned}\bar{H}_{\varepsilon,\varepsilon_1}(\mu,\varepsilon_0) := {} & |i^\star(\mu)| \max\{\sqrt{2}\Delta_\mu(\varepsilon)^{-1}, C_{\mu,\varepsilon_0}(\varepsilon,\varepsilon_1)\}^2 \\ & + |(\mathcal{I}_\varepsilon(\mu) \setminus i^\star(\mu)) \cup ([K] \setminus \mathcal{I}_{\varepsilon_1}(\mu))| \max\{C_{\mu,\varepsilon_0}(\varepsilon), C_{\mu,\varepsilon_0}(\varepsilon,\varepsilon_1)\}^2 \\ & + \sum_{i \in \mathcal{I}_{\varepsilon_1}(\mu) \setminus \mathcal{I}_\varepsilon(\mu)} \max\{C_{\mu,\varepsilon_0}(\varepsilon), C_{\mu,\varepsilon_0}(\varepsilon,\varepsilon_1), \sqrt{2}\Delta_i^{-1}\}^2 \,. \end{aligned} \tag{5.8}$$

Let us define $T_{\varepsilon,\varepsilon_1,\varepsilon_0,\mu}(\delta) = \sup\left\{n \mid n \leq \frac{4\bar{H}_{\varepsilon,\varepsilon_1}(\mu,\varepsilon_0)}{\min\{\beta,1-\beta\}} \tilde{f}_2(n,\delta) + 3K^2/2\right\}$ *. For all* $n > T_{\varepsilon,\varepsilon_1,\varepsilon_0,\mu}(\delta)$ *, under the event* $\tilde{\mathcal{E}}_{n,\delta}$ *, we have* $U_{\varepsilon,\varepsilon_1,n}(n,\delta) = \emptyset$ *.*

Taking $\beta = 1/2$ and combining Lemmas 5.13 and 5.15, one can show that $\tilde{\mathcal{E}}_{n,\delta} \cap \{B_n^{\mathrm{EB}} \notin \mathcal{I}_{\varepsilon_1}(\mu)\} = \emptyset$ for all $n > E_{\varepsilon_1,\varepsilon_0,\mu}(\delta)$ with

$$E_{\varepsilon_1,\varepsilon_0,\mu}(\delta) := \sup\left\{n \mid n \leq 8 \inf_{\varepsilon \in [0,\varepsilon_1]} \max\{\bar{H}_{\varepsilon,\varepsilon_1}(\mu,\varepsilon_0), \tilde{H}_{\varepsilon,\varepsilon_1}(\mu,\varepsilon_0)\} \tilde{f}_2(n,\delta) + 5K^2/2\right\} .$$

Direct manipulation yields that $H_{i_\mu(\varepsilon_1)}(\mu,\varepsilon_0) \geq \inf_{\varepsilon \in [0,\varepsilon_1]} \max\{\bar{H}_{\varepsilon,\varepsilon_1}(\mu,\varepsilon_0), \tilde{H}_{\varepsilon,\varepsilon_1}(\mu,\varepsilon_0)\}$. An inversion formula allows to conclude the proof (Lemma E.5 in Appendix E.8).

5.5 Experiments

We assess the performance of the EB-TC$_{\varepsilon_0}$ algorithm on Gaussian instances both in terms of its empirical stopping time and its empirical simple regret, and we show that it performs favorably

compared to existing algorithms in both settings. For the sake of space, we only show the results for large sets of arms and for a specific instance with $|i^\star(\mu)| = 2$.

Empirical stopping time We study the impact of large sets of arms (up to $K = 1000$) in ε-BAI for $(\varepsilon, \delta) = (0.1, 0.01)$ on the "$\alpha = 0.3$" scenario of Jamieson and Nowak [2014] which sets $\mu_i = 1 - ((i-1)/(K-1))^\alpha$ for all $i \in [K]$. EB-TC$_{\varepsilon_0}$-IDS with slack $\varepsilon_0 = \varepsilon$ is compared to existing ε-BAI algorithms having low computational cost. This precludes algorithms such as ε-Track-and-Stop (TaS) [Garivier and Kaufmann, 2021], Sticky TaS [Degenne and Koolen, 2019] or ε-BAI adaptation of FWS [Wang et al., 2021] and DKM [Degenne et al., 2019]. We also compare EB-TC$_\varepsilon$-IDS to those algorithms on benchmarks with smaller number of arms (see Appendix J.2.2 in Jourdan et al. [2023b]). We show that EB-TC$_\varepsilon$ performs on par with ε-TaS and ε-FWS, but outperforms ε-DKM. As Top Two benchmarks with fixed $\beta = 1/2$, we consider T3C [Shang et al., 2020] (*i.e.* TS-TC-$1/2$), EB-TCI-$1/2$ [Jourdan et al., 2022] and UCB-TC-β (*i.e.* TTUCB [Jourdan and Degenne, 2023]). To provide a fair comparison, we adapt them to tackle ε-BAI by using the stopping rule (5.3) and by adapting their sampling rule to use the TC$_\varepsilon$ challenger from (5.2) (with a penalization $\log N_{n,i}$ for EB-TCI$_\varepsilon$). We use the heuristic threshold $c(n, \delta) = \log((1 + \log n)/\delta)$. While it is too small to ensure the (ε, δ)-PAC property, it still yields an empirical error which is several orders of magnitude lower than δ . Finally, we compare with LUCB [Kalyanakrishnan et al., 2012] and uniform sampling. For a fair comparison, LUCB uses $\sqrt{2c(n-1,\delta)/N_{n,i}}$ as bonus, which is also too small to yield valid confidence intervals. Our results are averaged over 100 runs, and the standard deviations are displayed. In Figure 5.1(a), we see that EB-TC$_\varepsilon$ performs on par with the ε-T3C heuristic, and significantly outperforms the other algorithms. While the scaling in K of ε-EB-TCI and LUCB appears to be close to the one of EB-TC$_\varepsilon$, UCB-TC$_\varepsilon$ and uniform sampling obtain a worse one. Figure 5.1(a) also reveals that the regularization ensured by the TC$_\varepsilon$ challenger is sufficient to ensure enough exploration, hence other exploration mechanisms are superfluous (TS/UCB leader or TCI$_\varepsilon$ challenger).

Anytime empirical simple regret The EB-TC$_{\varepsilon_0}$-$1/2$ algorithm with $\varepsilon_0 = 0.1$ is compared to existing algorithms on the instance $\mu = (0.6, 0.6, 0.55, 0.45, 0.3, 0.2)$ from Garivier and Kaufmann [2021], which has two best arms. As benchmark, we consider Doubling Successive Reject (DSR) and Doubling Sequential Halving (DSH), which are adaptations of the elimination based algorithms SR [Audibert et al., 2010] and SH [Karnin et al., 2013]. SR eliminates one arm with the worst empirical mean at the end of each phase, and SH eliminated half of them but drops past observations between each phase. These doubling-based algorithms have empirical error decreasing by steps: they change their recommendation only before they restart. In Figure 5.1(b), we plot the average of the simple regret over 10000 runs and the standard deviation of that average (which is too small to see clearly). We observe that EB-TC$_{\varepsilon_0}$-$1/2$ outperforms uniform sampling, as well as DSR and DSH, which both perform worse due to

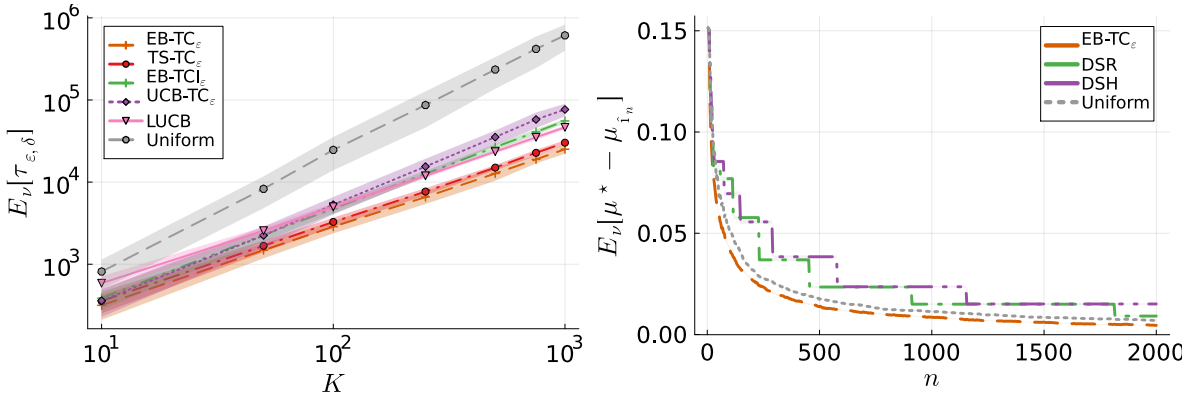

Figure 5.1 – (a) Empirical stopping time on "$\alpha = 0.3$" instances for varying K of EB-TC$_\varepsilon$-IDS with stopping rule (5.3) using $(\varepsilon, \delta) = (0.1, 0.01)$. (b) Empirical simple regret on instance $\mu = (0.6, 0.6, 0.55, 0.45, 0.3, 0.2)$ of EB-TC$_{\varepsilon_0}$-1/2 with slack $\varepsilon_0 = 0.1$.

the dropped observations. The favorable performance of EB-TC$_{\varepsilon_0}$-1/2 is confirmed on other instances from Garivier and Kaufmann [2021], and for "two-groups" instances with varying $|i^\star(\mu)|$ (see Figures 10 and 12 in Jourdan et al. [2023b]).

Supplementary experiments Extensive experiments and implementation details are available in Appendix J of Jourdan et al. [2023b]. In Appendix J.2.1, we compare the performance of EB-TC$_{\varepsilon_0}$ with different slacks ε_0 for IDS and $\beta = 1/2$. In Appendix J.2.2, we demonstrate the good empirical performance of EB-TC$_{\varepsilon_0}$ compared to state-of-the art methods in the fixed-confidence ε-BAI setting, compared to DSR and DSH for the empirical simple regret, and compared to SR and SH for the probability of 0-error in the fixed-budget setting (Figure 13). We consider a wide range of instances: random ones, benchmarks from the literature [Jamieson and Nowak, 2014, Garivier and Kaufmann, 2021] and "two-groups" instances with varying $|i^\star(\mu)|$.

5.6 Discussion

In Chapter 5, we proposed the EB-TC$_{\varepsilon_0}$ algorithm, which is easy to understand and implement. EB-TC$_{\varepsilon_0}$ is the first algorithm to be simultaneously asymptotically optimal in the fixed-confidence ε_0-BAI setting (Theorem 5.5), have finite-confidence guarantees (Theorem 5.6), and have also anytime guarantees on the probability of error at any level ε (Theorem 5.10), hence on the simple regret (Corollary 5.11). Furthermore, we demonstrated that the EB-TC$_{\varepsilon_0}$ algorithm achieves superior performance compared to other algorithms, in benchmarks where the number of arms is moderate to large.

While our results hold for general 1-sub-Gaussian distributions, the EB-TC$_{\varepsilon_0}$-IDS algorithm with slack $\varepsilon_0 > 0$ only achieves asymptotic optimality for ε_0-BAI with Gaussian bandits. It

would be interesting to have similar guarantees for other classes of distributions (see Chapters 3 and 4) Likewise, our non-asymptotic guarantees on $\mathbb{E}_\nu[\tau_{\varepsilon_0,\delta}]$ and $\mathbb{E}_\nu[\mu_\star - \mu_{\hat{\imath}_n}]$ were obtained for the EB-TC$_{\varepsilon_0}$-1/2 algorithm. Since better empirical performance are observed when using IDS, deriving similar (or better) non-asymptotic guarantees for IDS is an interesting avenue for future work.

The EB-TC$_{\varepsilon_0}$ algorithm is a promising method to tackle structured bandits. While heuristics exist for some structured bandits such as Top-k , it would be interesting to efficiently adapt Top Two methods to deal with sophisticated structure, *e.g.* linear bandits. In Part III, we will provide some elements of answer for ε-BAI in linear bandits. In particular, we extend the EB-TC$_{\varepsilon_0}$ algorithm in Chapter 8.

In this chapter, our initial motivation stems from a practical consideration, namely that we are often satisfied by any "good enough" answer if it avoids the wasteful queries required by BAI (see Part I). The ε-BAI problem is a natural setting in which the answer is good enough compared to the other answers. This instance-dependent goodness is defined relatively to other answers. Studied in Chapter 6, the good arm identification (GAI) problem is another possible setting in which the answer is good enough compared to a fixed threshold. This instance-independent goodness is defined globally, *i.e.* agnostic to the other arms.

Chapter 6

Good Arm Identification

In Chapter 6, we study the GAI problem for vanilla bandits in the anytime setting, as described in Chapter 1. The presented results are currently under review, see Jourdan and Réda [2023] for a preprint version.

In good arm identification (GAI), the goal is to identify one arm whose average performance exceeds a given threshold, referred to as good arm, if it exists. Few works have studied GAI in the fixed-budget setting, when the sampling budget is fixed beforehand, or the anytime setting, when a recommendation can be asked at any time. We propose APGAI, an anytime and parameter-free sampling rule for GAI in stochastic bandits. APGAI can be straightforwardly used in fixed-confidence and fixed-budget settings. First, we derive upper bounds on its probability of error at any time. They show that adaptive strategies are more efficient in detecting the absence of good arms than uniform sampling. Second, when APGAI is combined with a stopping rule, we prove upper bounds on the expected sampling complexity, holding at any confidence level. Finally, we show good empirical performance of APGAI on synthetic and real-world data. Our work offers an extensive overview of the GAI problem in all settings.

Contents

6.1 Introduction

As described in Sections 1.5 and 5.1, the motivation to study the anytime setting comes from practical considerations. Properly choosing the constraint on δ or T for the fixed-confidence and fixed-budget settings is challenging for the practitioner since a "good" choice typically depends on unknown quantities. Moreover, in medical applications (*e.g.* clinical trials or outcome scoring), the maximal budget is limited but might not be fixed beforehand. When the collected data shows sufficient evidence in favor of one answer, an experiment is often stopped before the initial budget is reached, referred to as *early stopping*. When additional sampling budget have been obtained due to new funding, an experiment can continue after the initial budget has been consumed, referred to as *continuation*. While early stopping and continuation are common practices, both fixed-confidence and fixed-budget settings fail to provide useful guarantees for them. Recently, the *anytime* setting has received increased scrutiny as it fills this gap between theory and practice. When the candidate answer has anytime guarantees, the practitioners can use continuation or early stopping (when combined with a stopping rule).

As described in Chapter 1, the motivation of the GAI problem stems from the sampling cost of the BAI problem To avoid wasteful queries, practitioners might be interested in easier tasks that identify one "good enough" option. While Chapter 5 tackled the ε-BAI problem, we address the good arm identification (GAI) problem in this chapter. The agent aims to obtain a *good arm*, which is defined as an arm whose average performance exceeds a given threshold γ , *i.e.* $\mu_i \geq \gamma$. For instance, in our outcome scoring problem (see Section 6.5), practitioners have enough information about the distributions to define a meaningful threshold beforehand. GAI and variants have been studied in the fixed-confidence setting [Kaufmann et al., 2018, Kano et al., 2019, Tabata et al., 2020], but algorithms for fixed-budget or anytime GAI are missing despite their practical relevance. In this chapter, we fill this gap by introducing APGAI, an anytime and parameter-free sampling rule for GAI which is independent of a parameter T or δ and can be used in the fixed-budget and fixed-confidence settings.

We consider the set $\mathcal{D}_\sigma$ of σ-sub-Gaussian distributions and assume that $\sigma_i = 1$ for all $i \in [K]$ by scaling, hence $\mathcal{D}^K = \mathcal{D}_1^K$. Let $\nu \in \mathcal{D}^K$ with mean vector $\mu \in \mathbb{R}^K$. Given a threshold $\gamma \in \mathbb{R}$, the set of good arms is defined as $\mathcal{I}_\gamma^{\text{thr}}(\mu) := \{i \in [K] \mid \mu_i \geq \gamma\}$, which we shorten to $\mathcal{I}_\gamma^{\text{thr}}$ when μ is unambiguous. In GAI, we consider $\mathcal{S} = \{\mu \in \mathcal{R}^K \mid \min_{i\in[K]} |\mu_i - \gamma| > 0\}$. Let the gap of arm i compared to γ be $\Delta_{\gamma,i} := |\mu_i - \gamma| > 0$. Let $\Delta_{\gamma,\min} = \min_{i\in[K]} \Delta_{\gamma,i}$ be the minimum gap over all arms. Let

$$H_1(\mu) := \sum_{i\in[K]} \Delta_{\gamma,i}^{-2} \quad \text{and} \quad H_\gamma(\mu) := \sum_{i\in\mathcal{I}_\gamma^{\text{thr}}(\mu)} \Delta_{\gamma,i}^{-2} \,. \tag{6.1}$$

At time n , we denote by $\hat{\imath}_n \in [K] \cup \{\emptyset\}$ the candidate answer and by $I_n \in [K]$ the arm to pull next. The probability of error $P^{\text{err}}_{\nu,\mathfrak{A}}(n) := \mathbb{P}_\nu(\mathcal{E}^{\text{err}}_\mu(n))$ of algorithm $\mathfrak{A}$ on instance ν at time n is the probability of the error event $\mathcal{E}^{\text{err}}_\mu(n) = \{\hat{\imath}_n \in \{\emptyset\} \cup ([K] \setminus \mathcal{I}^{\text{thr}}_\gamma(\mu))\}$ when $\mathcal{I}^{\text{thr}}_\gamma(\mu) \neq \emptyset$, otherwise $\mathcal{E}^{\text{err}}_\mu(n) = \{\hat{\imath}_n \neq \emptyset\}$ when $\mathcal{I}^{\text{thr}}_\gamma(\mu) = \emptyset$.

Fixed-confidence GAI For fixed-confidence GAI, the algorithm is augmented by a stopping rule (and a stopping time $\tau^{\text{thr}}_{\gamma,\delta}$) using a fixed confidence level $1-\delta \in (0,1)$ which ensures δ-correctness, *i.e.* $\mathbb{P}_\nu(\{\tau^{\text{thr}}_{\gamma,\delta} < +\infty\} \cap \mathcal{E}^{\text{err}}_\mu(\tau^{\text{thr}}_{\gamma,\delta})) \leq \delta$ for all instances $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathcal{S}$. That requirement leads to a lower bound on the expected sample complexity on any instance.

Lemma 6.1 (Theorem 1 in Degenne and Koolen [2019]). *Let $\delta \in (0,1)$ and $\gamma \in \mathbb{R}$. For all δ-correct algorithms and all instances $\nu = \mathcal{N}(\mu, 1_K)$ with mean $\mu \in \mathcal{S}$, $\liminf_{\delta\to 0} \mathbb{E}_\nu[\tau^{\text{thr}}_{\gamma,\delta}]/\log(1/\delta) \geq T^{\text{thr}}_\gamma(\nu)$, where $T^{\text{thr}}_\gamma(\nu) = 2\min_{i\in\mathcal{I}^{\text{thr}}_\gamma(\mu)} \Delta^{-2}_{\gamma,i}$ when $\mathcal{I}^{\text{thr}}_\gamma(\mu) \neq \emptyset$, and $T^{\text{thr}}_\gamma(\nu) = 2H_1(\mu)$ otherwise.*

A fixed-confidence algorithm is said to be *asymptotically optimal* if it is δ-correct, and its expected sample complexity matches the lower bound, *i.e.* $\limsup_{\delta\to 0} \mathbb{E}_\nu[\tau^{\text{thr}}_{\gamma,\delta}]/\log(1/\delta) \leq T^{\text{thr}}_\gamma(\nu)$ for all instances $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathcal{S}$.

Contribution 6.2. *In Chapter 6, we propose the APGAI algorithm, an anytime and parameter-free sampling rule for GAI in stochastic bandits, which is independent of a budget T or a confidence δ . APGAI is the first algorithm which can be employed without modification for fixed-budget GAI (and without prior knowledge of the budget) and fixed-confidence GAI. Furthermore, it enjoys guarantees in both settings. As such, APGAI allows both continuation and early stopping.*

- *We show an upper bound on $P^{err}_{\nu,APGAI}(n)$ of the order $\exp(-\mathcal{O}(n/H_1(\mu)))$ which holds for any deterministic time n (Theorem 6.3). Adaptive strategies are more efficient in detecting the absence of good arms than uniform sampling (see Section 6.3).*
- *When combined with a GLR stopping rule, we derive an upper bound on $\mathbb{E}_\nu[\tau^{\text{thr}}_{\gamma,\delta}]$ holding at any confidence level (Theorem 6.13). In particular, APGAI is asympotically optimal for GAI with Gaussian distributions when there is no good arm.*
- *APGAI is easy to implement, computationally inexpensive and achieves good empirical performance in both settings on synthetic and real-world data with an outcome scoring problem for RNA-sequencing data (see Section 6.5).*

This chapter offers an overview of the GAI problem in all settings.

6.1.1 Related Work

GAI has never been studied in the fixed-budget or anytime setting. In the fixed-confidence setting, several questions have been studied which are closely connected to GAI. Given two thresholds $\gamma_L < \gamma_U$, Tabata et al. [2020] studies the Bad Existence Checking problem, in which the agent should output "negative" if $\mathcal{I}^{\mathrm{thr}}_{\gamma_L}(\mu) = \emptyset$ and "positive" if $\mathcal{I}^{\mathrm{thr}}_{\gamma_U}(\mu) \neq \emptyset$. They propose an elimination-based meta-algorithm called BAEC, and analyze its expected sample complexity when combined with several index-policy to define the sampling rule. Kano et al. [2019] considers identifying the whole set of good arms $\mathcal{I}^{\mathrm{thr}}_{\gamma}(\mu)$ with high probability, and returns the good arms in a sequential way. We refer to that problem as AllGAI. In Kano et al. [2019], they introduce three index-based GAI algorithms named APT-G, HDoC and LUCB-G, and show upper bounds on their expected sample complexity. A large number of algorithms from previously mentioned works bear a passing resemblance to the APT algorithm in Locatelli et al. [2016] which tackles the thresholding bandit problem in the fixed-budget setting. The latter should classify all arms into $\mathcal{I}^{\mathrm{thr}}_{\gamma}(\mu)$ and $\mathcal{I}^{\mathrm{thr}}_{\gamma}(\mu)^{\complement}$ at the end of the sampling phase. This resemblance lies in that those algorithms rely on an arm index for sampling, yet the arm indices in BAEC [Tabata et al., 2020], APT-G, HDoC and LUCB-G Kano et al. [2019] are different.

Degenne et al. [2019] tackle the "any low arm" problem, which is a GAI problem for threshold $-\gamma$ on instance $-\mu$. They introduce Sticky Track-and-Stop, which is asymptotically optimal in the fixed-confidence setting. In Kaufmann et al. [2018], the "bad arm existence" problem aims to answer "no" when $\mathcal{I}^{\mathrm{thr}}_{-\gamma}(-\mu) = \emptyset$, and "yes" otherwise. They adapt Thompson Sampling by conditioning on the "worst event" (named Murphy Sampling). The empirical pulling proportions are shown to converge towards the allocation realizing $T^{\mathrm{thr}}_{\gamma}(\nu)$ in Lemma 6.1. Another related framework is the identification with high probability of k arms from $\mathcal{I}^{\mathrm{thr}}_{\gamma}(\mu)$ [Katz-Samuels and Jamieson, 2020]. They introduce the *unverifiable sample complexity*. It is the minimum number of samples after which the algorithm always outputs a correct answer with high probability. It does not require to certify that the output is correct.

6.2 Anytime Parameter-free Sampling Rule

We propose the APGAI (Anytime Parameter-free GAI) algorithm, which is independent of a budget T or a confidence δ and is summarized in Algorithm 6.1.

Recommendation rule Let $N_{n,i} = \sum_{t\in[n-1]} \mathbb{1}(I_t = i)$ and $\mu_{n,i} = N_{n,i}^{-1} \sum_{t\in[n-1]} \mathbb{1}(I_t = i) X_{n,i}$ be the empirical count and empirical mean of arm $i \in [K]$ before time n . At time $n > K$, the recommendation rule depends on whether the highest empirical mean lies below the threshold γ or not. When $\max_{i\in[K]} \mu_{n,i} \leq \gamma$, we recommend the empty set, *i.e.* $\hat{\imath}_n = \emptyset$. Otherwise, our

Input: Threshold γ .
Output: Next arm to sample I_n and next recommendation $\hat{\imath}_n$.
Set $\hat{\imath}_n \in \begin{cases} \{\emptyset\} & \text{if } \max_i \mu_{n,i} \leq \gamma \\ \arg\max_i \sqrt{N_{n,i}}(\mu_{n,i} - \gamma)_+ & \text{otherwise} \end{cases}$; // Recommendation
Set $I_n \in \begin{cases} \arg\min_{i \in [K]} \sqrt{N_{n,i}}(\gamma - \mu_{n,i})_+ & \text{if } \max_{i \in [K]} \mu_{n,i} \leq \gamma \\ \{\hat{\imath}_n\} & \text{otherwise} \end{cases}$; // Arm to pull

Algorithm 6.1: APGAI algorithm.

candidate answer is the arm which is the most likely to be a good arm given the collected evidence, *i.e.* $\hat{\imath}_n \in \arg\max_{i \in [K]} \sqrt{N_{n,i}}(\mu_{n,i} - \gamma)_+$.

Sampling rule The next arm to pull is based on the APT$_P$ indices introduced by [Tabata et al., 2020] as a modification to the APT indices [Locatelli et al., 2016]. At time $n > K$, we pull arm $I_n \in \arg\max_{i \in [K]} \sqrt{N_{n,i}}(\mu_{n,i} - \gamma)$. To emphasize the link with our recommendation rule, this sampling rule can also be written as $I_n \in \arg\min_{i \in [K]} \sqrt{N_{n,i}}(\gamma - \mu_{n,i})_+$ when $\max_{i \in [K]} \mu_{n,i} \leq \gamma$, and $I_n \in \arg\max_{i \in [K]} \sqrt{N_{n,i}}(\mu_{n,i} - \gamma)_+$ otherwise. Ties are broken arbitrarily at random, up to the constraint that $\hat{\imath}_n = I_n$ when $\max_{i \in [K]} \mu_{n,i} > \gamma$. This formulation better highlights the dual behavior of APGAI. When $\max_{i \in [K]} \mu_{n,i} \leq \gamma$, APGAI collects additional observations to verify that there are no good arms, hence pulling the arm which is the least likely to not be a good arm. Otherwise, APGAI gathers more samples to confirm its current belief that there is at least one good arm, hence pulling the arm which is the most likely to be a good arm.

Differences to BAEC While both APGAI and BAEC(APT$_P$) rely on the APT$_P$ indices [Tabata et al., 2020], they differ significantly. BAEC is an elimination-based meta-algorithm which samples active arms and discards arms whose upper confidence bounds (UCB) on the empirical means are lower than γ_U . The recommendation rule of BAEC is only defined at the stopping time, and it depends on lower confidence bounds (LCB) and UCB. Since the UCB/LCB indices depend inversely on the gap $\gamma_U - \gamma_L > 0$ and on the confidence δ , BAEC is neither anytime nor parameter-free. More importantly, APGAI can be used without modification for fixed-confidence or fixed-budget GAI. In contrast, BAEC can solely be used in the fixed-confidence setting when $\gamma_U > \gamma_L$, hence not for GAI itself (*i.e.* $\gamma_U = \gamma_L$).

6.3 Anytime Guarantees on the Probability of Error

To allow continuation or (deterministic) early stopping, the candidate answer of APGAI should be associated with anytime theoretical guarantees. Theorem 6.3 shows an upper bound of

the order $\exp(-\mathcal{O}(n/H_1(\mu)))$ for $P^{\text{err}}_{\nu,\mathfrak{A}}(n)$ that holds for any deterministic time n . The proof of Theorem 6.3 is sketch in Section 6.3.3.

Theorem 6.3. *Let $p(x) = x - 0.5\log x$. The APGAI algorithm $\mathfrak{A}$ satisfies than, for all $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathcal{S}$, for all $n > K + 2|\mathcal{I}^{\text{thr}}_\gamma(\mu)|$,*

$$P^{err}_{\nu,\mathfrak{A}}(n) \leq Ke\sqrt{2}\log(e^2 n)\exp\left(-p\left(\frac{n-K-2|\mathcal{I}^{\text{thr}}_\gamma(\mu)|}{2\alpha_{i_\mu}H_1(\mu)}\right)\right)$$

where $H_1(\mu)$ as in (6.1), $(\alpha_1, \alpha_\gamma) = (9, 2)$ and $i_\mu = 1 + (\gamma - 1)\mathbb{1}\left(\mathcal{I}^{\text{thr}}_\gamma(\mu) \neq \emptyset\right)$.

Theorem 6.3 holds for any deterministic time $n > K + 2|\mathcal{I}^{\text{thr}}_\gamma(\mu)|$ and any 1-sub-Gaussian instance ν . In the asymptotic regime where $n \to +\infty$, Theorem 6.3 shows that APGAI satisfies

$$\limsup_{n\to+\infty} \frac{n}{-\log P^{\text{err}}_{\nu,\mathfrak{A}}(n)} \leq 2\alpha_{i_\mu}H_1(\mu) \quad \text{with} \quad (\alpha_1, \alpha_\gamma) = (9, 2)\,.$$

Comparison with uniform sampling Despite the practical relevance of anytime and fixed-budget guarantees, APGAI is the first algorithm enjoying guarantees on the probability of error in GAI at any time n (hence at a given budget T). As baseline, we consider the uniform round-robin algorithm, named Unif, which returns the best empirical arm at time n if its empirical mean is higher than γ , and returns $\emptyset$ otherwise. Let n such that $(n-1)/K \in \mathbb{N}$, the recommendation of Unif is equivalent to the one used in APGAI, *i.e.* $\arg\max_{i\in[K]} \sqrt{N_{n,i}}(\mu_{n,i} - \gamma)_+ = \arg\max_{i\in[K]} \mu_{n,i}$. As the two algorithms only differ by their sampling rule, we can measure the benefits of adaptive sampling. It is possible to derive anytime upper bounds on $P^{\text{err}}_{\nu,\text{Unif}}(n)$ (see Theorem 4 in Appendix C of Jourdan and Réda [2023]). In the asymptotic regime, Unif achieves a rate in $2K\Delta^{-2}_{\gamma,\min}$ when $\mathcal{I}^{\text{thr}}_\gamma(\mu) = \emptyset$, and $4K\min_{i\in\mathcal{I}^{\text{thr}}_\gamma(\mu)} \Delta^{-2}_{\gamma,i}$ otherwise. While the latter rate is better than $2H_1(\mu)$ when arms have dissimilar gaps, APGAI has better guarantees than Unif when there is no good arm. Our experiments shows that APGAI outperforms Unif on most instances (Figures 6.1 and 6.2), and is on par with it otherwise.

Worst-case lower bound Degenne [2023] recently studied the existence of a complexity in fixed-budget pure exploration. While there is a complexity $T^{\text{thr}}_\gamma(\nu)$ as in Lemma 6.1 for the fixed-confidence setting, Theorem 6 from Degenne [2023] shows that a sequence of fixed-budget algorithms $(\mathfrak{A}_T)_T$ (where $\mathfrak{A}_T$ denotes the algorithm using fixed budget T) cannot have a better asymptotic rate than $KT^{\text{thr}}_\gamma(\nu)$ on all Gaussian instances

$$\exists \nu \in \mathcal{D}^K, \quad \limsup_{T\to+\infty} \frac{T}{-\log P^{\text{err}}_{\nu,\mathfrak{A}_T}(T)} \geq KT^{\text{thr}}_\gamma(\nu)\,. \tag{6.2}$$

Unif achieves the rate $KT_\gamma^{\text{thr}}(\nu)$ when $\mathcal{I}_\gamma^{\text{thr}}(\mu) \neq \emptyset$, but suffers from worse guarantees otherwise. Conversely, APGAI achieves the rate in $T_\gamma^{\text{thr}}(\nu)$ when $\mathcal{I}_\gamma^{\text{thr}}(\mu) = \emptyset$, but has sub-optimal guarantees otherwise. It does not conflict with (6.2) *e.g.* considering μ with $\mathcal{I}_\gamma^{\text{thr}} \neq \emptyset$ and such that there exists an arm $i \in [K]$ with $\Delta_{\gamma,i} \leq \max_{i \in \mathcal{I}_\gamma^{\text{thr}}(\mu)} \Delta_{\gamma,i}/\sqrt{K/2-1}$. Experiments in Section 6.5 suggest that the sub-optimal dependency when $\mathcal{I}_\gamma^{\text{thr}}(\mu) \neq \emptyset$ is not aligned with the good practical performance of APGAI. Formally proving better guarantees when $\mathcal{I}_\gamma^{\text{thr}}(\mu) \neq \emptyset$ is a direction for future work.

In fixed-budget GAI, a good strategy has different sampling modes depending on whether there is a good arm or not. Since wrongfully committing to one of those modes too early will incur higher error, it is challenging to find the perfect trade-off adaptively. Designing an algorithm whose guarantees are comparable to (6.2) for all instances is an open problem.

6.3.1 Benchmark: Other GAI Algorithms

To go beyond the comparison with Unif, we propose and analyze additional GAI algorithms. A summary of the comparison with APGAI is shown in Table 6.1.

From BAI to GAI Algorithms

Since a BAI algorithm outputs the arm with highest mean, it can be adapted to GAI by comparing the mean of the returned arm to the known threshold. We study the GAI adaptations of two fixed-budget BAI algorithms: Successive Rejects (SR) [Audibert et al., 2010] and Sequential Halving (SH) [Karnin et al., 2013]. SR-G and SH-G return $\hat{\imath}_T = \emptyset$ when $\mu_{n,i_T} \leq \gamma$ and $\hat{\imath}_T = i_T$ otherwise, where i_T is the arm that would be recommended for the BAI problem, *i.e.* the last arm that was not eliminated.

It is possible to derive an upper bound on $P^{\text{err}}_{\nu,\text{SR-G}}(T)$ and $P^{\text{err}}_{\nu,\text{SH-G}}(T)$ at the fixed budget T (see Theorems 5 and 6 in Appendix C of Jourdan and Réda [2023]). In the asymptotic regime, their rate is in $4\log(K)\Delta^{-2}_{\gamma,\min}$ when $\mathcal{I}_\gamma^{\text{thr}}(\mu) = \emptyset$, otherwise

$$\mathcal{O}(\log(K)\max\{\max_{i \in \mathcal{I}_\gamma^{\text{thr}}} \Delta^{-2}_{\gamma,i}, \max_{k > I^\star} k(\max_{i \in [K]} \mu_i - \mu_{(k)})^{-2}\})$$

with $I^\star = |\arg\max_{i \in [K]} \mu_i|$ and $\mu_{(k)}$ be the k^{th} largest mean in vector μ . Recently, Zhao et al. [2023] have provided a finer analysis of SH. Using their result yields improved rates. Those rates are better than $2H_1(\mu)$ when there is one good arm with large mean and the remaining arms have means slightly smaller than γ . However, APGAI has better guarantees than SR-G and SH-G when there is one good arm with mean slightly smaller than the largest mean.

Table 6.1 – Asymptotic error rate $C(\mu)$ of algorithm $\mathfrak{A}$ on ν , *i.e.* $\limsup_{n\to+\infty} n\log(1/P^{\mathrm{err}}_{\nu,\mathfrak{A}}(n))^{-1} \leq C(\mu)$. (†) Fixed-budget algorithm $\mathfrak{A}_{n,\nu}$ with prior knowledge on ν . $H_1(\mu)$ as in (6.1), $\Delta_{\gamma,\min} := \min_{i\in[K]} \Delta_{\gamma,i}$, $\widetilde{\Delta}^{-2} := \max\{\max_{i\in\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)} \Delta^{-2}_{\gamma,i}, \max_{k>I^\star} k(\max_{i\in[K]}\mu_i - \mu_{(k)})^{-2}\}$, $\overline{\Delta}_{\max} := \max_{i\in\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)} \Delta_{\gamma,i}$, $I^\star = |\arg\max_{i\in[K]}\mu_i|$ and $\hat{\Delta} := \max_{i\in\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)} \Delta_{\gamma,i} + \min_{i\notin\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)} \Delta_{\gamma,i}$.

Algorithm $\mathfrak{A}$	$\mathcal{I}^{\mathrm{thr}}_\gamma(\mu) = \emptyset$	$\mathcal{I}^{\mathrm{thr}}_\gamma(\mu) \neq \emptyset$
APGAI [Th. 6.3]	$18H_1(\mu)$	$4H_1(\mu)$
Unif [Th. 4 in Jourdan and Réda [2023]]	$2K\Delta^{-2}_{\gamma,\min}$	$4K\overline{\Delta}^{-2}_{\max}$
DSR-G [Th. 5 in Jourdan and Réda [2023]]	$16\log K\Delta^{-2}_{\gamma,\min}$	$4\log K\widetilde{\Delta}^{-2}$
DSH-G [Th. 6 in Jourdan and Réda [2023]]	$16\log K\Delta^{-2}_{\gamma,\min}$	$4\log K\widetilde{\Delta}^{-2}$
PKGAI($\star$) [Th. 7 in Jourdan and Réda [2023]]†	$2H_1(\mu)$	$2H_1(\mu)$
PKGAI(Unif) [Th. 8 in Jourdan and Réda [2023]]†	$2H_1(\mu)$	$2K\hat{\Delta}^{-2}$

Doubling trick The doubling trick allows the conversion of any fixed-budget algorithm into an anytime algorithm. It considers a sequence of algorithms that are run with increasing budgets $(T_k)_{k\geq 1}$, and recommends the answer outputted by the last instance. Zhao et al. [2023] shows that Doubling SH obtains the same guarantees than SH in BAI at the cost of a multiplicative factor 4 in the rate, similar results would hold for its GAI counterpart DSH-G (as well as for DSR-G). Empirically, our experiments show that APGAI is always better than DSR-G and DSH-G (Figures 6.1 and 6.2).

Prior Knowledge-based GAI Algorithms

Several fixed-budget BAI algorithms assume that the agent has access to some prior knowledge on unknown quantities to design upper/lower confidence bounds (UCB/LCB), *e.g.* UCB-E [Audibert et al., 2010] and UGapEb [Gabillon et al., 2012]. While this assumption is often not realistic, it yields better guarantees. We also investigate those approaches for fixed-budget GAI by proposing an elimination-based meta-algorithm for fixed-budget GAI called PKGAI (Prior Knowledge-based GAI, see Appendix D of Jourdan and Réda [2023]). As for BAEC, PKGAI($\star$) takes as input an index policy $\star$ which is used to define the sampling rule. The main difference to BAEC lies in the definition of the UCB/LCB since they depend both on the budget T and on knowledge of $H_1(\mu)$ and $H_\gamma(\mu)$.

We provide upper confidence bounds on the probability of error at time T holding for any choice of indices and for uniform round-robin sampling (Theorems 7 and 8 in Appendix D of Jourdan and Réda [2023]). The obtained upper bounds on $P^{\mathrm{err}}_{\nu,\mathrm{PKGAI}}(T)$ are marginally lower than the ones obtained for APGAI, while APGAI does not require to know $H_1(\mu)$ or $H_\gamma(\mu)$.

6.3.2 Unverifiable Sample Complexity

The *unverifiable sample complexity* was defined in Katz-Samuels and Jamieson [2020] as the smallest stopping time $\tau^{\text{thr}}_{U,\gamma,\delta}$ after which an algorithm always outputs a correct answer with probability at least $1-\delta$. In GAI, this means that algorithm $\mathfrak{A}$ satisfies $\mathbb{P}_\nu(\bigcup_{n\geq\tau^{\text{thr}}_{U,\gamma,\delta}} \mathcal{E}^{\text{err}}_\mu(n)) \leq \delta$. Compared to the fixed-confidence setting, it does not require to certify that the candidate answer is correct. Authors in Zhao et al. [2023] notice that anytime bounds on the error can imply an unverifiable sample complexity bound. It is possible to derive a deterministic upper bound on the unverifiable sample complexity $\tau^{\text{thr}}_{U,\gamma,\delta}$ of APGAI, *i.e.*

$$U_\delta(\mu) =_{\delta\to 0} 2\alpha_{i_\mu} H_1(\mu)\log(1/\delta) + \mathcal{O}(\log\log(1/\delta)) \,,$$

with $i_\mu = 1 + (\gamma-1)\mathbb{1}\left(\mathcal{I}^{\text{thr}}_\gamma(\mu)\neq\emptyset\right)$ and $(\alpha_1,\alpha_\gamma) = (9,2)$ (see Theorem 3 in Appendix B.3 of Jourdan and Réda [2023]). While such upper bounds are known in BAI [Katz-Samuels and Jamieson, 2020, Zhao et al., 2023, Jourdan et al., 2023b], this is the first result for GAI.

6.3.3 Proof of Theorem 6.3

To prove Theorem 6.3, we use a similar technique as the one detailed in Section 5.4.1 of Chapter 5. Instead of relying on Lemma 5.8, the key technical tool is Lemma 6.4 (proven in Appendix F.1).

Lemma 6.4. *Let $\delta\in(0,1]$ and $n>K$. Assume there exists a sequence of events $(A_t(n,\delta))_{K<t\leq n}$ and positive reals $(D_i(n,\delta))_{i\in[K]}$ such that, for all $t\in\{K+1,\cdots n\}$, under the event $A_t(n,\delta)$,*

$$\exists i_t\in[K],\quad N_{t,i_t}\leq D_{i_t}(n,\delta)\quad \textit{and}\quad N_{t+1,i_t} = N_{t,i_t}+1\,.$$

Then, we have $\sum_{t=K+1}^n \mathbb{1}\left(A_t(n,\delta)\right) \leq \sum_{i\in[K]} D_i(n,\delta)$.

We first define concentration events to control the deviations of the random variables used by APGAI. For all $n>K$ and $\delta\in(0,1)$, let

$$\tilde{\mathcal{E}}_{n,\delta} = \left\{\forall i\in[K],\forall t\leq n,\ |\mu_{n,i}-\mu_i| < \sqrt{2\tilde{f}_1(n,\delta)/N_{n,i}}\right\}\,,$$

with $\tilde{f}_1(n,\delta) = \frac{1}{2}\overline{W}_{-1}(2\log(1/\delta)+2\log(2+\log n)+2)$ with $\overline{W}_{-1}(x) = -W_{-1}(-e^{-x})$ for all $x\geq 1$. We recall that $\overline{W}_{-1}(x)\approx x+\log(x)$ (see Appendix A). Using concentration arguments, it is straightforward to show that $\mathbb{P}_\nu(\tilde{\mathcal{E}}^{\complement}_{n,\delta})\leq K\delta$.

Using Lemma 5.12 in Chapter 5, the proof boils down to constructing a time $T_\mu(\delta)$ such that $\tilde{\mathcal{E}}_{n,\delta}\subseteq\mathcal{E}^{\text{err}}_\mu(n)$ for $n>T_\mu(\delta)$ since it yields that $P^{\text{err}}_{\nu,\mathfrak{A}}(n)\leq K\inf\{\delta\mid n>T_\mu(\delta)\}$.

Instances where $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) = \emptyset$ In that case, we have $\mathcal{E}_\mu^{\mathrm{err}}(n) = \{\hat{\imath}_n \neq \emptyset\}$. Let us denote by $U_t(n,\delta) = \left\{i \in [K] \mid N_{t,i} \leq 2\tilde{f}_1(n,\delta)\Delta_{\gamma,i}^{-2}\right\}$ the set of undersampled arms. Lemma 6.5 shows that a necessary condition for an error to occur at time n is that there are undersampled arms.

Lemma 6.5. *Under the event $\tilde{\mathcal{E}}_{n,\delta}$, $\hat{\imath}_n \neq \emptyset$ implies that $U_n(n,\delta) \neq \emptyset$.*

Proof. Not recommending $\emptyset$ only happens when the largest empirical mean exceeds γ . Let $\hat{\imath}_n = \arg\max_{i\in[K]} \sqrt{N_{n,i}}(\mu_{n,i} - \gamma)_+$ which satisfies $\mu_{n,\hat{\imath}_n} > \gamma$. Under $\tilde{\mathcal{E}}_{n,\delta}$, we have $\gamma < \mu_{n,\hat{\imath}_n} \leq \mu_{\hat{\imath}_n} + \sqrt{2\tilde{f}_1(n,\delta)/N_{n,\hat{\imath}_n}}$ hence $\hat{\imath}_n \in U_n(n,\delta)$. ■

Lemma 6.6 shows that if there are still undersampled arms at time t , then I_t has not been sampled enough. The proof of Lemma 6.6 is detailed in Appendix F.2.

Lemma 6.6. *Under event $\tilde{\mathcal{E}}_{n,\delta}$, for all $t \leq n$ such that $U_t(n,\delta) \neq \emptyset$, we have $N_{t,I_t} \leq 18\tilde{f}_1(n,\delta)\Delta_{\gamma,I_t}^{-2}$ and $N_{t+1,I_t} = N_{t,I_t} + 1$.*

Lemma 6.7 provides a time after which all arms are sampled enough, hence no error will be made. The proof of Lemma 6.7 is detailed in Appendix F.3.

Lemma 6.7. *Let us define $T_\mu(\delta) = \sup\left\{n \mid n \leq 18H_1(\mu)\tilde{f}_1(n,\delta) + K\right\}$. For all $n > T_\mu(\delta)$, under the event $\tilde{\mathcal{E}}_{n,\delta}$, we have $U_n(n,\delta) = \emptyset$.*

Combining Lemmas 6.7 and 6.5, an inversion formula (Lemma E.5 in Appendix E.8) yields

$$\mathbb{P}_\nu(\hat{\imath}_n \neq \emptyset) \leq K \inf\{\delta \mid n > T_\mu(\delta)\} \leq Ke\sqrt{2}(2+\log n)\sqrt{\frac{n-K}{18H_1(\mu)}}\exp\left(-\frac{n-K}{18H_1(\mu)}\right) .$$

Instances where $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \neq \emptyset$ In that case, we have $\mathcal{E}_\mu^{\mathrm{err}}(n) = \{\hat{\imath}_n \in \{\emptyset\} \cup ([K] \setminus \mathcal{I}_\gamma^{\mathrm{thr}}(\mu))\}$. Let us denote by $U_t(n,\delta) = \left\{i \in [K] \mid N_{n,i} \leq \left(\sqrt{2\tilde{f}_1(n,\delta)\Delta_{\gamma,i}^{-2}} + 1\right)^2\right\}$ the set of undersampled arms. Lemma 6.8 shows that a necessary condition to recommend $\emptyset$ at time n is that all the good arms are undersampled arms, and that a necessary condition to recommend an arm in $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)^{\complement}$ at time n is that this arm is undersampled and will be sampled next.

Lemma 6.8. *Under the event $\tilde{\mathcal{E}}_{n,\delta}$, $\hat{\imath}_n = \emptyset$ implies that $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \subseteq U_n(n,\delta)$, and $\hat{\imath}_n \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)^{\complement}$ implies that $\hat{\imath}_n = I_n \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)^{\complement} \cap U_n(n,\delta)$.*

Proof. Suppose that $\hat{\imath}_n = \emptyset$, hence $\max \mu_{n,i} \leq \gamma$. Then, for all $i \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$, we have $\gamma \geq \mu_{n,i} \geq \mu_i - \sqrt{2\tilde{f}_1(n,\delta)/N_{n,i}}$, hence $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \subseteq U_n(n,\delta)$. Suppose that $\hat{\imath}_n \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)^\complement$, hence $\max_i \mu_{n,i} > \gamma$. Since $\hat{\imath}_n = I_n \in \arg\max_{i\in[K]} \sqrt{N_{n,i}}(\mu_{n,i} - \gamma)_+$, we have $\mu_{n,\hat{\imath}_n} > \gamma$. Then, we have $\gamma < \mu_{n,I_n} \leq \mu_{I_n} + \sqrt{2\tilde{f}_1(n,\delta)/N_{n,I_n}}$, hence $\hat{\imath}_n = I_n \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)^\complement \cap U_n(n,\delta)$. ∎

Lemma 6.9 shows that having all the good arms undersampled implies that the next arm we will pull has not been sampled enough. The proof of Lemma 6.9 is detailed in Appendix F.4.

Lemma 6.9. *Under the event $\tilde{\mathcal{E}}_{n,\delta}$, for all $t \leq n$ such that $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \subseteq U_t(n,\delta)$, we have $N_{t,I_t} \leq D_{I_t}(n,\delta)$ and $N_{t+1,I_t} = N_{t,I_t} + 1$, where $D_i(n,\delta) = \left(\Delta_{\gamma,i}^{-1}\sqrt{2\tilde{f}_1(n,\delta)} + 1\right)^2$ for all $i \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$ and $D_i(n,\delta) = 2\tilde{f}_1(n,\delta)\Delta_{\gamma,i}^{-2}$ for all $i \notin \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$.*

Lemma 6.10 shows that having a good arm which is sampled enough is a sufficient condition to recommend a good arm at time n . The proof of Lemma 6.10 is detailed in Appendix F.5.

Lemma 6.10. *Under the event $\tilde{\mathcal{E}}_{n,\delta}$, $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \cap U_n(n,\delta)^\complement \neq \emptyset$ implies that $\hat{\imath}_n \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$.*

Lemma 6.11 provides a time after which there exists a good arms which is sampled enough, hence no error will be made. The proof of Lemma 6.11 is detailed in Appendix F.6.

Lemma 6.11. *Let us define $S_\mu(\delta) = \sup\left\{n \mid n \leq 4H_1(\mu)\tilde{f}_1(n,\delta) + K + 2|\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)|\right\}$. For all $n > S_\mu(\delta)$, under the event $\tilde{\mathcal{E}}_{n,\delta}$, we have $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \cap U_n(n,\delta)^\complement \neq \emptyset$ and $\hat{\imath}_n \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$.*

Combining Lemmas 6.11 and 6.8, an inversion formula (Lemma E.5 in Appendix E.8) yields

$$\begin{aligned}
&\mathbb{P}_\nu(\{\hat{\imath}_n = \emptyset\} \cup \{\hat{\imath}_n \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)^\complement\}) \leq K \inf\{\delta \mid n > S_\mu(\delta)\} \\
&\quad \leq Ke\sqrt{2}(2 + \log n)\sqrt{\frac{n - K - 2|\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)|}{4H_1(\mu)}} \exp\left(-\frac{n - K - 2|\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)|}{4H_1(\mu)}\right) .
\end{aligned}$$

6.4 Fixed-confidence Guarantees

In some applications, the practitioner has a strict constraint on the confidence δ associated with the candidate answer. This constraint simultaneously supersedes any limitation on the sampling budget and allows early stopping when enough evidence is collected (random since data-dependent). In the fixed-confidence setting, an identification strategy should define a stopping rule in addition of the sampling and recommendation rules.

Stopping rule We couple APGAI with the GLR stopping rule [Garivier and Kaufmann, 2016] for GAI, which coincides with the Box stopping rule introduced in Kaufmann et al. [2018]. At fixed confidence δ , we stop at $\tau^{\text{thr}}_{\gamma,\delta} := \min(\tau_{>,\delta}, \tau_{<,\delta})$ where

$$\begin{aligned} \tau_{>,\delta} &:= \inf\left\{ n \mid \max_{i\in[K]} \sqrt{N_{n,i}}(\mu_{n,i} - \gamma)_+ > \sqrt{2c(n-1,\delta)} \right\} , \\ \tau_{<,\delta} &:= \inf\left\{ n \mid \min_{i\in[K]} \sqrt{N_{n,i}}(\gamma - \mu_{n,i})_+ > \sqrt{2c(n-1,\delta)} \right\} , \end{aligned} \tag{6.3}$$

and $c : \mathbb{N} \times (0,1) \to \mathbb{R}_+$ is a threshold function. Lemma 6.12 gives a threshold ensuring that the GLR stopping rule (6.3) is δ-correct for all $\delta \in (0,1)$, independently of the sampling rule. The proof of Lemma 6.12 is detailed in Appendix F.7.

Lemma 6.12. *Let $\overline{W}_{-1}(x) = -W_{-1}(-e^{-x})$ for all $x \geq 1$, where W_{-1} is the negative branch of the Lambert W function. It satisfies $\overline{W}_{-1}(x) \approx x + \log x$. Let $\delta \in (0,1)$. Given any sampling rule, using the threshold*

$$2c(n,\delta) = \overline{W}_{-1}(2\log(K/\delta) + 4\log\log(e^4 n) + 1/2) \tag{6.4}$$

in the GLR stopping rule (6.3) yields a δ-correct algorithm for 1-sub-Gaussian distributions with mean in $\mathcal{S}$.

Non-asymptotic upper bound Theorem 6.13 gives an upper bound on the expected sample complexity of the resulting algorithm holding for any confidence δ . The proof of Theorem 6.13 is detailed in Appendix F.8.

Theorem 6.13. *Let $\delta \in (0,1)$. Combined with GLR stopping (6.3) using threshold (6.4), the APGAI algorithm is δ-correct and it satisfies that, for all $\nu \in \mathcal{D}^K$ with mean $\mu \in \mathcal{S}$,*

$$\mathbb{E}_\nu[\tau^{\text{thr}}_{\gamma,\delta}] \leq C_\mu(\delta) + K\pi^2/6 + 1 ,$$

where $i_\mu := 1 + (\gamma - 1)\mathbb{1}\left(\mathcal{I}^{\text{thr}}_\gamma(\mu) \neq \emptyset\right)$ and

$$C_\mu(\delta) := \sup\left\{ n \mid n - 1 \leq 2H_{i_\mu}(\mu)\left(\sqrt{c(n-1,\delta)} + \sqrt{3\log n}\right)^2 + D_{i_\mu}(\mu) \right\} ,$$

with $H_1(\mu)$ and $H_\gamma(\mu)$ as in (6.1). $D_1(\mu)$ and $D_\gamma(\mu)$ satisfy $D_1(\mu) \approx D_\gamma(\mu) = \mathcal{O}(H_1(\mu)\log H_1(\mu))$. In the asymptotic regime, we obtain $\limsup_{\delta\to 0} \mathbb{E}_\nu[\tau^{\text{thr}}_{\gamma,\delta}]/\log(1/\delta) \leq 2H_{i_\mu}(\mu)$ since $C_\mu(\delta) =_{\delta\to 0} 2H_{i_\mu}(\mu)\log(1/\delta) + \mathcal{O}(\log\log(1/\delta))$.

Table 6.2 – Asymptotic upper bound $2C(\mu)$ on the expected sample complexity of algorithm $\mathfrak{A}$ on ν , *i.e.* $\limsup_{\delta\to 0} \mathbb{E}_\nu[\tau^{\mathrm{thr}}_{\gamma,\delta}]/\log(1/\delta) \leq 2C(\mu)$. (§) Requires an ordering on the possible answers $[K]\cup\{\emptyset\}$. $H_1(\mu)$ and $H_\gamma(\mu)$ as in (6.1), $\overline{\Delta}_{\gamma,\max} := \max_{i\in\mathcal{I}^{\mathrm{thr}}_\gamma} \Delta_{\gamma,i}$.

Algorithm $\mathfrak{A}$	$\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)=\emptyset$	$\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)\neq\emptyset$
APGAI[Th. 6.13]	$H_1(\mu)$	$H_\gamma(\mu)$
S-TaS [Degenne et al., 2019] §	$H_1(\mu)$	$\overline{\Delta}^{-2}_{\gamma,\max}$
HDoC [Kano et al., 2019]	$H_1(\mu)$	$\overline{\Delta}^{-2}_{\gamma,\max}$
APT-G, LUCB-G [Kano et al., 2019]	$H_1(\mu)$	–

Most importantly, Theorem 6.13 holds for any confidence $\delta\in(0,1)$ and any 1-sub-Gaussian instance ν . Theorem 6.13 shows that $\limsup_{\delta\to 0} \mathbb{E}_\nu[\tau^{\mathrm{thr}}_{\gamma,\delta}]/\log(1/\delta) \leq 2H_{i_\mu}(\mu)$ in the asymptotic regime. This implies that APGAI is asymptotically optimal for Gaussian distributions when $\mathcal{I}^{\mathrm{thr}}_\gamma = \emptyset$. When there are good arms, our upper bound scales as $H_\gamma(\mu)\log(1/\delta)$, which is better than the scaling in $H_1(\mu)\log(1/\delta)$ obtained for the unverifiable sample complexity.

However, when $\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)\neq\emptyset$, our upper bound is sub-optimal compared to $2\min_{i\in[K]}\Delta^{-2}_{\gamma,i}$ (see Lemma 6.1). This sub-optimal scaling stems from the greediness of APGAI when $\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)\neq\emptyset$ since there is no mechanism to detect an arm that is easiest to verify, *i.e.* $\arg\max_{i\in\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)}\Delta_{\gamma,i}$. Empirically, we observe that APGAI can suffer from large outliers when there are good arms with dissimilar gaps, and that adding forced exploration or randomization (with Thompson Sampling) circumvent this issue. Intuitively, a purely asymptotic analysis of APGAI would yield the dependency $2\max_{i\in\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)}\Delta^{-2}_{\gamma,i}$ which is independent from $|\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)|$. Compared to asymptotic results, our non-asymptotic guarantees hold for reasonable values of δ (not necessarily close to 0), with a δ-independent scaling of the order $\mathcal{O}(H_1(\mu)\log H_1(\mu))$.

Comparison with existing upper bounds Table 6.2 summarizes the asymptotic scaling of the upper bound on the expected sample complexity of existing GAI algorithms. While most GAI algorithms have better asymptotic guarantees when $\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)\neq\emptyset$, APGAI is the only one of them which has anytime guarantees on the probability of error (Theorem 6.3). However, we emphasize that APGAI is not the best algorithm to tackle fixed-confidence GAI since it is designed for anytime GAI. Sticky Track-and-Stop (S-TaS) is asymptotically optimal for the "any low arm" problem [Degenne et al., 2019], hence for GAI as well. Even though GAI is one of the few setting where S-TaS admits a computationally tractable implementation, its empirical performance heavily relies on the fixed ordering for the set of possible answers. This partly explains the lack of non-asymptotic guarantees for S-TaS which is asymptotic by nature, while APGAI has non-asymptotic guarantees. For the "bad arm existence" problem, Kaufmann et al. [2018] proves that the empirical proportion $(N_{n,i}/(n-1))_{i\in[K]}$ of Murphy Sampling converges almost surely towards the optimal allocation realizing the asymptotic lower bound of Lemma 6.1. While their result implies that $\lim_{\delta\to 0}\tau^{\mathrm{thr}}_{\gamma,\delta}/\log(1/\delta) = T^{\mathrm{thr}}_\gamma(\nu)$ almost surely, the

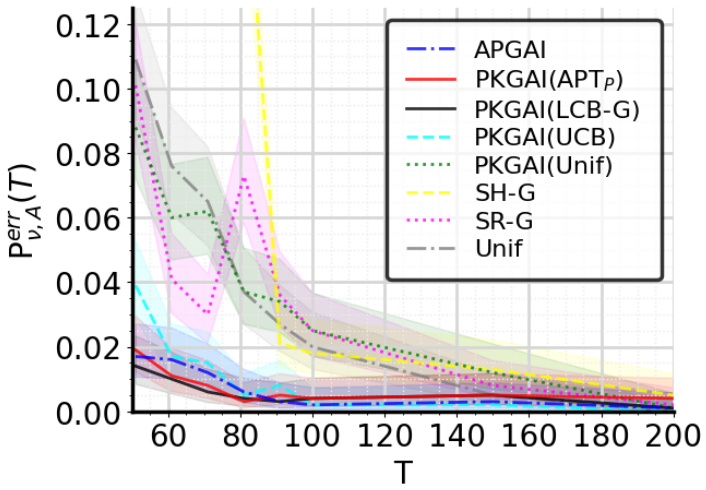

Figure 6.1 – Fixed-budget empirical error on our outcome scoring application.

authors provide no upper bound on the expected sample complexity of Murphy Sampling. Finally, we consider the AllGAI algorithms introduced in Kano et al. [2019] (HDoC, LUCB-G and APT-G) which enjoy theoretical guarantees for some GAI instances as well. When $\mathcal{I}_\gamma^{\text{thr}}(\mu) = \emptyset$, all three algorithms have an upper bound of the form $2H_1(\mu)\log(1/\delta) + \mathcal{O}(\log\log(1/\delta))$. When $\mathcal{I}_\gamma^{\text{thr}}(\mu) \neq \emptyset$, only HDoC admits an upper bound on the expected number of time to return one good arm, which is of the form $2\min_{i\in\mathcal{I}_\gamma^{\text{thr}}(\mu)} \Delta_{\gamma,i}^{-2}\log(1/\delta) + \mathcal{O}(\log\log(1/\delta))$.

The indices used for the elimination and recommendation in BAEC [Tabata et al., 2020] have a dependence in $\mathcal{O}(-\log(\gamma_U - \gamma_L))$, hence BAEC is not defined for GAI where $\gamma_U = \gamma_L$. While it is possible to use UCB/LCB which are agnostic to the gap $\gamma_U - \gamma_L > 0$, these choices have not been studied in Tabata et al. [2020]. Extrapolating the theoretical guarantees of BAEC when $\gamma_L \to \gamma_U$, one would expect an upper bound on its expected sample complexity of the form $2H_1(\mu)\log(1/\delta) + \mathcal{O}((\log(1/\delta))^{2/3})$.

6.5 Experiments

We assess the empirical performance of the APGAI in terms of empirical error, as well as empirical stopping time. Overall, APGAI perform favorably compared to other algorithms in both settings. Moreover, its empirical performance exceeds what its theoretical guarantees would suggest. This discrepancy between theory and practice paves the way for interesting future research. Extensive experiments and implementation details are available in Appendix I of Jourdan and Réda [2023].

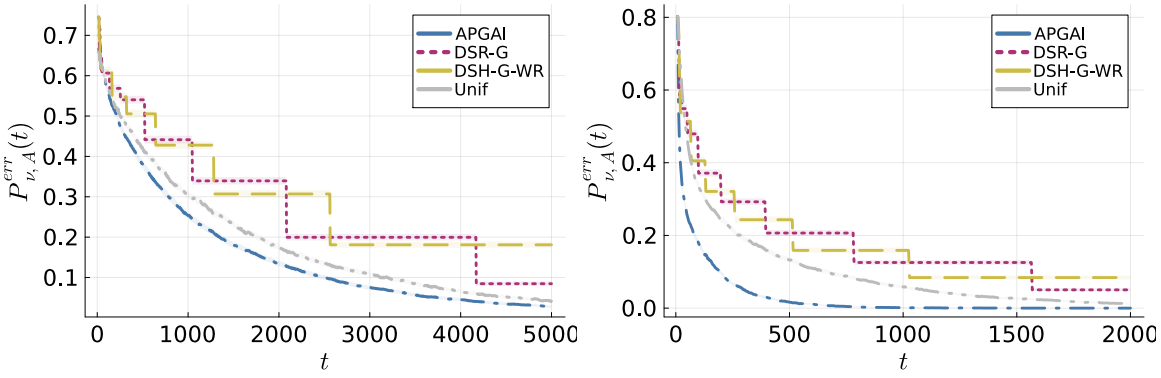

Figure 6.2 – Anytime empirical error on Gaussian instances (a) $\mu \in \{0.55, 0.45\}^{10}$ where $|\mathcal{I}_\gamma^{\text{thr}}(\mu)| = 3$ for $\gamma = 0.5$ and (b) $\mu = -(0.1, 0.4, 0.5, 0.6)$ for $\gamma = 0$.

Outcome scoring application Our real-life motivation is outcome scoring from gene activity (transcriptomic) data. This application is focused on the treatment of encephalopathy of prematurity in infants. The goal is to determine the optimal protocol for the administration of stem cells among $K = 18$ realistic possibilities. Our collaborators tested all treatments, and made RNA-related measurements on treated samples. Computed on 3 technical replicates, the mean value in $[-1, 1]$ corresponds to a cosine score computed between gene activity changes in treated and healthy samples, *i.e.* $\mu = (0.8, 0.791, 0.676, 0.545, 0.538, 0.506, 0.36, 0.329, 0.306, 0.274, 0.241$
,
$0.203, 0.112, 0.084, 0.081, 0.007, -0.018, -0.120)$. When the mean is higher than $\gamma = 0.5$, the treatment is considered significantly positive. Traditional approaches use grid-search with a uniform allocation. We model this application as a Bernoulli instance, *i.e.* observations from arm i are drawn from a Bernoulli distribution with mean $(\mu_i)_+$ (which is $1/2$-sub-Gaussian).

Fixed-budget empirical error The APGAI algorithm is compared to fixed-budget GAI algorithms: SR-G, SH-G, PKGAI and Unif. For a fair comparison, the threshold functions in PKGAI do not use prior knowledge. Several index policies are considered for PKGAI: Unif, APT$_P$, UCB and LCB-G. At time n , the latter selects among the set $\mathcal{S}_n$ of active candidates $I_n \in \arg\max_{i \in \mathcal{S}_n} \sqrt{N_{n,i}}\text{LCB}(n, i)$, where $\text{LCB}(n, i)$ is the lower confidence bound on $\mu_i - \gamma$ at time n . For a budget T up to 200 , our results are averaged over $1,000$ runs, and confidence intervals are displayed. On our outcome scoring application, Figure 6.1 first shows that all uniform samplings (SH-G, SR-G, Unif and PKGAI(Unif)) are less efficient at detecting one of the good arms contrary to the adaptive strategies. Moreover, APGAI actually performs as well as the elimination-based algorithms PKGAI($\star$), while allowing early stopping as well. In Appendix I.3 of Jourdan and Réda [2023], we confirm the good performance of APGAI in terms of fixed-budget empirical error on other instances.

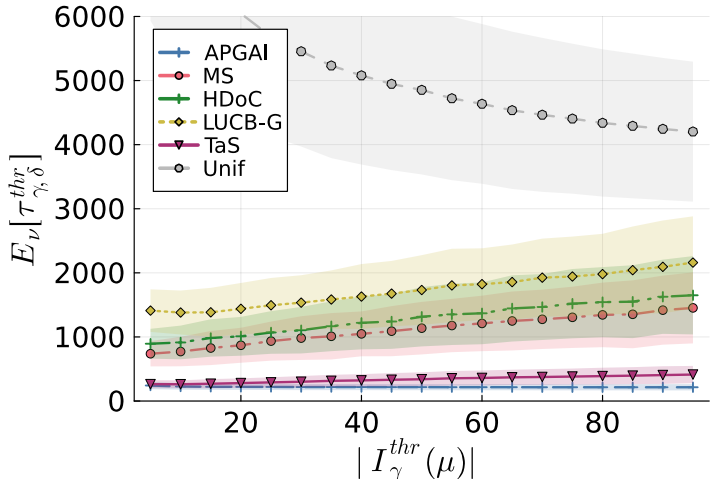

Figure 6.3 – Empirical stopping time ($\delta = 0.01$) on Gaussian instances $\mu \in \{0.5, -0.5\}^{100}$ where $|\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)| \in \{5k\}_{k\in[19]}$ for $\gamma = 0$.

Anytime empirical error The APGAI algorithm is compared to anytime GAI algorithms: DSR-G, DSH-G (see Section 6.3.1) and Unif. Since DSH-G has poor empirical performance, we consider the heuristic DSH-G-WR where each SH instance keeps its history instead of discarding it. On two Gaussian instances ($\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \neq \emptyset$ and $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) = \emptyset$), Figure 6.2 shows that APGAI has significantly smaller empirical error compared to Unif, which is itself better than DSR-G and DSH-G-WR. Our results are averaged over $10,000$ runs, and confidence intervals are displayed. In Appendix I.4 of Jourdan and Réda [2023], we confirm the good performance of APGAI in terms of anytime empirical error on other instances, *e.g.* when $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \neq \emptyset$ and when $|\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)|$ varies. Overall, APGAI appears to have better empirical performance than suggested by Theorem 6.3 when $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \neq \emptyset$.

Empirical stopping time The APGAI algorithm is compared to fixed-confidence GAI algorithms using the GLR stopping rule (6.3) with threshold (6.4) and confidence $\delta = 0.01$: Murphy Sampling (MS [Kaufmann et al., 2018]), HDoC, LUCB-G [Kano et al., 2019], Track-and-Stop for GAI (TaS [Garivier and Kaufmann, 2016]) and Unif. In Figure 6.3, we study the impact of the number of good arms by considering Gaussian instances with two groups of arms. Our results are averaged over $1,000$ runs, and the standard deviations are displayed. Figure 6.3 shows that the empirical performance of APGAI is invariant to varying $|\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)|$, and comparable to the one of TaS. In comparison, the other algorithms have worse performance, and they suffer from increased $|\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)|$ since they have an exploration bonus for each good arm. In contrast, APGAI is greedy enough to only focus its allocation to one of the good arms. While APGAI achieves the best performance when there is no good arm, it can suffer from large outliers when good arms have dissimilar means. To circumvent this problem, it is enough to add forced exploration to APGAI or randomization (with Thompson Sampling). While APGAI was designed for

anytime GAI, it is remarkable that it also has theoretical guarantees in fixed-confidence GAI, and relatively small empirical stopping time.

6.6 Discussion

In Chapter 6, we proposed APGAI, the first anytime and parameter-free sampling strategy for GAI in stochastic bandits, which is independent of a budget T or a confidence δ . In addition to showing its good empirical performance, we also provided guarantees on its probability of error at any deterministic time n (Theorem 6.3) and on its expected sample complexity at any confidence δ when combined with the GLR stopping time (6.3) (Theorem 6.13). As such, APGAI allows both continuation and early stopping. We reviewed and analyzed a large number of baselines for each GAI setting for comparison.

When there are good arms, the optimal allocation in GAI is supported solely on $i^\star(\mu)$. This sparsity of the optimal allocation also appears in BAI when considering the limit of $\Delta_{\min} \to 0$, *i.e.* allocation supported on the arms with the two highest means. Since the arguments in Appendix C of Komiyama et al. [2022] mainly rely on this sparsity, we conjecture that, for any asymptotically optimal GAI algorithm, there exists instances in which the error probability cannot decay exponentially with the horizon. Given that ε-BAI has dense optimal allocation (even asymptotically for $\Delta_{\min} \to 0$), their argument does not apply, hence we could provide strong anytime guarantees for the EB-TC$_{\varepsilon_0}$ algorithm (Chapter 5). However, in GAI, we do not believe that an algorithm could have such strong anytime guarantees. This explains why the APGAI algorithm appears to be less satisfactory in tackling GAI.

While we considered unstructured multi-armed bandits, many applications have a known structure. Investigating the GAI problem on *e.g.* linear or infinitely-armed bandits, would be an interesting subsequent work. In particular, working in a structured framework when facing a possibly infinite number of arms would bring out more compelling questions about how to explore the arm space in a both tractable and meaningful way. While linear bandits is the topic of Part III, we will discuss ε-BAI instead of GAI.

Part III

Epsilon Best Arm Identification in Linear Bandits

Chapter 7

Choosing the Furthest Answer

In Chapter 7, we study the ε-BAI problem for linear bandit in the fixed-confidence setting, as described in Chapter 1 and studied in Chapter 5 for the vanilla setting. The presented results were published in Jourdan and Degenne [2022].

While best-arm identification for linear bandits has been extensively studied in recent years, few works have been dedicated to identifying one arm that is ε-close to the best one (and not exactly the best one). In this problem with several correct answers, an identification algorithm should focus on one candidate among those answers and verify that it is correct. We demonstrate that picking the answer with highest mean does not allow an algorithm to reach asymptotic optimality in terms of expected sample complexity. Instead, a *furthest answer* should be identified. Using that insight to choose the candidate answer carefully, we develop a simple procedure to adapt best-arm identification algorithms to tackle ε-best-answer identification in transductive linear stochastic bandits. Finally, we propose an asymptotically optimal algorithm for this setting, which is shown to achieve competitive empirical performance against existing modified best-arm identification algorithms.

Contents

7.1 Introduction

Since ε-BAI for linear bandits is the focus of Part III, we recall the problem as described in Chapter 1. In the linear bandit problem, each arm $i \in [K]$ is associated with a known context vector $a_i \in \mathbb{R}^d$ and has a mean which is a linear function of unknown regression parameter $\theta \in \mathcal{M}$ where $\mathcal{M} \subseteq \mathbb{R}^d$ is a bounded set, *i.e.* $\mu_i = \langle a_i, \theta \rangle$. In the transductive setting [Fiez et al., 2019], each answer $j \in [Z]$ is associated with a known context vector $z_j \in \mathbb{R}^d$ and the mean is also linear, *i.e.* $\mu_j = \langle z_j, \theta \rangle$. Taking $\mathcal{Z} = \mathcal{A}$ recover the linear bandits setting. We assume that $\mathcal{A}$ spans $\mathbb{R}^d$ and by denote $L_{\mathcal{A}}$ (resp. $L_{\mathcal{Z}}$ and $L_{\mathcal{M}}$) is the maximum ℓ_2-norm of vectors in $\mathcal{A}$ (resp. $\mathcal{Z}$ and $\mathcal{M}$), *i.e.* $L_X := \max_{x \in X} \|x\|_2$ where $\|\cdot\|_2$ is the ℓ_2-norm. Note that μ is fully characterized by the regression parameter θ , the set of arms vector $\mathcal{A} = \{a_i\}_{i \in [K]}$ and the set of answers vector $\mathcal{Z} = \{z_j\}_{j \in [Z]}$. Moreover, we will use their context vectors to refer to arms $a \in \mathcal{A}$ and answers $z \in \mathcal{Z}$ instead of using their integer indices. The focus of this chapter is to highlight a phenomenon which is orthogonal to the distribution, therefore we restrict ourselves to the set $\mathcal{D}_{\mathcal{N}_\sigma}$ of Gaussian distributions with known variance and assume that $\sigma_a^2 = 1$ for all $a \in \mathcal{A}$ by scaling, hence $\mathcal{D}^K = \mathcal{D}_1^K$. Let $\nu \in \mathcal{D}^K$ with regression parameter $\theta \in \mathcal{M}$.

At time n , we denote by $I_n \in \mathcal{A}$ the next arm to pull and by $\hat{z}_n$ the candidate answer. Conditioned on I_n , the observation X_{n,I_n} satisfies that $X_{n,I_n} \sim \mathcal{N}(\langle I_n, \theta \rangle, 1)$. Recall that the history is defined as the σ-algebra $\mathcal{F}_n := \sigma\left(U_1, I_1, X_{1,I_1}, \cdots, I_{n-1}, X_{n-1,I_{n-1}}, U_n\right)$, where $U_n \sim \mathcal{U}([0,1])$ materializes the possible independent randomness used by the algorithm at time n . However, in this chapter, we present a deterministic algorithm.

In the ε-BAI for transductive linear bandits, the agent aims at identifying an answer whose mean is ε-close to the highest one, *i.e.* $z \in \mathcal{Z}_\varepsilon(\theta) := \{z \in \mathcal{Z} \mid \langle \theta, z \rangle \geq \max_{z \in \mathcal{Z}} \langle \theta, z \rangle - \varepsilon\}$ where $\varepsilon \geq 0$. In the multiplicative ε-BAI problem, the means are non-negative and one aims at returning an answer $z \in \mathcal{Z}_\varepsilon^{\text{mul}}(\theta) := \{z \in \mathcal{Z} \mid \langle \theta, z \rangle \geq (1-\varepsilon) \max_{z \in \mathcal{Z}} \langle \theta, z \rangle\}$ where $\varepsilon \in [0, 1)$. The set of *greedy* answers is defined as the set of answers with highest mean, *i.e.* $z^\star(\theta) := \arg\max_{z \in \mathcal{Z}} \langle \theta, z \rangle$. While most of our contributions (and statements) will hold for both the additive and the multiplicative settings, we note that the additive ε-BAI has received more scrutiny [Garivier and Kaufmann, 2021, Kocák and Garivier, 2021]. Taking $\varepsilon = 0$ recovers the BAI problem, in which there is a unique correct answer. As we extensively mentioned in previous chapters, ε-BAI is often seen as a more practical objective than BAI in cases where getting an answer close to optimal is enough. While a BAI algorithm will spend many samples distinguishing between the answer with highest mean and an ε-close one, an ε-BAI algorithm will be able to stop quickly.

We consider the fixed-confidence setting (see Section 1.4). Let $\delta \in (0,1)$ be given to the agent, and $\tau_{\varepsilon,\delta}$ denote the stopping time. A strategy is said to be (ε, δ)-PAC if, for all $\nu \in \mathcal{D}^K$ with regression parameter $\theta \in \mathcal{M}$, $\mathbb{P}_\nu\left(\tau_{\varepsilon,\delta} < +\infty, \hat{z}_{\tau_{\varepsilon,\delta}} \notin \mathcal{Z}_\varepsilon(\theta)\right) \leq \delta$. Among the class of (ε, δ)-PAC algorithms, our goal is to minimize the expected sample complexity $\mathbb{E}_\nu[\tau_{\varepsilon,\delta}]$.

Contribution 7.1. *The contributions of Chapter 7 are the following.*

- *We provide an analysis of ε-BAI for transductive linear bandits and highlight a phenomenon which was overlooked by previous work. The choice of the candidate answer is crucial to reach asymptotic optimality in terms of expected sample complexity and one should identify the furthest answer instead of using the greedy answers.*
- *By carefully choosing the candidate answer and leaving the sampling rule unchanged, we develop a simple procedure to adapt BAI algorithms to be (ε, δ)-PAC and empirically competitive for ε-BAI in transductive linear stochastic bandits.*
- *By leveraging the concept of furthest answer in the sampling rule, we propose an asymptotically optimal algorithm which has competitive empirical performance.*

7.2 Comparing Correct Answers

7.2.1 Lower Bound

For any $w \in (\mathbb{R}^+)^K$, we define the design matrix $V_w := \sum_{a \in \mathcal{A}} w_a a a^\intercal \in \mathbb{R}^{d \times d}$, which is symmetric and positive semi-definite, and definite if and only if $\mathrm{Span}(\{a \in \mathcal{A} \mid w_a \neq 0\}) = \mathbb{R}^d$. For any symmetric positive semi-definite matrix $V \in \mathbb{R}^{d \times d}$, we define the semi-norm $\|x\|_V := \sqrt{x^\intercal V x}$ for $x \in \mathbb{R}^d$, which is a norm if V is positive definite.

Alternative set Given an answer $z \in \mathcal{Z}$, the *alternative to* z is defined as the set of parameters $\lambda \in \mathcal{M}$ for which z is not a correct answer for λ , *i.e.* $\neg_\varepsilon z = \overline{\{\lambda \in \mathcal{M} \mid \langle \lambda, z \rangle < \max_{z \in \mathcal{Z}} \langle \lambda, z \rangle - \varepsilon\}}$. For multiplicative ε-BAI, we have $\neg_\varepsilon^{\mathrm{mul}} z = \overline{\{\lambda \in \mathcal{M} \mid \langle \lambda, z \rangle < (1-\varepsilon) \max_{z \in \mathcal{Z}} \langle \lambda, z \rangle\}}$.

Asymptotic lower bound Lemma 7.2 gives an asymptotic lower bound on the expected sample complexity of any (ε, δ)-PAC strategy for both additive and multiplicative ε-BAI. This is a corollary of Theorem 1 in Degenne and Koolen [2019], which holds for any multiple answer instance and σ-sub-Gaussian distributions.

Lemma 7.2 (Theorem 1 in Degenne and Koolen [2019])**.** *Let $\delta \in (0,1)$ and $\varepsilon \in \mathbb{R}$. For all (ε, δ)-PAC strategy, for all $\nu \in \mathcal{D}^K$ with regression parameter $\theta \in \mathcal{M}$, $\liminf_{\delta \to 0} \mathbb{E}_\nu[\tau_{\varepsilon,\delta}] / \log(1/\delta) \geq T_\varepsilon(\nu)$ with $T_\varepsilon(\nu) = \min_{z \in \mathcal{Z}_\varepsilon(\theta)} T_\varepsilon(\nu, z)$ and $T_\varepsilon(\nu, z)^{-1} := \max_{w \in \Sigma_K} \inf_{\lambda \in \neg_\varepsilon z} \frac{1}{2} \|\theta - \lambda\|^2_{V_w}$.*

An (ε, δ)-PAC algorithm is said to be *asymptotically optimal* if the bound is tight: for all $\nu \in \mathcal{D}^K$ with regression parameter $\theta \in \mathcal{M}$, $\limsup_{\delta \to 0} \mathbb{E}_\nu[\tau_{\varepsilon,\delta}] / \log(1/\delta) \leq T_\varepsilon(\nu)$. We refer the reader to Section 1.4.1 for more details on lower bounds on the expected sample complexity.

As noted by Chernoff [1959], the complexity $T_\varepsilon(\nu)^{-1}$ is the value of a zero-sum game between two players. The agent chooses a correct answer and a pulling proportion over arms, $(z, w) \in \mathcal{Z}_\varepsilon(\theta) \times \Sigma_K$. The nature plays the most confusing alternative $\lambda \in \neg_\varepsilon z$ with respect to a reweighted Kullback-Leibler divergence ($\|\cdot\|^2_{V_w}$ for Gaussian) in order to fool the agent into rejecting this answer. Our algorithm, named LεBAI (Linear ε-BAI), is based on this formulation. Even for known θ , computing $T_\varepsilon(\nu)^{-1}$ is in general intractable due to the non-convexity of $\neg_\varepsilon z$ and the additional maximization over $\mathcal{Z}_\varepsilon(\theta)$. When ε is large enough to have $\mathcal{Z}_\varepsilon(\theta) = \mathcal{Z}$ for all $\theta \in \mathcal{M}$, then $T_\varepsilon(\nu) = 0$, *i.e.* it is *so easy* that no sample is needed.

Since $T_\varepsilon(\nu) \leq T_0(\nu)$ for all $\nu \in \mathcal{D}^K$ such that $\theta \in \mathcal{M}$ (because $\neg_\varepsilon z \subseteq \neg_0 z$), ε-BAI is easier than BAI. There exists arbitrarily hard BAI instances that can be solved if seen as an ε-BAI problem, *e.g.* when the gap between the best and the second best arm is arbitrarily small.

7.2.2 Furthest Answer

The contributions in this chapter are linked with the concept of furthest answers: it should be leveraged in the recommendation-stopping pair (see Section 7.3) and in the sampling rule (see Section 7.4). In a nutshell, to reach asymptotic optimality in terms of sample complexity one should identify the unique furthest answer instead of simply using the greedy answers: *all correct answers are not equivalent*.

The set $z_F(\nu)$ of furthest (or easiest-to-verify) answer is defined as the answers which maximizes the dissimilarity involved in the definition $T_\varepsilon(\nu)^{-1}$, when using an optimal allocation over arms $w_F(\nu) \in \Sigma_K$. Introduced in Degenne and Koolen [2019] and Garivier and Kaufmann [2021], it is defined as

$$(z_F(\nu), w_F(\nu)) := \operatorname*{arg\,max}_{(z,w) \in \mathcal{Z}_\varepsilon(\theta) \times \Sigma_K} \inf_{\lambda \in \neg_\varepsilon z} \frac{1}{2} \|\theta - \lambda\|^2_{V_w} \ . \tag{7.1}$$

Both $z_F(\nu)$ and $z^\star(\theta)$ are subsets of $\mathcal{Z}_\varepsilon(\theta)$, however they might differ. In BAI with a unique best arm the set $\mathcal{Z}_\varepsilon(\theta)$ is a singleton, hence those two notions coincide.

We assume there is a unique furthest answer for the unknown θ , *i.e.* $|z_F(\nu)| = 1$. When $|z_F(\nu)| > 1$, some function of θ has to have exactly the same value for all answers of the set. This happens with probability 0 if θ arises from an absolutely continuous distribution. Almost all BAI algorithms make the assumption that $|z^\star(\theta)| = 1$, which implies $|z_F(\nu)| = 1$ in the BAI case. Since the furthest answer is assumed unique, we abuse notation and denote by $z_F(\nu)$

both that answer, and the singleton containing it as in (7.1). $z^\star(\theta)$ denotes a set as we do not assume that $|z^\star(\theta)| = 1$. The dependence of $z_F(\nu)$ and $w_F(\nu)$ on ε is omitted.

Asymptotic sub-optimality of $z^\star(\theta)$ An (ε, δ)-PAC strategy is said to be *asymptotically greedy* if the only correct answers for which the algorithm will stop asymptotically are the greedy answers $z^\star(\theta)$, *i.e.* for all $\nu \in \mathcal{D}^K$ with regression parameter $\theta \in \mathcal{M}$,

$$\lim_{\delta \to 0} \mathbb{P}_\nu \left(\tau_{\varepsilon,\delta} < +\infty, \hat{z}_{\tau_{\varepsilon,\delta}} \in \mathcal{Z}_\varepsilon(\theta) \setminus z^\star(\theta) \right) = 0 \,. \tag{7.2}$$

Lemma 7.3 shows that any asymptotically greedy (ε, δ)-PAC strategy is asymptotically sub-optimal whenever $z_F(\nu) \notin z^\star(\theta)$, *i.e.* it can only reach $T_{g,\varepsilon}(\nu)$ which is strictly higher than $T_\varepsilon(\nu)$. The proof of Lemma 7.3 is detailed in Appendix G.1.

Lemma 7.3. *For all asymptotically greedy (ε, δ)-PAC strategy, for all $\nu \in \mathcal{D}^K$ such that $\theta \in \mathcal{M}$,*

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau_{\varepsilon,\delta}]}{\log(1/\delta)} \geq T_{g,\varepsilon}(\nu) \quad \textit{with} \quad T_\varepsilon(\nu) = \min_{z \in z^\star(\theta)} T_\varepsilon(\nu, z) \,,$$

where $T_{g,\varepsilon}(\nu) > T_\varepsilon(\nu)$ if and only if $z_F(\nu) \notin z^\star(\theta)$.

Lemma 7.4 shows that any (ε, δ)-PAC strategy recommending any greedy answers $\hat{z}_{\tau_{\varepsilon,\delta}} \in z^\star(\theta_{\tau_{\varepsilon,\delta}})$ which succeeds in identifying $z^\star(\theta)$ is asymptotically greedy. Since $\theta \mapsto z^\star(\theta)$ is continuous and $\mathcal{Z}$ is finite, it is sufficient to have a sampling rule ensuring that $\lim_{n \to +\infty} \theta_n = \theta$.

Lemma 7.4. *Any (ε, δ)-PAC strategy recommending $\hat{z}_{\tau_{\varepsilon,\delta}} \in z^\star(\theta_{\tau_{\varepsilon,\delta}})$ is asymptotically greedy if the sampling rule ensures that $\lim_{\delta \to 0} \mathbb{P}_\nu(\tau_{\varepsilon,\delta} < +\infty, z^\star(\theta_{\tau_{\varepsilon,\delta}}) = z^\star(\theta)) = 1$.*

Asymptotic optimality of $z_F(\nu)$ The furthest answer has by definition a central role in the characteristic time. Among the oracles that first choose an answer $z \in \mathcal{Z}_\varepsilon(\theta)$ and then sample according to the optimal proportions to verify that $\theta \notin \neg_\varepsilon z$, the only one achieving asymptotic optimality is the one picking $z_F(\nu)$. To be asymptotically optimal, an ε-BAI algorithm has to implicitly identify $z_F(\nu)$.

The definition of $z_F(\nu)$ comes from an asymptotic lower bound, and no finite time lower bounds are available for ε-BAI. It could be that for larger δ (hence small stopping times), identifying $z_F(\nu)$ among $\mathcal{Z}_\varepsilon(\theta)$ is too costly to be done before stopping. In that regime, it could be that an algorithm cannot do better than picking any ε-optimal answer. Strong moderate

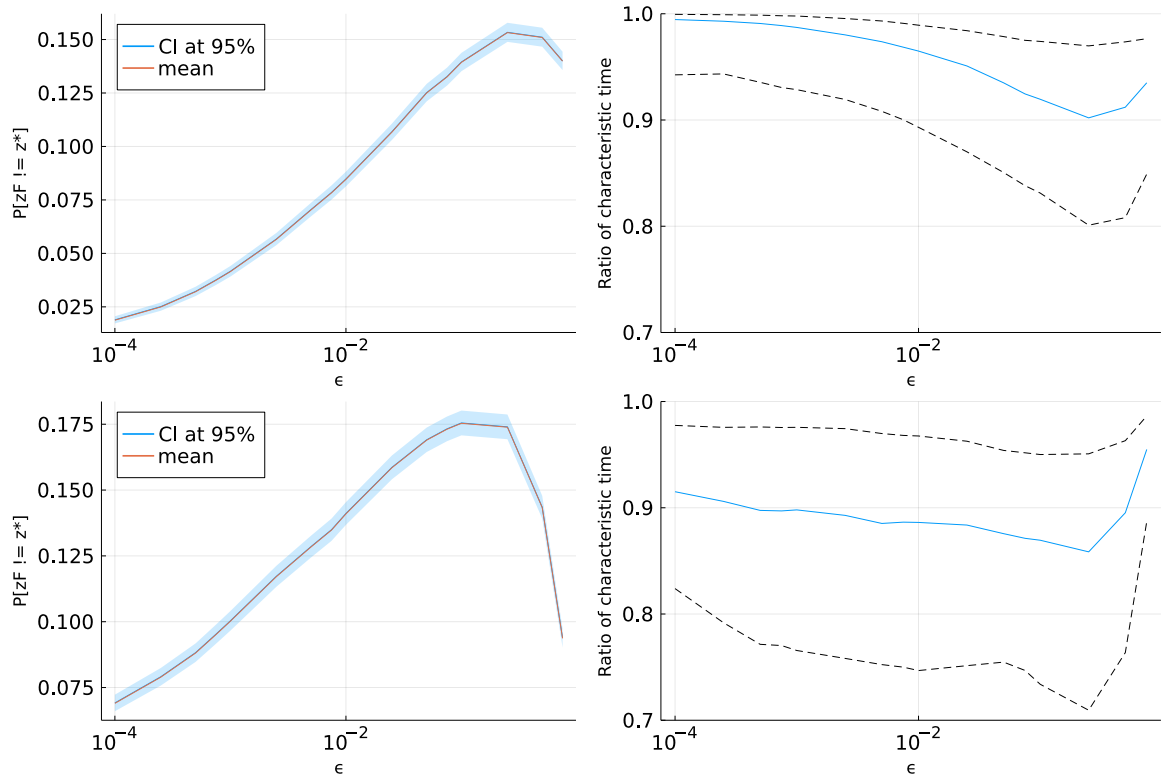

Figure 7.1 – Influence of ε on (left) the proportion of draws where $z_F(\nu) \notin z^\star(\theta)$, (right) the median (and first/third quartile) of (top) $T_\varepsilon(\nu)/T_{g,\varepsilon}(\nu)$ and (bottom) $T_\varepsilon^{\mathrm{mul}}(\nu)/T_{g,\varepsilon}^{\mathrm{mul}}(\nu)$, when $z_F(\nu) \notin z^\star(\theta)$.

confidence terms (independent of δ) affecting the sample complexity have been shown in different settings [Katz-Samuels and Jamieson, 2020, Mason et al., 2020].

Numerical simulations We compare the furthest and the greedy answers for additive and multiplicative ε-BAI. We consider $d = 2$, $\mathcal{M} = \mathbb{R}^2$ and $\mathcal{Z} = \mathcal{A}$ with $K = 4$. We use $\theta = (1, 0)$ and generate 25000 random instances. In each one of them, we consider $z_1 = \theta$ and draw uniformly at random $z_2 \in \{(\cos(\phi), \sin(\phi)) \mid \phi \in [-\phi_\varepsilon, \phi_\varepsilon]\}$ and $z_3, z_4 \in \{(\cos(\phi), \sin(\phi)) \mid \phi \in (-\pi, -\phi_\varepsilon) \cup (\phi_\varepsilon, \pi]\}$, where $\phi_\varepsilon := \arccos(1 - \varepsilon)$. This yields $z_1 = z^\star(\theta)$, $z_2 \in \mathcal{Z}_\varepsilon(\theta)$ and $z_3, z_4 \in \mathcal{Z} \setminus \mathcal{Z}_\varepsilon(\theta)$. To approximate $(T_\varepsilon(\nu), z_F(\nu))$, we discretize $\triangle_4$ with 10000 vectors. This is repeated for several values of ε . We never observed $|z^\star(\theta)| > 1$ or $|z_F(\nu)| > 1$.

Figure 7.1 reveals that the proportion of draws where $z^\star(\theta) \neq z_F(\nu)$ is not negligible. On those instances, Figure 7.1(b) shows that $T_\varepsilon(\nu)/T_{g,\varepsilon}(\nu)$ and $T_\varepsilon^{\mathrm{mul}}(\nu)/T_{g,\varepsilon}^{\mathrm{mul}}(\nu)$ is on average 0.95 and 0.9 . Therefore, when they are different, the furthest answer has a 5% and 10% lower characteristic time than greedy answers.

7.3 From BAI to ε-BAI Algorithms

We propose a simple procedure to convert any BAI algorithm into an (ε, δ)-PAC algorithm. While leaving the original sampling rule unchanged, the stopping-recommendation rule are carefully chosen thanks to the concept of furthest answer.

Structure Since ε-BAI is easier than BAI, the stopping rule of BAI algorithms has to be modified for ε-BAI. Instead of stopping whenever a single best arm is identified, it is enough to stop when we know that an answer is ε-close to the ones with highest mean. In most (ε-)BAI algorithms, the stopping-recommendation pair and the sampling rule can be thought as two independent blocks. There exists stopping-recommendation pairs that guarantee the strategy to be (ε, δ)-PAC regardless of the sampling rule (*e.g.* see Lemma 7.5). Therefore, we can take the sampling rule from a BAI algorithm and couple it with a stopping-recommendation pair with this property.

We will now describe such a stopping-recommendation pair for ε-BAI in transductive linear Gaussian bandits. Due to its generality, this procedure can be readily adapted to tackle general distributions (*e.g.* σ-sub-Gaussian) and different structures (*e.g.* spectral bandits) by simply adapting the stopping rule and its associated threshold.

7.3.1 Recommendation and Stopping Rules

Estimator Let N_n be the empirical allocation before time n , *i.e.* $N_{n,a} = \sum_{t\in[n-1]} \mathbb{1}(I_t = a)$. We denote the Ordinary Least Square (OLS) estimator by $\theta_n = V_{N_n}^{-1} \sum_{t\in[n-1]} X_{t,I_t} I_t$. When $\theta_n \in \mathcal{M}$, this is also the Maximum Likelihood Estimator (MLE). We denote by ν_n a bandit instance with regression parameter θ_n .

GLR stopping rule As detailed in Section 1.4.2, we adopt the GLR stopping rule. Given a candidate answer $\hat{z}_n \in \mathcal{Z}_\varepsilon(\theta_n)$, the algorithm stops as soon as

$$\inf_{\lambda\in\neg_\varepsilon \hat{z}_n} \|\theta_n - \lambda\|^2_{V_{N_n}} > 2c(n-1,\delta) \,. \tag{7.3}$$

In Lemma 7.5, we show that combining a recommendation rule such that $\hat{z}_n \in \mathcal{Z}_\varepsilon(\theta_n)$ and this stopping rule is sufficient to obtain a (ε, δ)-PAC strategy regardless of the sampling rule. This holds even when the stopping criterion is checked only on an infinite subset of $\mathbb{N}$. The proof leverages the concentration inequalities of Kaufmann and Koolen [2021].

Lemma 7.5. *Let $\mathcal{T} \subseteq \mathbb{N}$ with $|\mathcal{T}| = \infty$. Given any sampling and recommendation rule such that $\hat{z}_n \in \mathcal{Z}_\varepsilon(\theta_n)$ for all $n \in \mathcal{T}$, then evaluating the stopping criterion (7.3) at each time $n \in \mathcal{T}$ with the threshold*

$$c(n, \delta) = 2K \log\left(4 + \log\left(n/K\right)\right) + K\mathcal{C}_G\left(\log\left(1/\delta\right)/K\right) \tag{7.4}$$

yields an (ε, δ)-PAC strategy for linear Gaussian distributions with unit variance and regression parameter in $\mathcal{M}$. The function $\mathcal{C}_G$ is defined in (B.1). It satisfies $\mathcal{C}_G(x) \approx x + \log(x)$.

Proof. The proof is the same as the one detailed in Appendix B.1. ■

Since this result holds for any sampling and recommendation rules satisfying one mild requirement, $\hat{z}_n \in \mathcal{Z}_\varepsilon(\theta_n)$ for all $n \in \mathcal{T}$, this leaves open the question on how to design those two rules to stop as early as possible. Algorithms that are agnostic to the choice of the candidate answer might have a higher expected sample complexity than the ones aiming at identifying the furthest answer.

Recommendation rule Taking a greedy answer $\hat{z}_n \in z^\star(\theta_n)$ is a direct choice. Thanks to its efficient implementation, using a greedy answer is the only computationally feasible recommendation rule for combinatorial or continuous answers sets. Unfortunately, when $z_F(\nu) \notin z^\star(\theta)$, this approach leads to sub-optimal algorithms in terms of asymptotic sample complexity (Lemmas 7.3 and 7.4).

When Z is not too large or when we disregard the computational cost, a more careful choice than the greedy one alleviates this sub-optimality. The set of correct answers for which the GLR (*l.h.s.* of (7.3)) is maximized are the *instantaneous furthest answers*

$$z_F(\nu_n, N_n) := \operatorname*{arg\,max}_{z \in \mathcal{Z}_\varepsilon(\theta_n)} \inf_{\lambda \in \neg_\varepsilon z} \|\theta_n - \lambda\|^2_{V_{N_n}} . \tag{7.5}$$

By definition, $z_F(\nu_n, N_n)$ are the correct answers for which we have the most evidence that they are correct at time n . At a lower computational cost than using a furthest answer for the current estimator $\hat{z}_n \in z_F(\nu_n)$, we will see that using an instantaneous furthest answer enjoys similar empirical performance (sample complexity). For all the above sets of candidate answers, the ties are broken arbitrarily. Empirically, we only observed singletons.

Dependence in K In linear bandits, when K is large, dependencies in K can be replaced by d [Lattimore and Szepesvari, 2019]. The focus of our work is to highlight the importance of carefully choosing answers, therefore having K instead of d is a price we are willing to pay for

simpler arguments. Prior works removed the K dependency in the analysis of game-based algorithms [Degenne et al., 2020a, Tirinzoni et al., 2020, Réda et al., 2021].

7.3.2 Modified BAI Algorithms

Modification procedure Given any BAI algorithm for transductive linear Gaussian bandits, we modify it to use (7.3) as stopping rule while leaving the sampling rule unchanged. By Lemma 7.5, the resulting algorithm is an (ε, δ)-PAC strategy. For the recommendation rule, theory (Lemmas 7.3-7.4) and experiments (Figure 7.2) both suggest to use $\hat{z}_n \in z_F(\nu_n, N_n)$ instead of $\hat{z}_n \in z^\star(\theta_n)$. We do not prove any theoretical guarantees on the sample complexity of the modified algorithms since such results depend heavily on each sampling rule.

BAI benchmarks Lots of algorithms have been designed to tackle the BAI setting and we mention below the ones used in the experiments as benchmarks. Soare et al. [2014] proposed a static allocation design $\mathcal{XY}$-Static and its elimination-based improvement $\mathcal{XY}$-Adaptive, which are linked to a G-optimal design. In Xu et al. [2018], LinGapE was introduced as the first gap-based BAI algorithm. All the above BAI algorithms are not shown to be asymptotically optimal and depend on δ (except $\mathcal{XY}$-Static). Algorithm such as DKM [Degenne et al., 2019] and LinGame [Degenne et al., 2020a] are asymptotically optimal and their sampling rule does not depend on δ .

Other stopping rules For all BAI algorithm using a GLR stopping rule, the stopping condition (7.3) is a natural extension. Some other stopping rules (which are not based on the GLR) have a direct extension to ε-BAI. This is the case for the gap-based stopping rule for additive ε-BAI employed by LinGapE, where we can stop when the gap is smaller than ε instead of stopping when it is negative.

7.3.3 Experiments

We perform experiments to highlight the empirical performance of the modified BAI algorithms on additive ε-BAI problems. Moreover, we show that using $\hat{z}_n \in z_F(\nu_n, N_n)$ in (7.3) achieves lower empirical stopping time compared to $\hat{z}_n \in z^\star(\theta_n)$, and outperforms the ε-gap stopping rule with $\hat{z}_n \in z^\star(\theta_n)$. We consider linear bandits ($\mathcal{A} = \mathcal{Z}$), $\mathcal{M} = \mathbb{R}^d$ and $(\varepsilon, \delta) = (0.05, 0.01)$, and perform 5000 runs. The stopping-recommendation pair is updated at each time n .

Hard instances We adapt the usual hard instance studied in BAI for linear bandits to enforce the existence of multiple correct answers, *i.e.* $|\mathcal{Z}_\varepsilon(\theta)| > 1$. Taking $\theta = e_1$ with $e_i = (\mathbb{1}(j = i))_{j\in[d]}$, the answers set is defined as $\mathcal{Z} = \{e_1, \cdots, e_d, a_{d+1}, a_{d+2}\}$ where $a_{d+1} =$

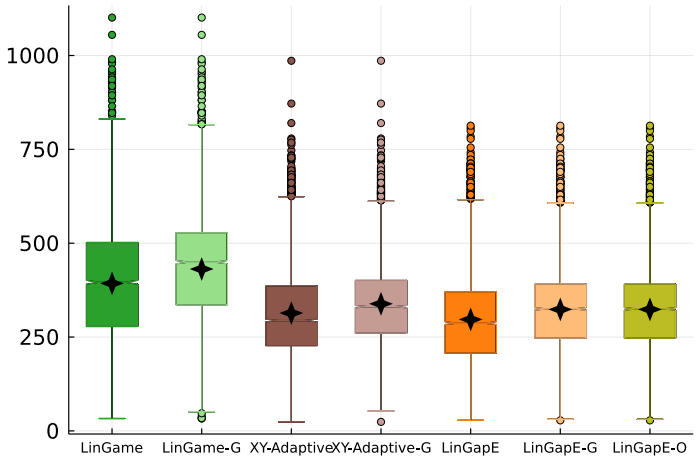

Figure 7.2 – Empirical stopping time of the modified BAI algorithms with $\hat{z}_n \in z_F(\nu_n, N_n)$ on the hard instance (star is mean). "-G" denotes $\hat{z}_n \in z^\star(\theta_n)$. "-O" denotes the ε-gap stopping rule for $\hat{z}_n \in z^\star(\theta_n)$.

$\cos(\phi_1)e_1 + \sin(\phi_1)e_2 \in \mathcal{Z}_\varepsilon(\theta)$ and $a_{d+2} = \cos(\phi_2)e_1 + \sin(\phi_2)e_2 \notin \mathcal{Z}_\varepsilon(\theta)$. Considering $d = 2$, we use $\phi_1 = r_\varepsilon\theta_\varepsilon$ and $\phi_2 = (1 + r_\varepsilon)\theta_\varepsilon$ with $\theta_\varepsilon = \arccos(1 - \varepsilon)$ and $r_\varepsilon = 0.1$.

On this instance, the BAI algorithms without modification require on average 545 times more samples than compared to their modified version. The discrepancy is particularly striking since the hard instance for ε-BAI is even harder for BAI.

Figure 7.2 reveals that, for all modified BAI, considering an instantaneous furthest answer instead of a greedy answer leads to lower empirical stopping time. Their ratio is 0.92 on average. This matches the asymptotic observations when computing $T_\varepsilon(\nu)/T_{g,\varepsilon}(\nu)$ for additive ε-BAI (as done in Figure 7.1 for multiplicative ε-BAI, see Jourdan and Degenne [2022]). The modified LinGapE using $\hat{z}_n \in z_F(\nu_n, N_n)$ outperforms the ε-gap extension of the original stopping rule, which is equivalent to using (7.3) with $\hat{z}_n \in z^\star(\theta_n)$. While guided by the asymptotic regime, using $z_F(\nu_n, N_n)$ instead of $z^\star(\theta_n)$ for the stopping-recommendation pair has practical utility in the moderate confidence regime with a 10% speed-up in terms of sample complexity.

7.4 LεBAI Algorithm

Leveraging the concept of furthest answer in the sampling rule, we present LεBAI (Linear ε-BAI), an asymptotically optimal algorithm for (ε, δ)-PAC best-answer identification in transductive linear bandits. It deals with both the multiplicative and the additive ε-BAI problems. Similarly to works on linear bandits [Abbasi-Yadkori et al., 2011, Soare et al., 2014], we assume that the set of parameters is bounded, *i.e.* $L_\mathcal{M} < +\infty$.

Input: Learner $\mathcal{L}^{\mathcal{A}}$ and $\mathcal{Z}$-oracle $\mathcal{L}^{\mathcal{Z}}$.
Pull once each arm $a \in \mathcal{A}$, set $n_0 = K$ and $W_{n_0+1} = 1_K$;
for $n > n_0$ **do**
  Get $\hat{z}_n \in z_F(\mu_n, N_n)$; // Candidate answer
  If $\inf_{\lambda \in \neg_\varepsilon \hat{z}_n} \|\theta_n - \lambda\|^2_{V_{N_n}} > 2c(n-1, \delta)$ **then return** $\hat{z}_n$; // GLR stopping rule
  Get $\left(\tilde{z}_n, w_n^{\mathcal{L}^{\mathcal{A}}}\right)$ from $\mathcal{L}^{\mathcal{Z}} \times \mathcal{L}^{\mathcal{A}}$; // Learner plays
  Set $w_n = \frac{1}{nK} 1_K + \left(1 - \frac{1}{n}\right) w_n^{\mathcal{L}^{\mathcal{A}}}$ and $W_{n+1} = W_n + w_n$; // Forced exploration
  Set $\lambda_n \in \arg\min_{\lambda \in \neg_\varepsilon \tilde{z}_n} \|\theta_n - \lambda\|^2_{V_{w_n}}$; // Closest alternative
  Set $U_{n,a} = \left(\|\theta_n - \lambda_n\|_{aa^\intercal} + \sqrt{c_{n,a}}\right)^2$ for all $a \in \mathcal{A}$; // Optimistic gains
  Feed $\mathcal{L}^{\mathcal{A}}$ with gain $g_n(w) = (1 - \frac{1}{n})\langle w, U_n \rangle$; // Update learner
  Pull $I_n \in \arg\min_{a \in \mathcal{A}} \{N_{n,a} - W_{n,a}\}$, observe X_{n,I_n} and update (θ_n, N_n) ;
end for

Algorithm 7.1: LεBAI algorithm.

Structure After pulling each arm once, at each round $n > n_0$, if the stopping condition (7.3) for the candidate answer $\hat{z}_n \in z_F(\theta_n, N_n)$, we return $\hat{z}_n$; else, the sampling rule returns an arm I_n to pull. Then, the statistics are updated based on this new observation.

Sampling rule The algorithmic ingredients used in the sampling rule of LεBAI build upon the ones in LinGame [Degenne et al., 2020a]. It is a saddle-point algorithm approximating a two-player zero-sum game. At each round $n > n_0$, if the algorithm hasn't stopped yet, the agent chooses a candidate answer and a pulling proportion over arms $\left(\tilde{z}_n, w_n^{\mathcal{L}^{\mathcal{A}}}\right) \in \mathcal{Z}_\varepsilon(\theta_n) \times \Sigma_K$, where $\tilde{z}_n$ can be different from $\hat{z}_n$. A mild logarithmic forced exploration is added, *i.e.* $w_n = \frac{1}{nK} 1_K + \left(1 - \frac{1}{n}\right) w_n^{\mathcal{L}^{\mathcal{A}}}$. The agent will play by combining a no-regret learner on Σ_K (*e.g.* AdaHedge of de Rooij et al. [2014]), denoted by $\mathcal{L}^{\mathcal{A}}$, and a $\mathcal{Z}$-oracle, denoted by $\mathcal{L}^{\mathcal{Z}}$. While Theorem 7.6 was proven for $\tilde{z}_n \in z_F(\nu_n)$, we obtain similar empirical performance with the heuristic $\tilde{z}_n \in z_F(\nu_n, N_n)$ at a much lower computational cost.

Given $(\tilde{z}_n, w_n)$ from $\mathcal{L}^{\mathcal{Z}} \times \mathcal{L}^{\mathcal{A}}$, the nature plays the most confusing alternative parameter $\lambda_n \in \arg\min_{\lambda \in \neg_\varepsilon \tilde{z}_n} \|\theta_n - \lambda\|^2_{V_{w_n}}$. To update $\mathcal{L}^{\mathcal{A}}$, the agent uses gains $g_n(w) = (1 - \frac{1}{n})\langle w, U_n \rangle$ where the optimistic gains are defined for all $a \in \mathcal{A}$ as $U_{n,a} = \left(\|\theta_n - \lambda_n\|_{aa^\intercal} + \sqrt{c_{n,a}}\right)^2$ with $c_{n,a} = \min\left\{2c\left(n^2, n^{2/3}\right) \|a\|^2_{V_{N_n}^{-1}}, 4L^2_{\mathcal{M}} L^2_{\mathcal{A}}\right\}$. Under a good event, the quantity $\langle w, U_n \rangle$ is an upper bound on the unknown $\inf_{\lambda \in \neg_\varepsilon z_F(\nu)} \|\theta - \lambda\|^2_{V_w}$. Finally, I_n is obtained deterministically by *tracking*, *i.e.* $I_n \in \arg\min_{a \in \mathcal{A}} \{N_{n,a} - w_{n,a}\}$.

To obtain efficient implementations for combinatorial or large arms sets K , LεBAI should be modified by using existing improvements for game-based algorithms [Tirinzoni et al., 2020, Réda et al., 2021, Jourdan et al., 2021]. When $\varepsilon = 0$, LεBAI is close to LinGame, but uses one

learner instead of Z learners. Other differences are that LinGame uses regularization in the estimator and a stopping threshold featuring d .

7.4.1 Sample Complexity Upper Bound

For both the multiplicative and the additive ε-optimality, Theorem 7.6 shows that LεBAI yields an (ε, δ)-PAC and asymptotically optimal algorithm. The proof sketch of Theorem 7.6 is inspired by the one of LinGame [Degenne et al., 2020a], hence we will only highlight the novel technical difficulties that had to be addressed.

Theorem 7.6. *Let $\mathcal{L}^{\mathcal{A}}$ with sub-linear regret* (e.g. *AdaHedge*) *and $\mathcal{L}^{\mathcal{Z}}$ returning $\tilde{z}_n \in z_F(\nu_n)$. Using* (7.4) *as stopping threshold $c(n, \delta)$, LεBAI yields an (ε, δ)-PAC algorithm which satisfies that, for all $\nu \in \mathcal{D}^K$ such that $\theta \in \mathcal{M}$ and $|z_F(\nu)| = 1$, $\limsup_{\delta \to 0} \mathbb{E}_\nu [\tau_{\varepsilon,\delta}] / \log(1/\delta) \leq T_\varepsilon(\nu)$.*

Technical difficulties Since the proof of Theorem 7.6 is inspired by the one of LinGame [Degenne et al., 2020a], we highlight the novel technical difficulties that had to be addressed. In BAI, we have $|z^\star(\theta)| = 1$. The key property used in BAI proofs which does not hold in ε-BAI is that, for all $z \neq z^\star(\theta)$, θ belongs to the alternative $\neg_0 z$. The consequence of this is that whenever the answer used by the sampling rule $\tilde{z}_n$ is wrong, the correct parameter belongs to $\neg_0 \tilde{z}_n$, hence the algorithm will sample in order to try and exclude that true parameter, which cannot succeed and will at some point correct the mistake. In ε-BAI we can have $\tilde{z}_n \neq z_F(\nu)$ while having $\theta \notin \neg_\varepsilon \tilde{z}_n$ and there is a priori no such self-correction mechanism to enforce that $\tilde{z}_n = z_F(\nu)$ after a while.

Our analysis reveals that a similar self-correction mechanism can be obtained for LεBAI. Let $\neg_F z$ be the *furthest alternative* to z , *i.e.* the set of parameters for which z is not the unique furthest answer. Intuitively, as it uses $\tilde{z}_n \in z_F(\nu_n)$, LεBAI samples to asymptotically exclude $\neg_F \tilde{z}_n$. Leveraging the logarithmic forced exploration, this cannot succeed when $\tilde{z}_n \neq z_F(\nu)$. Those two choices yield a self-correction mechanism for ε-BAI. More formally, we show that, under a good concentration event, the event $\{\tilde{z}_t \neq z_F(\nu)\}_{t \in [n]}$ only happens sub-linearly in n .

Related work Tackling ε-BAI in MAB for additive ε-optimality, ε-TaS [Garivier and Kaufmann, 2021] recommends $z_F(\nu_n, N_n)$ and uses the associated GLR as stopping rule. The sampling rule computes $w_F(\nu_n)$ and then tracks it with added forced exploration. Addressing additive spectral bandits, SpectralTaS [Kocák and Garivier, 2021] recommends $z^\star(\theta_n)$ and uses the GLRT associated with $z_F(\nu_n, N_n)$ for the stopping rule. For the sampling rule, a mirror ascent algorithm is run based on a super-gradient of a function depending on any ε-optimal answer. While the choice of the answer is not discussed, it is our understanding that a greedy answer is

used (matching their candidate answer). When considering ε-BAI on the unit sphere, Jedra and Proutiere [2020] recommend $z^\star(\theta_n)$ and use the associated GLRT, however their sampling rule is uniform over a spanner.

Designed for the multiple-correct answer setting, Sticky TaS [Degenne and Koolen, 2019] is a modified TaS algorithm: at round n , they compute $\bigcup_{\kappa\in\mathcal{C}_n} z_F(\kappa)$ where $\mathcal{C}_n$ is a continuous confidence region around ν_n , and stick to one of those (given an arbitrary order). For some identification problems (*e.g.* GAI), it rewrites as computing a finite number of furthest answers. There is no such rewriting for ε-BAI, hence Sticky TaS is not computationally feasible. Experiments suggest that it performs on par with ε-TaS at a higher computational cost, *i.e.* solving the same optimization for each parameter in a confidence region.

7.4.2 Proof Sketch of Theorem 7.6

The proof scheme sketched below is inspired by the game approach studied in Degenne et al. [2020a]. First, we derive a non-asymptotic one, then take the limit $\delta \to 0$. Having multiple ε-optimal answers is a key difficulty in several arguments.

Let $f(n) := 2c(n-1, n^{1/s})$ with $c(n,\delta)$ as in (7.4) and $s > 1$. Let

$$\mathcal{E}_n := \left\{\forall t \leq n,\ \|\theta_t - \theta\|^2_{V_{N_t}} \leq f(n)\right\} . \tag{7.6}$$

Using concentration arguments, we have $\sum_n \mathbb{P}_\nu(\mathcal{E}_n^\complement) \leq \zeta(s)$ where ζ is the Riemann ζ function. Using Lemma 2.26 in Chapter 2, the proof boils down to construct a time $T_\nu(\delta) > K$ such that $\mathcal{E}_n \subseteq \{\tau_{\varepsilon,\delta} \leq n\}$ for $n \geq T_\nu(\delta)$ since it yields that $\mathbb{E}_\nu[\tau_{\varepsilon,\delta}] \leq T_\nu(\delta) + K\zeta(s)$. Then, a sufficient condition to conclude the proof is to show that $\limsup_{\delta\to 0} T_\nu(\delta)/\log(1/\delta) \leq T_\varepsilon(\nu)$. To construct such a $T_\nu(\delta)$, it is sufficient to show that under $\mathcal{E}_n$, if the algorithm does not stop at time n , then $nT_\varepsilon(\nu)^{-1} \leq \log(1/\delta) + \tilde{\mathcal{O}}\left(n^{1-\beta_1} + \log(1/\delta)^{1-\beta_2}\right)$ where $(\beta_1,\beta_2) \in (0,1)^2$.

The analysis distinguishes between two independent components. Under $\mathcal{E}_n$, if the algorithm does not stop at time n , we can show that the stopping-recommendation pair satisfy

$$2c(n-1,\delta) \geq \max_{z\in\mathcal{Z}} \inf_{\lambda\in\neg_\varepsilon z} \|\theta - \lambda\|^2_{V_{N_n}} - \tilde{\mathcal{O}}\left(n^{1-\beta_1} + \log(1/\delta)^{1-\beta_2}\right) . \tag{7.7}$$

The *r.h.s.* only features the empirical counts, and the proof is independent of the sampling rule itself. Then, it is possible to show that the sampling rule satisfy

$$\max_{z\in\mathcal{Z}} \inf_{\lambda\in\neg_\varepsilon z} \|\theta - \lambda\|^2_{V_{N_n}} \geq 2nT_\varepsilon(\nu)^{-1} - \tilde{\mathcal{O}}\left(n^{1-\beta_1} + \log(1/\delta)^{1-\beta_2}\right) . \tag{7.8}$$

Due to the similarity with existing proof techniques [Degenne et al., 2020a] and the fact that the Top Two approach is the main focus of this manuscript, the proofs of (7.7) and (7.8) are

omitted. For more details, we refer the reader to Lemma E.5 in Appendix E.2 [Jourdan and Degenne, 2022] for (7.7), and to Lemmas E.6 and E.7 in Appendix E.3 for (7.8).

7.5 Experiments

We show that LεBAI has competitive empirical performance compared to existing ε-BAI algorithms, which are computationally expensive, and that using an instantaneous furthest answer is efficient both in terms of computational cost and sample complexity. Moreover, LεBAI performs on par with the modified BAI algorithms, which are not asymptotically optimal, on hard and random instances.

As heuristic with lower computational cost (not supported by Theorem 7.6), the $\mathcal{Z}$-oracle in LεBAI returns an instantaneous furthest answer, *i.e.* $\tilde{z}_n \in z_F(\theta_n, N_n)$. The experiments below are considering the multiplicative ε-optimality. Similar experiments can be done in the additive setting (see Jourdan and Degenne [2022]). We use the same experimental setup as in Section 7.3.3. On the 5000 runs, we report the standard deviation of means by using sub-samples of 100 runs.

ε-BAI and candidate answer Even when K is small, algorithms based on solving the optimization problem $(z_F(\nu), w_F(\nu))$ are intractable, *i.e.* ε-TaS or recommending the furthest answer. We evaluate their performance empirically on the hard instance with $\mathcal{A} = \{e_1, e_2\}$, and discretize uniformly Δ_2 with 500 vectors.

In Table 7.1, we combine and compare four ε-BAI sampling rules with three candidate answers for the stopping rule (7.3). Comparing the rows of Table 7.1 reveals that LεBAI performs on par with ε-TaS and the "oracle" *fixed* algorithm, which tracks the unknown optimal allocation $w_F(\nu)$. It also consistently outperforms uniform sampling ($\approx 85\%$).

Based on Table 7.1, greedy answer is consistently worse than a (instantaneous) furthest answer, with a ratio of stopping time being on average 0.92 (coherent with Figure 7.1(b)). Moreover, it highlights that using an instantaneous furthest answer achieves similar perfor-

Table 7.1 – Empirical stopping time ($\pm$ standard deviation) on the hard instance with $\mathcal{A} = \{e_1, e_2\}$.

	$z^\star(\theta_n)$	$z_F(\nu_n)$	$z_F(\nu_n, N_n)$
LεBAI	416 ($\pm$13)	383 ($\pm$16)	381 ($\pm$17)
ε-TaS	400 ($\pm$14)	371 ($\pm$15)	371 ($\pm$15)
Fixed	401 ($\pm$14)	374 ($\pm$14)	374 ($\pm$14)
Uniform	492 ($\pm$16)	450 ($\pm$17)	449 ($\pm$17)

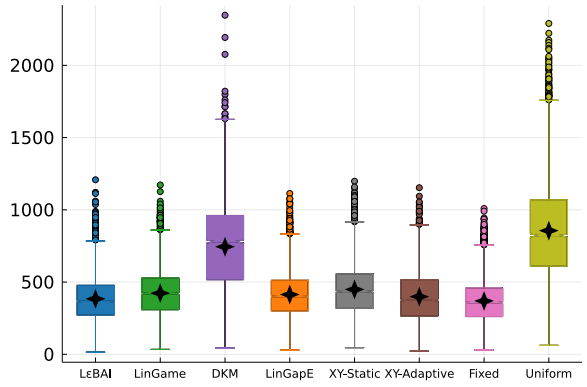

Figure 7.3 – Empirical stopping time on the hard instance with $\mathcal{A} = \mathcal{Z}$. The modified BAI algorithms use (7.3) with $\hat{z}_n \in z_F(\nu_n, N_n)$.

mance as a furthest answer at a lower computational cost. In the following experiments, the stopping-recommendation pair is (7.3) combined with $\hat{z}_n \in z_F(\nu_n, N_n)$.

Modified BAI algorithms Figure 7.3 compares LεBAI with the modified BAI algorithms, all using the same stopping-recommendation pair. We see that LεBAI slightly outperforms the modified LinGapE and $\mathcal{XY}$-Adaptive, performs better than the modified LinGame and $\mathcal{XY}$-Static and is on par with the "oracle" *fixed* algorithm. Uniform sampling and the modified DKM perform poorly.

Random instances To assess the impact of higher dimensions, random instances are considered (one per run). For the answer set, 19 vectors $(a_k)_{k\in[19]}$ are uniformly drawn from $\mathbb{S}^{d-1} := \left\{a \in \mathbb{R}^d : \|a\|_2 = 1\right\}$ and set $\theta = a_1$. To enforce multiple correct answers, a modification of the greedy answer is added such that $a_{20,i} = a_{1,i}$ for $i \neq i_0$ and $a_{20,i_0} = \frac{1-\|\theta\|_2^2+\mu_{i_0}^2-r_\varepsilon\varepsilon}{\theta_{i_0}}$

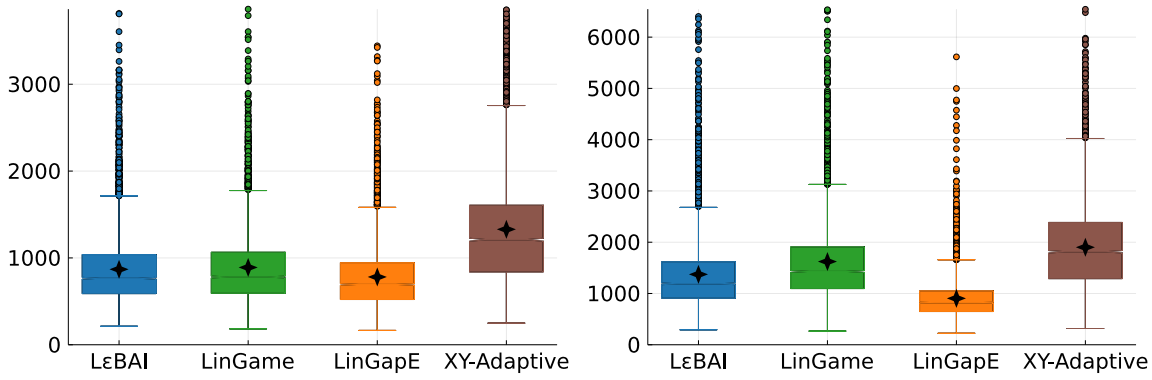

Figure 7.4 – Empirical stopping time on random instances ($\mathcal{A} = \mathcal{Z}$) for $d \in \{6, 12\}$ (from top left to bottom right). The modified BAI algorithms use (7.3) with $\hat{z}_n \in z_F(\nu_n, N_n)$.

where $i_0 = \arg\min_{i \in [d]} \theta_i$ and $r_\varepsilon = 0.1$. Those instances are motivated by a practical BAI example where a modified/corrupted version of the unique correct answer exists. Seeing the problem as an ε-BAI one allows to return an ε-optimal answer, while avoiding wasteful queries required by BAI algorithms.

In Figure 7.4, LεBAI shows similar empirical performance with modified BAI algorithms. Even though it is outperformed by the modified LinGapE, LεBAI is almost twice as fast as the modified $\mathcal{XY}$-Adaptive and appears to be slightly more robust than the modified LinGame to increasing dimension.

7.6 Discussion

In Chapter 7, we showed that the choice of the candidate answer is important for the sample complexity in fixed-confidence ε-BAI. Using an instantaneous furthest answer as candidate answer, we proposed a simple procedure to adapt existing BAI algorithms for ε-BAI problems. Leveraging it in the sampling rule as well, we introduced LεBAI which is asymptotically optimal and has competitive empirical performance.

Computing the furthest answer requires solving the closest alternative sub-problem $|\mathcal{Z}_\varepsilon(\theta_n)|$ times. While that number is small (in particular much less that Z) in the examples we considered, that computation can become an issue if many different answers are close to each other. If we extend the setting to continuous answers, the computation of the furthest answer by iterating becomes unfeasible. The question of finding an ε-close point of a reward function in a non-finite set is the general question of optimization, which is central to many areas of machine learning. Extending the problem-dependent approach of the bandit framework to that setting is an interesting research direction.

Since the existence of a tight finite-time lower bound for multiple-correct answer setting is still an open problem, it remains unclear how to assess the theoretical performance of algorithms in this regime. We believe that, once derived, this lower bound would reveal the existence of strong moderate confidence terms (independent of δ) affecting the sample complexity, which could then be used to design ε-BAI algorithms with theoretical guarantees in both regime.

The approach of adapting the transportation costs to tackle the underlying structure in linear bandits is successful, as it was for general class for distributions (see Part I) or for problems with multiple correct answers (see Part II). While most of the algorithms presented in this thesis are inspired by the Top Two approach, we built upon the game approach in Chapter 7. Compared to the Track-and-Stop approach, the game based approach is less computationally expensive. However, it still more costly than the Top Two approach, which is easier to implement and interpret. Understanding how to adapt the Top Two approach for linear bandits is the topic of Chapter 8.

Chapter 8

Extending the Top Two Approach

In Chapter 8, we also study the ε-BAI problem for linear bandit in the fixed-confidence setting, as done in Chapter 7. The presented results are currently unpublished since the challenges of the theoretical analysis are not solved yet.

Given the recent success of the Top Two approach for the vanilla ε-BAI problem, it is natural to wonder whether this approach could be extended for linear bandits. The transductive setting is particularly relevant to understand the different roles played by the arms and the answers. The arms should be pulled to verify that an answer is better than another one. Therefore, the Top Two approach should be seen as defining a candidate answer and the associated most confusing alternative answer (*i.e.* which challenges the most our current belief), and collecting additional observations to compare those top two answers. We propose LεTT which extends the Top Two approach to tackle ε-BAI for transductive linear bandits. After performing an empirical study showcasing the good empirical performance of the resulting algorithms, we highlight the challenges in the analysis of the expected sample complexity.

Contents

8.1 Introduction

We consider the exact same setting as in Chapter 7, hence we refer the reader to Section 7.1 for a detailed problem statement and to Sections 7.2 and 7.3 for additional notation. The unknown regression parameter is denoted by $\theta \in \mathbb{R}^d$, the set of known arms $\mathcal{A} \subset \mathbb{R}^d$ and known answers $\mathcal{Z} \subset \mathbb{R}^d$. Let $\mathcal{Z}_\varepsilon(\theta) = \{z \in \mathcal{Z} \mid \langle \theta, z \rangle \geq \max_{z \in \mathcal{Z}} \langle \theta, z \rangle - \varepsilon\}$ be the set of ε-good answers for $\varepsilon \geq 0$. For multiplicative ε-BAI, the means are non-negative and $\varepsilon \in [0, 1)$ and $\mathcal{Z}_\varepsilon^{\mathrm{mul}}(\theta) = \{z \in \mathcal{Z} \mid \langle \theta, z \rangle \geq (1 - \varepsilon) \max_{z \in \mathcal{Z}} \langle \theta, z \rangle\}$.

Characteristic times When $\mathcal{M} = \mathbb{R}^d$, the inverse of the characteristic times can be written as

$$T_\varepsilon(\nu, z)^{-1} = \max_{w \in \Sigma_K} \min_{x \in \mathcal{Z} \setminus \{z\}} C_\varepsilon(z, x; \nu, w) \quad \text{and} \quad T_\varepsilon^{\mathrm{mul}}(\nu, z)^{-1} = \max_{w \in \Sigma_K} \min_{x \in \mathcal{Z} \setminus \{z\}} C_\varepsilon^{\mathrm{mul}}(z, x; \nu, w) \,, \tag{8.1}$$

with the transportation costs between (z, x) with allocation $w \in \Sigma_K$ on instance ν are

$$C_\varepsilon(z, x; \nu, w) = \frac{(\varepsilon + \langle \theta, z - x \rangle)^2}{2\|z - x\|^2_{V_w^\dagger}} \mathbb{1}\left(\varepsilon + \langle \theta, z - x \rangle > 0\right) \,,$$
$$C_\varepsilon^{\mathrm{mul}}(z, x; \nu, w) = \frac{\langle \theta, z - (1 - \varepsilon)x \rangle^2}{2\|z - (1 - \varepsilon)x\|^2_{V_w^\dagger}} \mathbb{1}\left(\langle \theta, z - (1 - \varepsilon)x \rangle > 0\right) \,,$$

where $V_w = \sum_{a \in \mathcal{A}} w_a a a^\intercal$ and $V^\dagger$ denotes the Moore-Penrose pseudo-inverse of V .

GLR stopping rules Let ν_n denotes a bandit instance with regression parameter $\theta_n = V_{N_n}^{-1} \sum_{t \in [n-1]} X_{t, I_t} I_t$ where $N_{n,a} = \sum_{t \in [n-1]} \mathbb{1}\left(I_t = a\right)$. The empirical transportation costs are defined as

$$W_{\varepsilon,n}(z, x) = C_\varepsilon(z, x; \nu_n, N_n) \quad \text{and} \quad W_{\varepsilon,n}^{\mathrm{mul}}(z, x) = C_\varepsilon^{\mathrm{mul}}(z, x; \nu_n, N_n) \,.$$

Using the instantaneous furthest answer, the GLR_ε stopping rule can be written as

$$\tau_{\varepsilon,\delta} = \inf \left\{ n \mid \max_{z \in \mathcal{Z}_\varepsilon(\theta_n)} \min_{x \in \mathcal{Z} \setminus \{z\}} W_{\varepsilon,n}(z, x) > c(n - 1, \delta) \right\} \,, \tag{8.2}$$

and a similar formula would give $\tau_{\varepsilon,\delta}^{\mathrm{mul}}$ for multiplicative ε-BAI. The threshold function $c(n, \delta)$ can be chosen as in (7.4) to ensure that the algorithm is (ε, δ)-PAC regardless of the sampling rule (see Lemma 7.5 in Chapter 7).

Contribution 8.1. *The contributions of Chapter 8 are the following.*

- *We extend the Top Two approach to tackle structured bandits by proposing the Structured Top Two approach, which is also specified by four choices (leader answer, challenger answer, target allocation and mechanism to reach it). We propose several instances for ε-BAI in transductive linear bandits with Gaussian distributions, including LεTT which generalizes the EB-TC$_\varepsilon$ algorithm studied in Chapter 5.*
- *We highlight the challenges in the analysis of the expected sample complexity to obtain the asymptotic optimality of LεTT. When there is no randomness in the collected observations, LεTT can be written as a saddle-point algorithm to solve $T_\varepsilon(\nu)^{-1}$. While the analysis is still an open problem, this draws a connection between the Structured Top Two approach and the game based approach.*
- *Our empirical study showcases the competitive empirical performance of LεTT.*

8.2 Linear Top Two Algorithm

As summarized in Algorithm 2.1, a Top Two sampling rule is defined by four components: leader answer, challenger answer, target allocation and mechanism to reach it. Extending the Top Two approach to tackle a structured bandits such as transductive linear bandits requires some modification. While we still want to verify that the leader answer is better than the challenger answer, we cannot simply define a target allocation $(\tilde{\beta}_n(B_n, C_n), 1-\tilde{\beta}_n(B_n, C_n)) \in [0,1]^2$ for the leader/challenger pair (B_n, C_n) : (1) sampling any arm will reveal information on the vector θ and (2) an answer might not be an arm that can be pulled. Therefore, given a leader/challenger pair, we need to define a target allocation which might be supported on the whole set of arms $\mathcal{A}$, *i.e.* $\beta_n(B_n, C_n) \in \Sigma_K$. While we focus on ε-BAI in transductive linear bandits, it is possible to propose a Structured Top Two sampling rule in general, which is summarized in Algorithm 8.1.

As done in Chapter 7, we initialize by sampling each arm once, even though a smarter initialization is possible for linear bandits (*e.g.* sample a set of arms that covers $\mathbb{R}^d$). While we present the details for the additive setting, the formulas can also be straightforwardly adapted to tackle the multiplicative setting.

Naming convention As in Section 2.2.5, we use `{leader}-{challenger}-{target}` as naming convention. Recall that the mechanism to reach the target allocation (Section 8.2.2) is implicitly defined by the first three choices: using tracking only if the leader/challenger pair is deterministic, and randomization otherwise. Note that the exact formulas for each of those four choices explicitly depends on the considered structured bandits.

Input: Mechanisms to choose the leader answer $\mathcal{L}^B$, the challenger answer $\mathcal{L}^C$, the target allocation $\mathcal{L}^T$ and how to reach the target $\mathcal{L}^R$.
Output: Next arm to sample I_n .
Get $B_n \in \mathcal{Z}$ from $\mathcal{L}^B$; // Leader answer
Get $C_n \in \mathcal{Z} \setminus \{B_n\}$ from $\mathcal{L}^C$; // Challenger answer
Get $\boldsymbol{\beta_n(B_n, C_n) \in \Sigma_K}$ from $\mathcal{L}^T$; // Target allocation
Get $I_n \in [K]$ from $\mathcal{L}^R$ using $\beta_n(B_n, C_n)$; // Reaching the target

Algorithm 8.1: Structured Top Two sampling rule.

Leader answer In Section 2.2.1, we presented several choices of leader answer which can be straightforwardly adapted to identify $i^\star(\theta)$. The EB leader selects $B_n^{\mathrm{EB}} \in \arg\max_{z\in\mathcal{Z}} \langle \theta_n, z\rangle$. The UCB leader answer uses $B_n^{\mathrm{UCB}} \in \arg\max_{z\in\mathcal{Z}} U_{n,z}$ with

$$U_{n,z} = \langle \theta_n, z\rangle + \sqrt{g(n)}\|z\|_{V_n^{-1}} ,$$

where $g(n) = \Theta(\log n)$. Given a sampler Π_n on $\mathbb{R}^d$, the TS leader answer chooses $B_n^{\mathrm{TS}} \in i^\star(\lambda_n)$ with $\lambda_n \sim \Pi_n$. For Gaussian distributions with unit variance, using the improper prior $\Pi_1 = \mathcal{N}(0_d, +\infty I_d)$ yields $\Pi_n = \mathcal{N}(\theta_n, V_{N_n}^{-1})$ as posterior distribution before time n .

However, Chapter 7 advocates that one should identify the furthest answer $z_F(\nu)$ defined in (7.1) instead of the greedy answer $z^\star(\theta) = \arg\max_{z\in\mathcal{Z}}\langle \theta, z\rangle$, which is doomed to be a sub-optimal choice when $z_F(\nu) \neq z^\star(\theta)$. Therefore, we should not use the EB, UCB or TS leaders when the goal is to achieve asymptotic optimality. Inspired by Chapter 7, we could use the F_ε (Furthest) leader answer, meaning $B_n^{\mathrm{F}_\varepsilon} \in z_F(\nu_n)$, or the IF_ε (Instantaneous Furthest) leader answer, defined as

$$B_n^{\mathrm{IF}_\varepsilon} \in \operatorname*{arg\,max}_{z\in\mathcal{Z}_\varepsilon(\theta_n)} \min_{x\in\mathcal{Z}\setminus\{z\}} W_{\varepsilon,n}(z,x) = z_F(\nu_n, N_n) \quad \text{with } z_F(\nu_n, N_n) \text{ as in (7.5).} \tag{8.3}$$

Since the F_ε leader answer has a high computational cost, we advocate to use the IF_ε leader answer. Those two choices will be asymptotically equivalent provided that the sampling rule ensures convergence towards the set of optimal furthest allocation $w_F(\nu)$.

As for the EB leader, the IF_ε leader might be too greedy when no additional exploration is enforced (implicitly or explicitly). While one could derive an optimistic version of the IF_ε leader (by deriving concentration results), it is actually not clear how to use a sampler to obtain a randomized version of the IF_ε leader. Fortunately, Chapter 5 advocates that there is no need for additional exploration when considering ε-BAI since $\varepsilon > 0$ acts as a regularizer.

Remark 8.2 (Simple Regret). *When the goal is to have a vanishing simple regret on the candidate answer (as for EB-TC$_\varepsilon$-1/2 in Chapter 5), the instantaneous furthest answer is doomed to be sub-optimal on some instances. On instances θ such that $z_F(\nu) \notin z^\star(\theta) + Span(\{\theta\})^\perp$, we have $\lim_{n\to+\infty} \mathbb{E}_\nu[\langle \theta, z^\star(\theta) - \hat{z}_n\rangle] > 0$ for candidate answers $(\hat{z}_n)_n$ such that $\lim_{n\to+\infty} \hat{z}_n = z_F(\nu)$ (since $\langle \theta, z^\star(\theta) - z_F(\nu)\rangle > 0$). Therefore, we should use a leader which identifies $z^\star(\theta)$,* e.g. *the EB, UCB or TS leaders.*

Challenger answer In Section 2.2.2, we presented several choices of challenger answer which can be straightforwardly adapted. The TC$_\varepsilon$ and TCI$_\varepsilon$ challenger consider

$$C_n^{\mathrm{TC}_\varepsilon} \in \operatorname*{arg\,min}_{z\in\mathcal{Z}\setminus\{B_n\}} W_{\varepsilon,n}(B_n, z) \quad \text{and} \quad C_n^{\mathrm{TCI}_\varepsilon} \in \operatorname*{arg\,min}_{z\in\mathcal{Z}\setminus\{B_n\}} \{W_{\varepsilon,n}(B_n, z) + \log N_{n,z}\} \ .$$

The RS$_\varepsilon$ challenger takes *i.e.* $C_n^{\mathrm{RS}_\varepsilon} \in \arg\max_{z\in\mathcal{Z}} \langle \lambda_n, z\rangle$ with $\lambda_n \sim \Pi_n$ until $B_n \notin \mathcal{Z}_\varepsilon(\lambda_n)$. Note that the resampling step is even more computationally expensive when $\varepsilon > 0$.

Historically, LinGapE [Xu et al., 2018] is an extension of the LUCB algorithm to the linear bandit setting. As such, it is the first Top Two algorithm for structured bandits. LinGapE uses the EB leader and a UCB challenger, *i.e.* $C_n^{\mathrm{UCB}} \in \arg\max_{z\in\mathcal{Z}\setminus\{B_n\}} \{\langle \tilde{\theta}_n, z - B_n\rangle + \sqrt{g(n)}\|z - B_n\|_{V_{N_n}^{-1}}\}$ where $g(n)$ is a bonus calibrated with concentration results and $\tilde{\theta}_n$ is the regularized least-squared estimator, *i.e.* $\tilde{\theta}_n = (V_{N_n} + \lambda I_d)^{-1} \sum_{t\in[n-1]} X_{t,I_t} I_t$. Xu et al. [2018] proposed several target allocations given this pair of leader/challenger. While it remains a heuristic, the target allocation achieving the best empirical performance is a *greedy* (deterministic) choice, which samples $I_n \in \arg\min_{a\in\mathcal{A}} \|B_n - C_n\|_{(V_{N_n} + aa^\intercal)^{-1}}$. It stops when the empirical gap between the LCB of B_n and the UCB of C_n is lower than ε .

8.2.1 Target Allocation Over Arms

Due to the underlying structure, it is possible to collect information on any arms to verify that the leader is better than the challenger. Conditioned on (B_n, C_n) , one should define a target allocation over arms, which we will denote by $\beta_n(B_n, C_n) \in \Sigma_K$. As in Section 2.2.3, we first present several fixed designs, which arose out of convenience yet are doomed to be sub-optimal. Then, we extend the optimal design of IDS to the transductive linear bandit setting.

Fixed design approach Since answers are not arms that can be pulled, it is neither possible to sample both the leader and the challenger answers, nor possible to sample the leader with a fixed proportion β . Moreover, even when θ is known, it is not clear which arm is informative enough to allocate a fixed proportion β to it.

It is straightforward to generalize the idea of sampling the arm that yields the largest empirical transportation cost would the estimator θ_n be unchanged, *i.e.*

$$I_n \in \arg\max_{a\in\mathcal{A}} C_\varepsilon(B_n, C_n; \nu_n, N_n + 1_a) \,.$$

For Gaussian distribution, only the denominator of $C_\varepsilon(z, x; \nu, w)$ depends on the allocation w , and the means are "decoupled" from the allocation. Therefore, we obtain that

$$I_n \in \arg\min_{a\in\mathcal{A}} \|B_n - C_n\|_{(V_{N_n}+aa^\mathsf{T})^{-1}} = \arg\max_{a\in\mathcal{A}} \frac{\langle V_{N_n}^{-1}a, B_n - C_n\rangle^2}{1 + \|a\|^2_{V_{N_n}^{-1}}} \,, \tag{8.4}$$

where the second equality uses the Sherman-Morrison formula $(A+uv^\mathsf{T})^{-1} = A^{-1} - \frac{A^{-1}uv^\mathsf{T}A^{-1}}{1+v^\mathsf{T}A^{-1}u}$. The choice (8.4) is made by the greedy version of LinGapE [Xu et al., 2018]. Similarly, one can sample the arm for which the gradient of the empirical transportation cost is the largest, *i.e.*

$$I_n \in \arg\max_{a\in\mathcal{A}} \frac{\partial C_\varepsilon(B_n, C_n; \nu_n, N_n)}{\partial w_a} = \arg\max_{a\in\mathcal{A}} \langle V_{N_n}^{-1}a, B_n - C_n\rangle^2 \,, \tag{8.5}$$

where the second equality uses Lemma 8.3 (proved with standard linear algebra). While (8.5) differs slightly from (8.4), there are asymptotically equivalent since $\|a\|^2_{V_{N_n}^{-1}} \to_{n\to+\infty} 0$.

Lemma 8.3. *Let w such that V_w is invertible. For all $z \in \mathcal{Z}_\varepsilon(\theta)$, $x \neq z$ and $a \in \mathcal{A}$,*

$$\frac{\partial C_\varepsilon(z, x; \nu, w)}{\partial w_a} = C_\varepsilon(z, x; \nu, w)\frac{\langle V_w^{-1}a, z - x\rangle^2}{\|z - x\|^2_{V_w^{-1}}} \text{ and } \sum_{a\in\mathcal{A}} w_a \frac{\partial C_\varepsilon(z, x; \nu, w)}{\partial w_a} = C_\varepsilon(z, x; \nu, w) \,. \tag{8.6}$$

The choice (8.5) is made by BC-TE [Lee et al., 2023]. For vanilla bandits, BC-TE [Lee et al., 2023] is shown to be near optimal with a characteristic time $\underline{T}(\nu)$ defined in (2.9). Compared to $T^\star(\nu)$, the optimization over the simplex in $\underline{T}(\nu)$ is constrained with a problem-dependent condition, namely $w_{(2)}/(w_{i^\star} + w_{(2)}) = \gamma$ where γ is defined implicitly in (2.10). For transductive linear bandits, using (8.5) also yields a constraint on the asymptotic allocation. Providing an explicit formula for this constraint is beyond the scope of this thesis. Let $\underline{T}_\varepsilon(\nu, z)$ be the characteristic time defined as $T_\varepsilon(\nu, z)$ in (8.1) where the optimization over the simplex is restricted by this (conjectured) problem-dependent condition. We conjecture that (1) $\underline{T}_\varepsilon(\nu, z^\star(\theta))/T_\varepsilon(\nu, z^\star(\theta))$ is not too large, and (2) that $\underline{T}_\varepsilon(\nu, z^\star(\theta))$ can be reached asymptotically by using the EB-TC algorithm with (8.5). While the greedy LinGapE [Xu et al., 2018] relies on the UCB challenger, this might explain the good empirical performance of the greedy LinGapE [Xu et al., 2018] and pave the way to a theoretical analysis of this heuristic algorithm (which has state-of-the-art empirical performance).

In both (8.5) and (8.4), the ties are broken arbitrarily at random. Therefore, it is equivalent to using a randomized mechanism to reach the target allocation $\beta_n(B_n, C_n)$, which is a vector uniformly supported on the set of maximizers of the *r.h.s.* of each equation.

Optimal design IDS Similarly as in Section 2.2.3, it is possible to simplify the dual formulation of $T_\varepsilon(\nu, z)^{-1}$ in order to obtain a target allocation over arms extending IDS [You et al., 2023] for linear bandits. By computing the contribution of an arm a to the empirical transportation cost between (z, x) , we define $\beta_{n,a}(z, x) = 1/K$ for all $a \in \mathcal{A}$ if $\varepsilon + \langle \theta_n, z - x \rangle \leq 0$, and

$$\forall a \in \mathcal{A}, \quad \beta_{n,a}(z,x) = \frac{N_{n,a} \frac{\partial C_\varepsilon(z,x;\nu_n,N_n)}{\partial w_a}}{C_\varepsilon(z,x;\nu_n,N_n)} = N_{n,a} \frac{\langle V_{N_n}^{-1} a, z - x \rangle^2}{\|z - x\|^2_{V_{N_n}^{-1}}} \quad \text{otherwise} , \tag{8.7}$$

where the second equality uses Lemma 8.3 which also yields that $\beta_n(z, x) \in \Sigma_K$. As for the vanilla setting, the IDS proportions are independent of the empirical means and the slack ε for Gaussian with known variance.

Likewise, the IDS proportions are obtained by simplifying the dual formulation of the optimization problem $T_\varepsilon(\nu, z)^{-1} = \max_{w \in \Sigma_K} \min_{x \in \mathcal{Z} \setminus \{z\}} C_\varepsilon(z, x; \nu, w)$ which can be seen as the following convex optimization problem

$$T_\varepsilon(\nu,z)^{-1} = \max \left\{ \phi \mid \sum_{a \in \mathcal{A}} w_a = 1, \ \forall a \in \mathcal{A}, \ w_a \geq 0, \ \forall x \in \mathcal{Z} \setminus \{z\}, \ \phi - C_\varepsilon(z,x;\nu,w) \leq 0 \right\} . \tag{8.8}$$

Lemma 8.4 gives a necessary and sufficient condition for optimality in (8.8), which features a dual allocation vector $\gamma \in \Sigma_{Z-1}$. Intuitively, the dual variable γ_x should be thought as the conditional probability of selecting answer x as challenger given that the leader is z . Moreover, $\beta_a(z, x; \nu, w)$ represents the conditional probability of pulling arm a given that the leader/challenger pair of answers is (z, x) . Therefore, it is intuitive to take $\beta_n(z, x) = \beta(z, x; \nu_n, N_n)$. Lemma 8.4 recovers Lemma 2.5 for vanilla bandits (see proof in Appendix H.1).

Lemma 8.4. *Let $z \in \mathcal{Z}_\varepsilon(\theta)$. A feasible solution (ϕ, w) is optimal for (8.8) if and only if $\phi = T_\varepsilon(\nu, z)^{-1}$ and there exists a dual variable $\gamma \in \Sigma_{Z-1}$ such that, $\gamma_x(\phi - C_\varepsilon(z, x; \nu, w)) = 0$ for all $x \neq z$, and $w = \sum_{x \in \mathcal{Z} \setminus \{z\}} \gamma_x \beta(z, x; \nu, w)$ where $\beta(z, x; \nu, w) \in \Sigma_K$ is such that*

$$\forall a \in \mathcal{A}, \quad \beta_a(z,x;\nu,w) = \frac{w_a}{C_\varepsilon(z,x;\nu,w)} \frac{\partial C_\varepsilon(z,x;\nu,w)}{\partial w_a} .$$

ACC algorithm In a concurrent work, Qin and You [2023] introduced the Adaptive Culprit Curbing (ACC) and ACC-TS sampling rules for general pure exploration problems. Those algorithms are particular instances of Algorithm 8.1, which rely on randomization and use the IDS target allocation defined above. For the choice of the leader/challenger pair, ACC considers EB-TC and ACC-TS uses TS-RS. As for Algorithm 8.1, it is an open problem to analyze ACC and ACC-TS for linear bandits, yet they have good empirical performance.

While both ACC and Algorithm 8.1 were designed concurrently, we emphasize that Algorithm 8.1 arose by studying transductive linear bandit while ACC was aimed to tackle pure exploration problems in general. Due to its generality, ACC can tackle many interesting pure exploration problems. For example, ACC-TS recovers the Murphy sampling algorithm [Kaufmann et al., 2018] when considering GAI (see Chapter 6). Therefore, the analysis of these generic algorithms is an interesting research direction in the field of pure exploration problems.

Optimal design BOLD Similarly as in Section 2.2.3, one could attempt to generalize the BOLD target [Chen and Ryzhov, 2023, Bandyopadhyay et al., 2024] defined for vanilla BAI in (2.15), *i.e.*

$$I_n = B_n \quad \text{if} \quad \sum_{i \neq B_n} \frac{\frac{\partial C(B_n, i; \nu_n, N_n)}{\partial w_{B_n}}}{\frac{\partial C(B_n, i; \nu_n, N_n)}{\partial w_i}} > 1 \,, \quad \text{and} \quad I_n = C_n \quad \text{otherwise} \,.$$

Lemma 8.5 recovers Lemma 2.6 for vanilla bandits (see proof in Appendix H.2).

Lemma 8.5. *Let $z \in \mathcal{Z}_\varepsilon(\theta)$. An allocation w is optimal for $T_\varepsilon(\nu, z)^{-1}$ if and only there exists a dual variable $\gamma \in \Sigma_{Z-1}$ such that*

$$\begin{aligned} &\textit{Information balance:} && \forall x \in \mathcal{Z}_1, \quad C_\varepsilon(z, x; \nu, w) = T_\varepsilon(\nu, z)^{-1} \,, \\ &\textit{Overall balance:} && \forall a \in \mathcal{A}_1, \quad T_\varepsilon(\nu, z) \sum_{x \in \mathcal{Z}_1} \gamma_x \frac{\partial C_\varepsilon(z, x; \nu, w)}{\partial w_a} = 1 \,, \end{aligned}$$

where $\mathcal{A}_1 = \{a \in \mathcal{A} \mid w_a > 0\}$ and $\mathcal{Z}_1 = \{x \in \mathcal{Z} \setminus \{z\} \mid \gamma_x > 0\}$.

As in Lemma 2.6, Lemma 8.5 also features an information balance and an overall balance, which are expressed slightly differently. First, the information balance (or equality at equilibrium) will only holds for a subset of answers x , and we have $C_\varepsilon(z, x; \nu, w) > T_\varepsilon(\nu, z)^{-1}$ for the others. Intuitively, some answers directions might be redundant (or dominated by other directions). The situation is akin to the offline-online paradigm studied in Agrawal et al. [2023] where some answers directions are already saturated due to offline data. Second, the dual $\gamma \in \Sigma_{Z-1}$ has not been removed, and it is now defined implicitly as a solution of a system of

$|\text{supp}(w)|$ linear equations. The specificity of linear bandits lies in the fact that the support of an optimal allocation might not be dense (*i.e.* $|\text{supp}(w)| \leq K$). Intuitively, some arms might be redundant (or dominated by other directions) when collecting information on θ to return a correct answer $z \in \mathcal{Z}_\varepsilon(\theta)$. In the vanilla setting, there are also K equations to define $\gamma \in \Sigma_{K-1}$. Solving the first $K-1$ equations fully specifies γ , and plugging it in the last equation yields the overall balance (single) equation. Compared to the vanilla setting, it is less straightforward to derive a BOLD target out of the empirical version of the overall balance. There is not anymore a one-dimensional condition to be checked, and it is not straightforward to derive an empirical version of the dual parameter.

8.2.2 Mechanism to Reach the Target Allocation

As in Section 2.2.4, we need a mechanism to reach the target allocation $\beta_n(B_n, C_n) \in \Sigma_K$. The randomized approach will sample arm $I_n \sim \beta_n(B_n, C_n)$, *i.e.*

$$\forall a \in \mathcal{A}, \quad \mathbb{P}_{|n}(I_n = a \mid (B_n, C_n) = (z, x)) = \beta_{n,a}(z, x) \, .$$

Note that the fixed design approaches presented above rely on randomization. When the leader/challenger pair is deterministic, we can rely on the tracking approach when using the optimal design IDS. More precisely, we consider $Z(Z-1)$ tracking procedures, one per pair of answers $(z, x) \in \mathcal{Z}^2$ such that $z \neq x$. Given the leader/challenger pair of answers (B_n, C_n) at time n , the next arm to pull is

$$I_n = \arg\min_{a \in \mathcal{A}} \{N_{n,a}(B_n, C_n) - T_{n+1}(B_n, C_n)\overline{\beta}_{n+1,a}(B_n, C_n)\} \, , \tag{8.9}$$

with $T_n(z, x) = \sum_{t \in [n-1]} \mathbb{1}\left((B_t, C_t) = (z, x)\right)$, $N_{n,a}(z, x) = \sum_{t \in [n-1]} \mathbb{1}\left((B_t, C_t) = (z, x), I_t = a\right)$ and $\overline{\beta}_{n,a}(z, x) = T_n(z, x)^{-1} \sum_{t \in [n-1]} \mathbb{1}\left((B_t, C_t) = (z, x)\right) \beta_{t,a}(z, x)$. Using Theorem 6 in Degenne et al. [2020b] for the tracking procedure yields Lemma 8.6.

Lemma 8.6. *For all $n > K$, all $z \in \mathcal{Z}$, all $x \in \mathcal{Z} \setminus \{x\}$ and all $a \in \mathcal{A}$, we have $-\sum_{i=2}^{K} \frac{1}{i} \leq N_{n,a}(z, x) - T_n(z, x)\overline{\beta}_{n,a}(z, x) \leq 1$.*

8.2.3 LεTT Algorithm

The LεTT algorithm (see Algorithm 8.2) is a specific instance of Algorithm 8.1 which combines the IF$_\varepsilon$ leader, the TC$_\varepsilon$ challenger, the IDS target over arms and tracking to reach it. Using our naming convention, LεTT is IF$_\varepsilon$-TC$_\varepsilon$-IDS. For $\varepsilon = 0$, we recommend to use the TCI challenger to

alleviate a greedy behavior by fostering an implicit exploration over arms. For vanilla bandits, it recovers the EB-TC$_\varepsilon$-IDS algorithm studied in Chapter 5.

When the objective is to have vanishing simple regret, we recommend to use EB-TC$_\varepsilon$-IDS algorithm. Recall that the EB leader (or TS/UCB ones) should be used in that case since the IF$_\varepsilon$ leader answer is chosen to reach asymptotic optimality in the fixed-confidence setting.

8.3 Towards an Analysis of a Saddle-point Algorithm

At the moment, it is still an open problem to obtain any guarantees on any instances of the Structured Top Two approach. Even without the structural assumption of linear bandits, we should keep in mind that the asymptotic optimality of the Top Two approach in the fixed-confidence setting is only proven for Gaussian with known variance. For other classes of distributions, our proof only establishes asymptotic β-optimality. Likewise, anytime guarantees on the uniform probability of ε-error are only known for EB-TC$_\varepsilon$-β, but not for EB-TC$_\varepsilon$-IDS. Unfortunately, there is no natural notion of fixed design β for linear bandits. Due to the use of the optimal design IDS, there are still major challenges to be solved before showing that, for ε-BAI in transductive linear bandits, (1) the LεTT algorithm is asymptotically optimal and (2) the EB-TC$_\varepsilon$-IDS algorithm has anytime guarantees on the simple regret.

A natural first step to understand how to analyze a bandit algorithm is to consider the deterministic setting, where there is no randomness in the observations. In other words, we know the regression parameter θ exactly (*i.e.* $\theta_n = \theta$ for all n), hence we also know $\mathcal{Z}_\varepsilon(\theta)$. In the deterministic setting, the goal is now to solve the optimization problem $T_\varepsilon(\nu)^{-1}$ with a fully sequential algorithm. In other words, we would like to find a saddle-point of

$$T_\varepsilon(\nu)^{-1} = \max_{z \in \mathcal{Z}_\varepsilon(\theta)} \max_{w \in \Sigma_K} \min_{x \in \mathcal{Z} \setminus \{z\}} C_\varepsilon(z, x; \nu, w) \, .$$

Input: Slack $\varepsilon > 0$.
Output: Next arm to sample a_n and next recommendation $\hat{z}_n$.
Set $\hat{z}_n \in \arg\max_{z \in \mathcal{Z}_\varepsilon(\theta_n)} \min_{x \in \mathcal{Z} \setminus \{z\}} W_{\varepsilon,n}(z, x)$; `// Candidate answer`
Set $B_n = \hat{z}_n$ and $C_n \in \arg\min_{x \in \mathcal{Z} \setminus \{B_n\}} W_{\varepsilon,n}(B_n, x)$; `// Leader and challenger`
Set $w_{n,a}(B_n, C_n) = N_{n,a} \frac{\langle V_{N_n}^{-1} a, B_n - C_n \rangle^2}{\|B_n - C_n\|^2_{V_{N_n}^{-1}}}$, then update $(\overline{w}_{n+1,a}(B_n, C_n))_{a \in \mathcal{A}}$ and $T_{n+1}(B_n, C_n)$; `// Target allocation`
Set $a_n = \arg\min_{a \in \mathcal{A}} \{N_{n,a}(B_n, C_n) - T_{n+1}(B_n, C_n)\overline{w}_{n+1,a}(B_n, C_n)\}$; `// Tracking`

Algorithm 8.2: LεTT (or IF$_\varepsilon$-TC$_\varepsilon$-IDS) algorithm.

Game based approach For instance, LεBAI in Chapter 7 can be seen as a saddle-point algorithm. First, the deterministic LεBAI algorithm plays a leader answer $\tilde{z}_n \in \mathcal{Z}$ (Outer-Max player) using a $\mathcal{Z}$-oracle. In the analysis of Theorem 7.6, we used the F_ε leader answer as $\mathcal{Z}$-oracle. Since θ is known, the $\mathcal{Z}$-oracle is fixed to be $\tilde{z}_n \in z_F(\nu)$. Second, LεBAI sequentially learns the optimal allocation by playing an allocation over arms $w_n \in \Sigma_K$ (Inner-Max player) using a learner $\mathcal{L}^{\mathcal{A}}$ (*i.e.* an online optimization algorithm, *e.g.* AdaHedge [de Rooij et al., 2014]). For the deterministic setting, the forced exploration by mixing with the uniform allocation is not needed. Third, given w_n , the nature (Min player) plays the best response parameter which is the most confusing alternative regression parameter, *i.e.* $\lambda_n \in \arg\min_{\lambda \in \neg_\varepsilon z_F(\nu)} \|\theta_n - \lambda\|^2_{V_{w_n}}$. Since optimism is superfluous for the deterministic setting, the learner $\mathcal{L}^{\mathcal{A}}$ is updated with the gain vector $U_n \in \mathbb{R}^K_+$ where $U_{n,a} = \|\theta_n - \lambda_n\|^2_{aa^\intercal}$ for all $a \in \mathcal{A}$. Finally, the empirical allocation is updated by using tracking, *i.e.* $I_n \in \arg\min_{a \in \mathcal{A}} \{N_{n,a} - \sum_{t \in [n]} w_{t,a}\}$.

Since LεBAI uses the F_ε leader answer as $\mathcal{Z}$-oracle, LεBAI requires to first compute $z_F(\nu)$, hence to solve $T_\varepsilon(\nu)^{-1}$. Therefore, LεBAI is not a fully sequential saddle-point algorithm. As heuristic, we proposed to use the IF_ε leader answer as $\mathcal{Z}$-oracle, *i.e.* $\tilde{z}_n \in z_F(\nu, N_n)$. While the F_ε leader is fixed, the IF_ε leader answer depends on the current allocation N_n . This heuristic LεBAI is a fully sequential saddle-point algorithm, which achieve competitive empirical performance. Unfortunately, the proof still eludes us.

Structured Top Two approach The Structured Top Two approach can also be seen as a saddle-point algorithm. First, the deterministic Structured Top Two approach plays a leader answer $B_n \in \mathcal{Z}$ (Outer-Max player) using a leader mechanism $\mathcal{L}^B$. Since θ is known, the leader answer will belong to set of correct answers, *i.e.* $B_n \in \mathcal{Z}_\varepsilon(\theta)$ for all n , yet it might vary. Second, given the leader B_n , the nature plays a challenger answer $C_n \in \mathcal{Z} \setminus \{B_n\}$ (Min player) using a challenger mechanism $\mathcal{L}^C$. Third, given the leader/challenger pair (B_n, C_n) , the Structured Top Two approach plays an allocation over arms $\beta_n(B_n, C_n) \in \Sigma_K$ (Inner-Max player) using a target mechanism $\mathcal{L}^T$. Finally, the empirical allocation is updated by using a reaching mechanism $\mathcal{L}^R$.

Conceptually, the main difference with the game-based approach lies in the order in which each players acts. While the game-based approach considers the players (Outer-Max, Inner-Max, Min), the Structured Top Two approach uses the players (Outer-Max, Min, Inner-Max). In both cases, the players are acting sequentially in the sense that they can leverage the knowledge of the action of the previous players. The idea of using (Min, Max) players instead of (Max, Min) ones was already introduced in Degenne et al. [2019]. For example, in Section 3.2 of Degenne et al. [2019], the authors suggest to use the Follow-The-Perturbed-Leader for the Min player, followed by a best response for the Max player. It is worth noting that the (Max, Min) players approach has received more attention in subsequent works than their (Min, Max) players counterpart.

8.3.1 A Case Study: LεTT Algorithm

Even though it is appealing to study the Structured Top Two approach with a unified analysis as done in Chapter 2, a natural first step is to consider a specific instance. Due to its empirical success (see Section 8.4) and its link with the EB-TC$_\varepsilon$-IDS algorithm (analyzed in Chapter 5), we consider the LεTT algorithm in the deterministic setting.

Let $|\mathcal{Z}_\varepsilon(\theta)| = Z_\varepsilon$. The optimization problem $T_\varepsilon(\nu)^{-1}$ can be written as

$$\begin{aligned} T_\varepsilon(\nu)^{-1} &= \max_{p \in \Sigma_{Z_\varepsilon}} \sum_{z \in \mathcal{Z}_\varepsilon(\theta)} p(z) \left(\min_{\tilde{q}(z) \in \Sigma_{Z-1}} \max_{\tilde{w}(z) \in \Sigma_K} \sum_{x \in \mathcal{Z} \setminus \{z\}} \tilde{q}(z,x) C_\varepsilon(z,x;\nu,\tilde{w}(z)) \right) \\ &= \max_{p \in \Sigma_{Z_\varepsilon}} \min_{q \in (\Sigma_{Z-1})^{Z_\varepsilon}} \max_{\tilde{w} \in (\Sigma_K)^{Z_\varepsilon}} \sum_{z \in \mathcal{Z}_\varepsilon(\theta)} \sum_{x \in \mathcal{Z} \setminus \{z\}} p(z) \tilde{q}(z,x) C_\varepsilon(z,x;\nu,\tilde{w}(z)) \,, \end{aligned} \tag{8.10}$$

where we used Sion's lemma to swap the inner max-min in the first equality. Therefore, solving $T_\varepsilon(\nu)^{-1}$ is equivalent to finding a saddle-point $(p, \tilde{q}, \tilde{w}) \in \Sigma_{Z_\varepsilon} \times (\Sigma_{Z-1})^{Z_\varepsilon} \times (\Sigma_K)^{Z_\varepsilon}$ of (8.10).

The LεTT algorithm attempts to find a saddle-point by learning $(p_n, \tilde{q}_n, \tilde{w}_n)$ sequentially. For $z \in \mathcal{Z}_\varepsilon(\theta)$, we define $p_n(z) = T_n(z)/(n-1)$ with $T_n(z) = \sum_{x \neq z} T_n(z,x)$ hence $p_n \in \Sigma_{Z_\varepsilon}$; for all $x \neq z$, $\tilde{q}_n(z,x) = T_n(z,x)/(n-1)$; and for all $a \in \mathcal{A}$, $\tilde{w}_n(z,a) = N_n(z,a)/(n-1)$ with $N_n(z,a) = \sum_{x \neq z} N_{n,a}(z,x)$ and $\overline{w}_n(z,a) = \frac{1}{n-1} \sum_{t \in [n-1]} \mathbb{1}(B_t = z)\, \beta_{t,a}(z, C_t)$. Using Lemma 8.6, we have $\max_{z \in \mathcal{Z}_\varepsilon(\theta)} \max_{a \in \mathcal{A}} |\tilde{w}_n(z,a) - \overline{w}_n(z,a)| = \mathcal{O}(1)$. Due to their closeness, the convergence of $(p_n, \tilde{q}_n, \overline{w}_n)$ towards a saddle point should imply the one of $(p_n, \tilde{q}_n, \tilde{w}_n)$. Intuitively, $(p_n, \tilde{q}_n, \overline{w}_n)$ can be viewed as a *$\mathcal{A}$-continuous* version of $(p_n, \tilde{q}_n, \tilde{w}_n)$, where fractions of samples can be allocated to arms. Even though the algorithm is learning a vector of allocations over arms $\tilde{w} \in (\Sigma_K)^{Z_\varepsilon}$, it is effectively playing an allocation over arms $w \in \Sigma_K$. Let us define the *$\mathcal{A}$-continuous* version of the empirical allocation $N_n/(n-1)$ as $w_{n,a} = \frac{1}{n-1} \sum_{t \in [n-1]} \beta_{t,a}(B_t, C_t)$ for all $a \in \mathcal{A}$ hence $w_n \in \Sigma_K$. Using Lemma 8.6, we have $\|(n-1)w_n - N_n\|_\infty = \mathcal{O}(1)$. For simplicity, we consider the *$\mathcal{A}$-continuous* learning algorithm instead of its original version, which could be referred as *$\mathcal{A}$-discrete* since only one arm can be pulled at each time. Therefore, we replace $N_n/(n-1)$ by w_n in the algorithm. Similarly, we marginalize over the leader answer, *i.e.* $q_n(x) = \sum_{z \in \mathcal{Z}_\varepsilon(\theta)} \tilde{q}_n(z,x)$ for all $x \in \mathcal{Z}$ hence $q_n \in \Sigma_Z$.

At this point, we only used that the LεTT algorithm use tracking as in (8.9) to consider an $\mathcal{A}$-continuous optimization problem. As initialization, we have $p_1 = 1_{Z_\varepsilon}/Z_\varepsilon$, $q_1 = 1_Z/Z$ and $w_1 = 1_K/K$. In $\Sigma_{Z_\varepsilon} \times \Sigma_Z \times \Sigma_K$, the update step can be written as

$$\begin{bmatrix} p_{n+1} \\ q_{n+1} \\ w_{n+1} \end{bmatrix} = \left(1 - \frac{1}{n}\right) \begin{bmatrix} p_n \\ q_n \\ w_n \end{bmatrix} + \frac{1}{n} \begin{bmatrix} 1_{\{B_n\}} \\ 1_{\{C_n\}} \\ \beta_n(C_n, B_n) \end{bmatrix} .$$

The Outer-Max player selects the IF_ε leader and the Min player uses the TC_ε challenger, *i.e.* $B_n = \arg\max_{z \in \mathcal{Z}_\varepsilon(\theta)} \min_{x \in \mathcal{Z} \setminus \{z\}} C_\varepsilon(z, x; \nu, w_n)$ and *i.e.* $C_n = \arg\min_{x \in \mathcal{Z} \setminus \{B_n\}} C_\varepsilon(B_n, x; \nu, w_n)$. Given the current allocation w_n , this is a Frank-Wolfe step for the Outer-Max player and a best-response step for the Min player. The Inner-Max player considers the optimal design IDS, *i.e.*

$$\beta_n(B_n, C_n) = \frac{1}{C_\varepsilon(B_n, C_n; \nu_n, w_n)} w_n \odot \nabla_w C_\varepsilon(B_n, C_n; \nu, w_n) \,,$$

where $\odot$ denotes the coordinate-wise multiplication, *i.e.* $x \odot y = (x_a y_a)_{a \in \mathcal{A}}$. Given the current allocation w_n and the played leader/challenger pair (B_n, C_n) , this corresponds to a gradient step. More precisely, it is a convex combination between the current allocation w_n and a reweighted gradient, which is normalized by the empirical transportation cost.

Open problems In the deterministic setting, it is still an open problem to show that the above online optimization algorithm converges towards a saddle-point of $T_\varepsilon(\nu)^{-1}$. One could also attempt to study the special case where the Outer-Max player is virtually removed, either by assuming that we have access to the F_ε leader answer as $\mathcal{Z}$-oracle or by considering $\varepsilon = 0$ in which case $\mathcal{Z}_\varepsilon(\theta) = \{z^\star\}$. It is also an open problem to analyze either of those simpler cases which aims at solving the optimization problem $T_\varepsilon(\nu, z)^{-1}$ sequentially.

Dealing with stochastic observations Once those open problems are solved in the deterministic setting, one should still cope for the randomness of the observations if we want to analyze LεTT. Adding a forced exploration step (*i.e.* mixing with the uniform allocation) will yield an asymptotic analysis, *i.e.* we virtually know θ for n large enough. However, in the Top Two approach, we showed that the forced exploration is unnecessary. Adding it amounts to worsening the performance of the algorithm to ease the analysis. The promise of the Structured Top Two approach is that the combined choice of the leader, challenger and target are enough to ensure (1) sufficient exploration and (2) convergence towards an optimal allocation. While we believe that the slack $\varepsilon > 0$ would be enough to foster implicit exploration, it is also possible to rely on other approaches based on optimism or randomization.

8.4 Experiments

We show that LεTT has competitive empirical performance compared to existing ε-BAI algorithms on hard and random instances. The experimental setup is similar to the ones in Sections 5.5 and 7.5, hence we refer to Chapters 5 and 7 for more details. In particular, we consider the heuristic LεBAI which uses the IF_ε leader as $\mathcal{Z}$-oracle instead of the F_ε leader (see Section 7.5). We consider linear bandits ($\mathcal{A} = \mathcal{Z}$) and $\delta = 0.01$. Our results are averaged over 5000 runs.

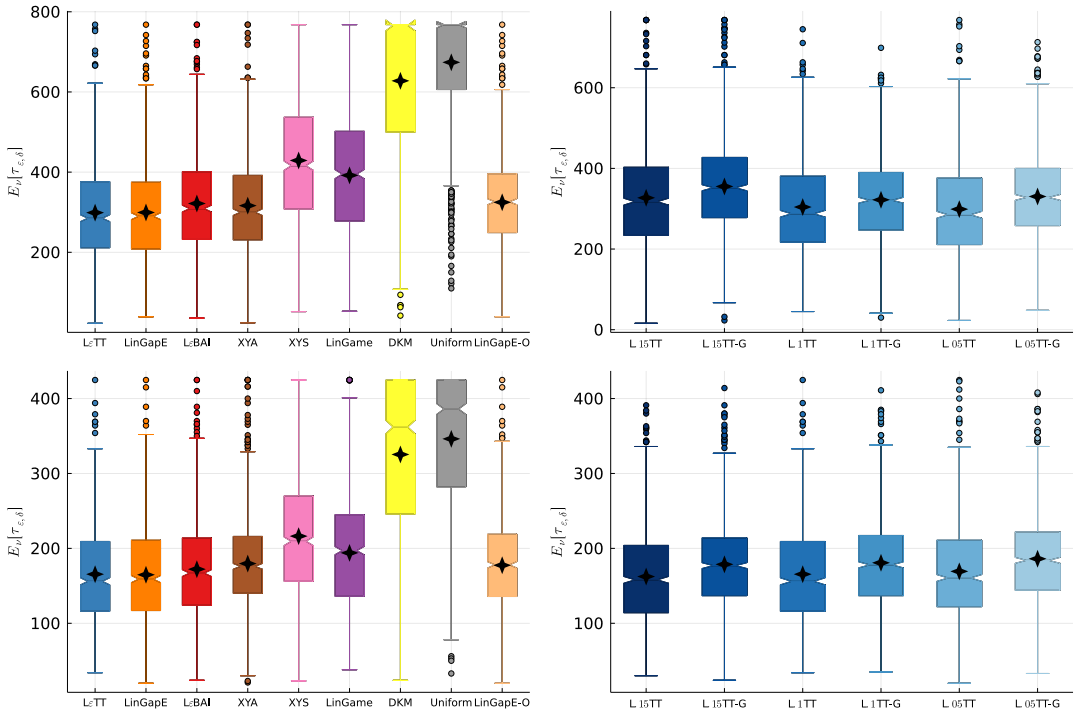

Figure 8.1 – Empirical stopping time on the hard instance ($\mathcal{A} = \mathcal{Z}$) for (top) $\varepsilon = 0.05$ and (bottom) $\varepsilon = 0.1$. All the algorithms use the GLR_ε stopping rule (8.2). "-G" denotes when the leader/challenger $(B_n^{\text{EB}}, C_n^{\text{TC}_{\varepsilon_0}})$ is used both for the sampling and the stopping rules. On the right, Lε_0TT with $\varepsilon_0 \in \{0.15, 0.1, 0.05\}$. "-O" denotes LinGapE with its original stopping rule.

Hard instances As in Section 7.3.3, we use hard instances with multiple correct answers, *i.e.* $|\mathcal{Z}_\varepsilon(\theta)| > 1$. Taking $\theta = e_1$ with $e_i = (\mathbb{1}(j = i))_{j\in[d]}$, the answers set is defined as $\mathcal{Z} = \{e_1, \cdots, e_d, a_{d+1}, a_{d+2}\}$ where $a_{d+1} = \cos(\phi_1)e_1 + \sin(\phi_1)e_2 \in \mathcal{Z}_\varepsilon(\theta)$ and $a_{d+2} = \cos(\phi_2)e_1 + \sin(\phi_2)e_2 \notin \mathcal{Z}_\varepsilon(\theta)$. Considering $d = 2$, we use $\phi_1 = r_\varepsilon\theta_\varepsilon$ and $\phi_2 = (1 + r_\varepsilon)\theta_\varepsilon$ with $\theta_\varepsilon = \arccos(1 - \varepsilon)$ and $r_\varepsilon = 0.1$.

In Figure 8.1, we see that LεTT performs on par with LinGapE and LεBAI, and that it outperforms other algorithms. Moreover, we observe that the performance are not too sensitive to the slack ε_0 used by the the $\text{TC}_{\varepsilon_0}$ challenger for ε-BAI. As expected, we observe better performance when using $(B_n^{\text{IF}_{\varepsilon_0}}, C_n^{\text{TC}_{\varepsilon_0}})$ instead of $(B_n^{\text{EB}}, C_n^{\text{TC}_{\varepsilon_0}})$. In Table 8.1, we observe that LεTT sample the direction a_2 the most. In particular, it allocates more samples to (a_3, a_4) than LinGapE which prefers to collect observations from a_1 .

Random instances As in Section 7.5, we use random instances to assess the impact of higher dimensions. For the answer set, 19 vectors $(a_k)_{k\in[19]}$ are uniformly drawn from $\mathbb{S}^{d-1} := \left\{a \in \mathbb{R}^d : \|a\|_2 = 1\right\}$ and set $\theta = a_1$. To enforce multiple correct answers, a modification

Table 8.1 – Average number of pulls per arm and empirical stopping time ($\pm\,\sigma$) on the hard instance ($\mathcal{A} = \mathcal{Z}$) for $\varepsilon = 0.05$. All the algorithms use the GLR_ε stopping rule (8.2). "-O" denotes LinGapE with its original stopping rule.

	a_1	a_2	a_3	a_4	**Total**
LεTT	29	250	17	3	299 ($\pm$119)
L2εTT	35	243	22	4	304 ($\pm$120)
L3εTT	61	223	38	6	327 ($\pm$130)
LinGapE	48	250	1	1	299 ($\pm$119)
LinGapE-O	50	273	1	1	325 ($\pm$111)
LεBAI	77	229	13	3	322 ($\pm$126)
$\mathcal{XY}$-Adaptive	77	238	1	1	316 ($\pm$119)
$\mathcal{XY}$-Static	215	216	1	1	433 ($\pm$171)
LinGame	111	221	50	11	393 ($\pm$150)
DKM	170	219	167	170	725 ($\pm$286)
Uniform	212	212	211	211	845 ($\pm$332)

of the greedy answer is added such that $a_{20,i} = a_{1,i}$ for $i \neq i_0$ and $a_{20,i_0} = \frac{1-\|\theta\|_2^2+\mu_{i_0}^2-r_\varepsilon\varepsilon}{\theta_{i_0}}$ where $i_0 = \arg\min_{i\in[d]} \theta_i$ and $r_\varepsilon = 0.1$.

In Figure 8.2, we see that LεTT is performing on par with LinGapE. Notably, it outperforms the LεBAI. When the dimension d increases, we already observed in Figure 7.4 of Chapter 7 that the empirical performance of LεBAI was scaling poorly compared to LinGapE. Even in the vanilla BAI setting, the empirical performance of the game based approach is known to suffer from increased dimension (despite being asymptotically optimal), see *e.g.* DKM [Degenne et al., 2019] in Figure 2.2 of Chapter 2.

8.5 Discussion

In Chapter 8, we extended the Top Two approach to tackle structured bandits by proposing the Structured Top Two approach, which is also specified by four choices (leader answer, challenger answer, target allocation and mechanism to reach it). In ε-BAI for transductive linear bandits, we proposed the LεTT algorithm which recovers EB-TC$_\varepsilon$ for vanilla BAI (see Chapter 5). We drew a connection between LεTT and a saddle-point algorithm aiming to solve $T_\varepsilon(\nu)^{-1}$ sequentially. Even in the deterministic setting where there is no randomness, the analysis of this procedure is still an open problem. Our empirical study showcased the good empirical performance of LεTT.

At the moment, this is one of the most exciting research direction as regards the Top Two approach. Solving it would pave the way to the analysis of Top Two algorithms for other pure

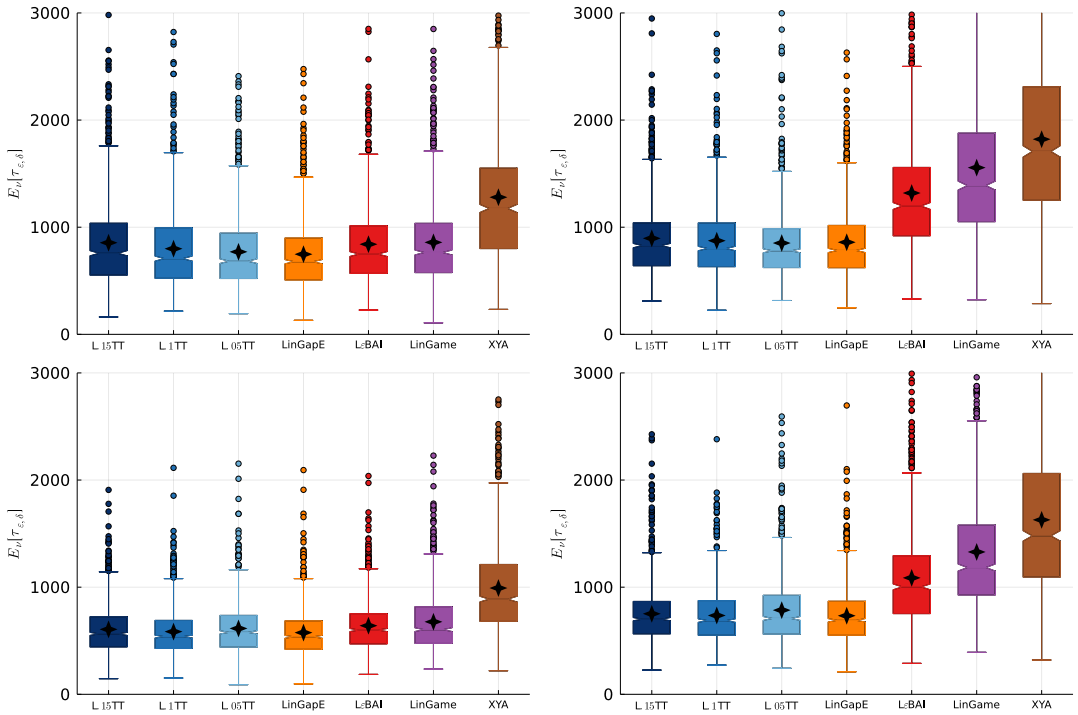

Figure 8.2 – Empirical stopping time on random instances ($\mathcal{A} = \mathcal{Z}$) with $d \in \{6, 12\}$ (from left to right) for (top) $\varepsilon = 0.05$ and (bottom) $\varepsilon = 0.1$. All the algorithms use the GLR_ε stopping rule (8.2). Lε_0TT with $\varepsilon_0 \in \{0.15, 0.1, 0.05\}$.

exploration problems and other types of structure, hence obtaining computationally efficient algorithms with good theoretical guarantees for any pure exploration problem.

Part IV

Conclusion and Appendices

Chapter 9

General Summary and Perspectives

9.1 Summary on our Contributions

In this thesis, we have made contributions to the field of pure exploration problems for stochastic multi-armed bandits. We endeavored to showcase the effectiveness of the Top Two approach in addressing these problems. In addition of its simplicity, interpretability, generalizability, and versatility, we have demonstrated that this principled methodology offers nearly optimal theoretical guarantees alongside state-of-the-art empirical performance.

In Part I, we delved into the fixed-confidence vanilla BAI setting. We proposed a unified perspective on the class of Top Two algorithms which puts forward four choices: leader answer, challenger answer, targeted allocation and the mechanism to reach it. We presented a unified asymptotic analysis of the Top Two approach, identifying desirable properties for each of these four choices, alongside a non-asymptotic analysis. We extended the algorithms and the asymptotic analysis to encompass other classes of distributions. For Gaussian distributions with unknown variance, we introduced and analyzed two approaches to handle unknown variances: plugging in the empirical variance or adapting the transportation costs. Notably, we derived new time-uniform concentration inequalities to calibrate our stopping rules. For the non-parametric class of bounded distributions, we proposed to adapt the transportation costs or to use the Dirichlet sampler, and derived new properties for the $\mathcal{K}_{\inf}$ functions and the Dirichlet sampler.

In Part II, we studied the impact of having multiple correct answers, and proved that algorithms can have good anytime guarantees. In ε-BAI, we introduced EB-TC$_\varepsilon$, an anytime sampling rule applicable without modification for fixed confidence or fixed budget identification (without prior knowledge of the budget). We established bounds on its expected sample complexity in the fixed confidence setting, notably demonstrating its asymptotic optimality when combined with an adaptively tuned exploration parameter. Additionally, we provided

upper bounds on its probability of error at any time and for any error parameter, which further yield upper bounds on its simple regret at any time. In GAI, we proposed APGAI, which simultaneously has anytime upper bounds on its probability of error and non-asymptotic upper bound on its expected sample complexity.

In Part III, we aimed to understand the influence of the structure in pure exploration problems with multiple correct answers. To this end, we considered ε-BAI for transductive linear bandits in the fixed-confidence setting. We emphasized the significance of the candidate answer choice and advocated for using the instantaneous furthest answer. We proposed a straightforward procedure to adapt existing BAI algorithms for ε-BAI, as well as an asymptotically optimal game-based algorithm. Lastly, we extended our unified perspective on the class of Top Two algorithms to address structured bandits. While the analysis of the Structured Top Two approach remains challenging, we highlighted some obstacles and linked them with the analysis of a saddle-point algorithm. Our empirical study underscored the favorable empirical performance of the LεTT algorithm, which recovers EB-TC$_\varepsilon$ for vanilla BAI.

9.2 Perspectives

Throughout this thesis, we have proposed potential avenues for future research. Below, we highlight three particularly promising directions.

Structured Top Two approach Chapter 8 introduced a unified Structured Top Two approach. Even within the confined scope of LεTT for transductive linear bandits in the deterministic setting, we encountered challenges in the analysis. Solving this simplified scenario is crucial for understanding how to study the Structured Top Two approach in linear settings, and it would pave the way to analyze the Structured Top Two approach in diverse pure exploration problems (*e.g.* , Top-k identification, Pareto set identification) and other types of structure (*e.g.* , combinatorial bandits, generalized linear bandits). The objective is to develop computationally efficient algorithms with robust theoretical guarantees for any pure exploration problem.

Anytime setting Thanks to its versatility, the anytime setting is a promising framework which captures the fact that constraints in decision-making scenarios can fluctuate unpredictably. The stream of recommendation can be leveraged by external actors on different downstream tasks, and the induced policies will inherit good theoretical properties. To derive anytime guarantees on existing algorithms, both their expected sample complexities and their probabilities of error have to be controlled. Better characterizing the Pareto front on the anytime performance with theoretical lower bounds would reveal the fundamental trade-off between achieving (i) a low sample complexity or a low probability of error, and be competitive in (ii) the asymptotic

regime (*i.e.* $\delta \to 0$ or $n \to +\infty$) or the moderate regime. Hopefully, this understanding will help deriving better anytime algorithms.

Privacy, safety and fairness Pure exploration problems are progressively used to model data-sensitive applications, such as adaptive clinical trials or hyper-parameters tuning. Given the privacy concerns inherent in these applications, it is imperative to enforce privacy constraints on algorithms. Although this manuscript did not delve into this area, we have proposed and examined differential privacy variants (global and local) of the fixed-confidence BAI problem. Extending these concepts to other pure exploration problems and structured bandits represents an exciting avenue for future research. Furthermore, pure exploration problems are relevant to applications involving safety and fairness constraints. Before human clinical trials, the safety of a drug must be rigorously assessed on multiple criteria. During human clinical trials, fast identification of promising drugs is essential while ensuring equitable allocation.

Appendix A

The Lambert W Function

The Lambert W function is implicitly defined by the equation $W(x)e^{W(x)} = x$. It defines two main branches W_{-1} (negative) and W_0 (positive).

- W_{-1} , defined on $[-e^{-1}, 0)$, is decreasing and $W_{-1}(-e^{-1}) = -1$.
- W_0 , defined on $[-e^{-1}, +\infty)$, is increasing and $W_0(-e^{-1}) = -1$.

The function W_0 satisfies for all $x \geq e$, $W_0\left(e^x\right) \leq x$ and

$$\frac{\log\log(x)}{2\log(x)} \leq W_0(x) - (\log(x) - \log\log(x)) \leq \frac{e}{e-1}\frac{\log\log(x)}{\log(x)} .$$

Lambert's branches are involved in the inversion of $h(x) = x - \log(x)$. When $x \geq 1$, it involves the negative branch. When $x \leq 1$, it involves the negative part of the positive branch. To make the notations clearer, we define for all $x \geq 1$

$$\overline{W}_{-1}(x) = -W_{-1}(-e^{-x}) \quad \text{and} \quad \overline{W}_0(x) = -W_0(-e^{-x}) . \tag{A.1}$$

Lemma A.1 gather useful properties on $\overline{W}_{-1}$ and $\overline{W}_0$ that we will use.

Lemma A.1. ***(1)*** *For $x \geq 1$, let $h(x) = x - \log(x)$. Then,*

$$\forall y \geq 1, \qquad y \leq h(x) \iff \begin{cases} \overline{W}_{-1}(y) \leq x & \text{if } x \geq 1 \\ \overline{W}_0(y) \geq x & \text{if } x \in (0,1] \end{cases} ,$$

$$\forall \delta > 0, \forall c > 0, \qquad e^{-c(h(x)-1)} \leq \delta \iff \begin{cases} \overline{W}_{-1}\left(1 + \frac{1}{c}\log\frac{1}{\delta}\right) \leq x & \text{if } x > 1 \\ \overline{W}_0\left(1 + \frac{1}{c}\log\frac{1}{\delta}\right) \geq x & \text{if } x \in (0,1) \end{cases} ,$$

$$\forall x > 1, \qquad \exp\left(-x + e^{-x}\right) \leq \overline{W}_0(x) \leq \exp\left(-x + e^{1-x}\right) ,$$

$$\forall x > 1, \qquad x + \log(x) \leq \overline{W}_{-1}(x) \leq x + \log(x) + \min\left\{\frac{1}{2}, \frac{1}{\sqrt{x}}\right\},$$
$$\forall u > 1, \forall t > 1, \qquad \overline{W}_0\left(1 + \frac{u}{t}\right) \geq \frac{1}{t} \iff t \geq \exp\left(1 + W_0\left(\frac{u-1}{e}\right)\right).$$

(2) *The function $\overline{W}_{-1}$ is increasing and strictly concave on $(1, +\infty)$. The function $\overline{W}_0$ is decreasing and strictly convex on $(1, +\infty)$. In particular,*

$$\forall x > 1, \qquad (\overline{W}_0)'(x) = \left(1 - \frac{1}{\overline{W}_0(x)}\right)^{-1} \text{ and } (\overline{W}_{-1})'(x) = \left(1 - \frac{1}{\overline{W}_{-1}(x)}\right)^{-1}.$$

Proof. **(1)** Let $y \geq 1$ and $x \in (0, 1]$. We obtain

$$\overline{W}_0(y) \geq x \iff W_0(-e^{-y}) \leq -x \iff -e^{-y} \leq -xe^{-x} \iff y \leq x - \log(x)$$

where the second equivalence uses that $-e^{-y} = W_0(-e^{-y})e^{W_0(-e^{-y})}$, $y \mapsto ye^y$ is increasing on $[-1, +\infty)$ and $W_0(x)$ has values on $[-1, 0)$ for $x \in [-e^{-1}, 0)$. Let $x \in (0, 1)$, $\delta, c > 0$. Then,

$$\overline{W}_0\left(1 + \frac{1}{c}\log\frac{1}{\delta}\right) \geq x \iff 1 + \frac{1}{c}\log\frac{1}{\delta} \leq h(x) \iff \exp\left(-c\left(h(x) - 1\right)\right) \leq \delta$$

Let $x > 1$ and $f(x) \in (0, 1)$. Then, we obtain

$$\overline{W}_0(x) \geq f(x) \iff x \leq f(x) - \log(f(x))$$

For $f(x) = e^{-x+e^{-x}}$, we have $x \leq f(x) - \log(f(x)) \iff e^{-x} \geq 0$, hence this condition holds and $\overline{W}_0(x) \geq f(x)$. For $f(x) = e^{-x+e^{1-x}}$, we have $x \leq f(x) - \log(f(x)) \iff x \leq 1$, hence this condition doesn't hold for $x > 1$, hence $\overline{W}_0(x) \leq f(x)$.

For $\overline{W}_{-1}(y)$, the same arguments yield the three results, which were first proven in Lemma A.1 and A.2 of Degenne [2019]).

We denote $v = \frac{u-1}{t} > 0$. Since $t > 1$, direct manipulations show that

$$\overline{W}_0\left(1 + \frac{u}{t}\right) \geq \frac{1}{t} \iff 1 + \frac{u}{t} \leq \frac{1}{t} - \log\left(\frac{1}{t}\right) \iff v + \log(v) \leq \log\left(\frac{u-1}{e}\right)$$
$$\iff ve^v \leq \frac{u-1}{e} \iff v \leq W_0\left(\frac{u-1}{e}\right) \iff t \geq \frac{u-1}{W_0\left(\frac{u-1}{e}\right)} = e^{1+W_0\left(\frac{u-1}{e}\right)}$$

The equivalence introducing W_0 uses that for $\alpha = \frac{u-1}{e} > 0$, $W_0(\alpha)e^{W_0(\alpha)} = \alpha$, $y \mapsto ye^y$ is increasing on $[-1, +\infty)$ and $v > 0$. The last equality uses that $e^{W_0(x)} = \frac{x}{W_0(x)}$.

(2) Let W denote W_0 or W_{-1} and $\overline{W}(x) = -W(-e^{-x})$. It is known (by implicit derivation) that $W'(z) = \frac{1}{z+e^{W(z)}}$ for $z \neq -e^{-1}$. Using that $e^{W(z)} = \frac{z}{W(z)}$, this yields that $zW'(z) = \left(1 + \frac{e^{W(z)}}{z}\right)^{-1} = \left(1 + \frac{1}{W(z)}\right)^{-1}$. For $x \neq 1$, using the above with $z = -e^{-x}$, we obtain

$$\overline{W}'(x) = -\frac{\mathrm{d}}{\mathrm{d}x}\left(W(-e^{-x})\right) = -e^{-x}W'(-e^{-x}) = \left(1 + \frac{1}{W(-e^{-x})}\right)^{-1} = \left(1 - \frac{1}{\overline{W}(x)}\right)^{-1}$$

Since $W_0(-e^{-x}) \in (-1, 0)$ for all $x > 1$ (positive branch on $(-e^{-1}, 0)$), we have $\overline{W}_0(x) \in (0, 1)$, hence $\overline{W}_0'(x) < 0$ for $x > 1$. Therefore, $\overline{W}_0$ is decreasing on $(1, +\infty]$. Using that $\overline{W}'(x) = \left(1 - \frac{1}{\overline{W}(x)}\right)^{-1}$ for $x \neq 1$, we obtain that $\overline{W}_0'$ is increasing on $(1, +\infty]$, hence strictly convex. The same arguments yield that $\overline{W}_{-1}$ is increasing and strictly concave on $(1, +\infty]$. ■

Lemma A.2 was proven in Degenne [2019]. It is needed when using the peeling method.

Lemma A.2 (Lemma A.3 in Degenne [2019])**.** *For $a, b \geq 1$, the minimal value of $f(\eta) = (1 + \eta)(a + \log(b + \frac{1}{\eta}))$ is attained at $\eta^\star$ such that $f(\eta^\star) \leq 1 - b + \overline{W}_{-1}(a + b)$. If $b = 1$, then there is equality.*

Appendix B

Complements on Chapter 2

B.1 Proof of Lemma 2.3

We build upon Theorem 9 of Kaufmann and Koolen [2021] which is restated below. While Theorem 9 was only stated for Gaussian distributions with variance $\sigma^2 = 1$, it is direct to notice that the result also holds for σ-sub-Gaussian distributions as mentioned by the authors.

Lemma B.1 (Theorem 9 of Kaufmann and Koolen [2021]). *Consider a σ-sub-Gaussian bandit ν with means $\mu \in \mathbb{R}^K$. Let $S \subseteq [K]$ and $x > 0$.*

$$\mathbb{P}_\nu \left(\exists n \in \mathbb{N},\ \sum_{k \in S} \frac{N_{n,k}}{2\sigma^2} (\mu_{n,k} - \mu_k)^2 > \sum_{k \in S} 2 \log \left(4 + \log \left(N_{n,k}\right)\right) + |S| \mathcal{C}_G \left(\frac{x}{|S|} \right) \right) \leq e^{-x}$$

where $\mathcal{C}_G$ is defined in Kaufmann and Koolen [2021] by $\mathcal{C}_G(x) = \min_{\lambda \in]1/2,1]} \frac{g_G(\lambda)+x}{\lambda}$ and

$$g_G(\lambda) = 2\lambda - 2\lambda \log(4\lambda) + \log \zeta(2\lambda) - \frac{1}{2} \log(1-\lambda) \,, \tag{B.1}$$

where ζ is the Riemann ζ function and $\mathcal{C}_G(x) \approx x + \log(x)$.

Using the computation from Section 1.4.2, for Gaussian with unit variance, we have

$$\{\tau_\delta < +\infty\} \cap \mathcal{E}_\mu^{\mathrm{err}}(\tau_\delta) \subseteq \bigcup_{n \in \mathbb{N}} \bigcup_{i \neq i^\star} \{N_{n,i}(\mu_{n,i} - \mu_i)^2 + N_{n,i^\star}(\mu_{n,i^\star} - \mu_{i^\star})^2 > 2c(n-1,\delta)\} \,.$$

By concavity of $x \to \log(4+\log x)$, we have $\sum_{k \in \{i^\star, i\}} \log\left(4 + \log N_{n,i}\right) \leq 2\log(4+\log((n-1)/2))$ for all $i \neq i^\star$ and all $n > K$. Using Lemma B.1 and a union bound over $i \neq i^\star$ yields $\mathbb{P}_\nu \left(\{\tau_\delta < +\infty\} \cap \mathcal{E}_\mu^{\mathrm{err}}(\tau_\delta) \right) \leq \delta$.

B.2 Proof of Lemma 2.4

Let $h_{i^\star,j}(\nu, w, u) = w_{i^\star}\mathcal{K}_{\inf}^{-}(\nu_{i^\star}, u) + w_j\mathcal{K}_{\inf}^{+}(\nu_j, u)$ and $u_{i^\star,j}(\nu, w) \in \arg\min_{u\in\mathbb{R}} h_{i^\star,j}(\nu, w, u)$. The optimality condition yields $\frac{\partial h_{i^\star,j}}{\partial u}(\nu, w, u_{i^\star,j}(\nu, w)) = 0$. By differentiating $C(i^\star, j; \nu, w) = h_{i^\star,j}(\nu, w, u_{i^\star,j}(w))$, we obtain

$$\frac{\partial C(i^\star, j; \nu, w)}{\partial w_{i^\star}} = \frac{\partial h_{i^\star,j}}{\partial w_{i^\star}}(\nu, w, u_{i^\star,j}(w)) + \frac{\partial h_{i^\star,j}}{\partial u}(\nu, w, u_{i^\star,j}(w))\frac{\partial u_{i^\star,j}}{\partial w_{i^\star}}(i^\star, i; \nu, w) ,$$
$$\frac{\partial C(i^\star, j; \nu, w)}{\partial w_j} = \frac{\partial h_{i^\star,j}}{\partial w_j}(\nu, w, u_{i^\star,j}(w)) + \frac{\partial h_{i^\star,j}}{\partial u}(\nu, w, u_{i^\star,j}(w))\frac{\partial u_{i^\star,j}}{\partial w_j}(i^\star, i; \nu, w) .$$

Therefore, we have

$$\frac{\partial C(i^\star, j; \nu, w)}{\partial w_{i^\star}} = \mathcal{K}_{\inf}^{-}(\nu_{i^\star}, u_{i^\star,j}(\nu, w)) \quad \text{and} \quad \frac{\partial C(i^\star, j; \nu, w)}{\partial w_j} = \mathcal{K}_{\inf}^{+}(\nu_j, u_{i^\star,j}(\nu, w)) .$$

B.3 Proof of Lemma 2.5

Since Slater's condition holds, the KKT conditions are necessary and sufficient for global optimality. Let $\lambda \geq 0$, $\alpha \in \mathbb{R}_+^K$ and $\gamma \in \mathbb{R}_+^{K-1}$ be the dual variables for the Lagrangian

$$\mathcal{L}(\phi, w; \lambda, \alpha, \gamma) = \phi + \lambda\left(\sum_{i\in[K]} w_i - 1\right) - \sum_{i\in[K]} \alpha_i w_i + \sum_{i\neq i^\star} \gamma_i(\phi - C(i^\star, i; \nu, w)) .$$

Using the complementary slackness condition, we have $\gamma_i(\phi - C(i^\star, i; \nu, w)) = 0$ for all $i \neq i^\star$. Combining it with the stationarity condition for arm $i^\star$, we obtain

$$0 = \lambda - \alpha_{i^\star} - \sum_{i\neq i^\star} \gamma_i \frac{\partial C(i^\star, i; \nu, w)}{\partial w_{i^\star}} \quad \text{hence} \quad 0 = \frac{\lambda}{\phi} - \frac{\alpha_{i^\star}}{\phi} - \sum_{i\neq i^\star} \frac{\gamma_i}{C(i^\star, i; \nu, w)} \frac{\partial C(i^\star, i; \nu, w)}{\partial w_{i^\star}} .$$

Multiplying by $w_{i^\star}$ and using that $\alpha_{i^\star} w_{i^\star} = 0$ yields that $w_{i^\star}\frac{\lambda}{\phi} = \sum_{i\neq i^\star} \gamma_i \frac{w_{i^\star}\frac{\partial C(i^\star,i;\nu,w)}{\partial w_{i^\star}}}{C(i^\star,i;\nu,w)}$. Similarly, one can show that $0 = \lambda - \alpha_i - \gamma_i \frac{\partial C(i^\star,i;\nu,w)}{\partial w_i}$ for all $i \neq i^\star$. Multiplying by w_i , using that $\alpha_{i^\star} w_{i^\star} = 0$ and $w_i \frac{\partial C(i^\star,i;\nu,w)}{\partial w_i} = C(i^\star, i; \nu, w) - w_{i^\star}\frac{\partial C(i^\star,i;\nu,w)}{\partial w_{i^\star}}$, we obtain $w_i\frac{\lambda}{\phi} = \gamma_i\left(1 - \frac{w_{i^\star}\frac{\partial C(i^\star,i;\nu,w)}{\partial w_{i^\star}}}{C(i^\star,i;\nu,w)}\right)$ for all $i \neq i^\star$. By summing and using that $\sum_{i\in[K]} w_i = 1$ (since $\lambda > 0$), we have $\lambda = \phi\sum_{i\neq i^\star}\gamma_i$. By scaling, we can consider $\tilde{\gamma}_i = \gamma_i(\sum_{i\neq i^\star}\gamma_i)^{-1}$ for all $i \neq i^\star$, *i.e.* $\tilde{\gamma} \in \Sigma_{K-1}$.

This completes the proof for the necessity of the conditions. Those conditions are also sufficient since it is direct to construct dual variables such that the KKT conditions hold.

B.4 Proof of Lemma 2.6

The proof boils down to giving an explicit dual vector $\gamma \in \Sigma_{K-1}$ for Lemma 2.5 by using the KKT conditions. It crucially relies on the fact that the optimal allocation can be shown to be have dense support (*i.e.* $\min_{i\in[K]} w_i > 0$), hence $\alpha_i = 0$ for all $i \in [K]$. Therefore, we have $\lambda = \gamma_i \frac{\partial C(i^\star,i;\nu,w)}{\partial w_i}$ for all $i \neq i^\star$, and $\lambda = \sum_{i\neq i^\star} \gamma_i \frac{\partial C(i^\star,i;\nu,w)}{\partial w_{i^\star}}$. Plugging in the last equation the explicit formula of γ given by the first $Z-1$ equations allows to conclude the proof by using (2.12). This completes the proof for the necessity of the conditions. Those conditions are also sufficient since it is direct to construct dual variables such that the KKT conditions hold.

B.5 Proof of Lemma 2.12

Definition B.2 introduces the notion of *asymptotically tight* threshold. All the thresholds proposed in Kaufmann and Koolen [2021] for one-parameter exponential families are asymptotically tight, including the stopping threshold (2.3).

Definition B.2. *A threshold $c : \mathbb{N} \times (0,1] \to \mathbb{R}_+$ is said to be asymptotically tight if there exists $\alpha \in [0,1)$, $\delta_0 \in (0,1]$, functions $f, \overline{T} : (0,1] \to \mathbb{R}_+$ and C independent of δ satisfying: (1) for all $\delta \in (0,\delta_0]$ and $n \geq \overline{T}(\delta)$, then $c(n,\delta) \leq f(\delta) + Cn^\alpha$, (2) $\limsup_{\delta\to 0} f(\delta)/\log(1/\delta) \leq 1$ and (3) $\limsup_{\delta\to 0} \overline{T}(\delta)/\log(1/\delta) = 0$.*

We only sketch the proof of the first result involving $T^\star(\nu)$, since the same argument will yield the result involving $T^\star_\beta(\nu)$. Let $\gamma_\nu > 0$ such that: for all $\gamma \in (0,\gamma_\nu]$, $\mathbb{E}_\nu[T_\gamma(w^\star)] < +\infty$ with $T_\gamma(w)$ as in (2.21). Let $\overline{T}_\gamma(w) := \inf\{T \geq 1 \mid \forall n \geq T,\ \|N_n - w\|_\infty \leq \gamma\}$. Then, we have $\mathbb{E}_\nu[\overline{T}_\gamma(w^\star)] < +\infty$ for $\gamma \in (0,\gamma_\nu/K]$. Let $\zeta > 0$. By continuity of $(\kappa, w) \to C(i,j;\kappa,w)$ and $\kappa \to i^\star(m(\kappa))$, there exists $\gamma_\zeta > 0$ such that: $\max\{\|w - w^\star\|_\infty, \|m(\kappa) - \mu\|_\infty\} \leq \gamma_\zeta$ implies that $i^\star(m(\kappa)) = \{i^\star\}$ and $\min_{j\neq i^\star} C(i^\star, i;\kappa,w) \geq (1-\zeta)T^\star(\nu)^{-1}$. Since $\min_{i\in[K]} w^\star_i > 0$ and $\mathbb{E}_\nu[\overline{T}_\gamma(w^\star)] < +\infty$, arms are sampled linearly, hence the empirical means μ_n converges towards the true mean μ . Therefore, there exists $\widehat{T}_{\gamma,\zeta}(w^\star)$ with $\mathbb{E}[\widehat{T}_{\gamma,\zeta}(w^\star)] < +\infty$ such that

$$\forall n \geq \widehat{T}_{\gamma,\zeta}(w^\star), \quad \hat{\imath}_n = i^\star \quad \text{and} \quad \min_{j\neq \hat{\imath}_n} W_n(\hat{\imath}_n, i) \geq n(1-\zeta)T^\star(\nu)^{-1}\,.$$

Let $\alpha \in [0,1)$, $\delta_0 \in (0,1]$, functions $f, \bar{T} : (0,1] \to \mathbb{R}_+$ and C as in the definition of an asymptotically tight threshold (Definition B.2). Let $\delta \leq \delta_0$, $\xi \in (0,1)$ and $T \geq \max\{\widehat{T}_{\gamma,\zeta}(w^\star), \bar{T}(\delta)\}/\xi$. By definition of the GLR stopping rule (2.2) with an asymptotically tight threshold, we have

$$\min\{\tau_\delta, T\} \leq \xi T + \sum_{n=\xi T}^{T} \mathbb{1}(\tau_\delta > n) \leq \xi T + \sum_{n=\xi T}^{T} \mathbb{1}\left(\min_{j\neq i^\star} W_n(i^\star, j) \leq c(n-1,\delta)\right)$$

$$\leq \xi T + \sum_{n=\xi T}^{T} \mathbb{1}\left(n(1-\zeta)T^\star(\nu)^{-1} \leq f(\delta) + CT^\alpha\right)$$
$$\leq \xi T + T^\star(\nu)(1-\zeta)^{-1}(f(\delta) + CT^\alpha)\,.$$

Let $T_\zeta(\delta) = \inf\{T \geq 1 \mid T^\star(\nu)(1-\zeta)^{-1}(1-\xi)^{-1}(f(\delta)+CT^\alpha) \leq T\}$. Then, $\tau_\delta \leq T$ for all $T \geq \max\{T_\zeta(\delta), \widehat{T}_{\gamma,\zeta}(w^\star)/\xi, \bar{T}(\delta)/\xi\}$, hence $\mathbb{E}_\nu[\tau_\delta] \leq T_\zeta(\delta) + \mathbb{E}_\nu[T_{\gamma,\zeta}(w^\star)]/\xi + \bar{T}(\delta)/\xi$. Then,

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\log(1/\delta)} \leq \limsup_{\delta\to 0} \frac{T_\zeta(\delta)}{\log(1/\delta)} \leq \frac{T^\star(\nu)}{(1-\zeta)(1-\xi)}\,,$$

where the last inequality is a known inversion result (see *e.g.* Lemma 13 in Jourdan et al. [2022]). Letting ζ and ξ go to zero concludes the proof.

B.6 Proof of Lemma 2.17

We start by presenting the result when using randomization to reach the target allocation, then sketch the one when using tracking for a fully deterministic algorithm. The result has the same proof for both optimal design IDS and fixed design β since they both satisfies Lemma 2.13.

Randomized When using randomization, Lemma B.3 gives a strictly positive lower bound on the probability of sampling the least sampled arm in the effective leader/challenger pair.

Lemma B.3. *Assume that the target satisfies Lemma 2.13. Then, for all $n > K$, $\mathbb{P}_{|n}\left(I_n \in \arg\min_{i\in\{\widehat{B}_n,\widehat{C}_n\}} N_{n,i}\right) \geq \beta_{\min}/K^2$.*

Proof. Assume that $N_{n,\widehat{B}_n} \leq N_{n,\widehat{C}_n}$, hence $\beta_n(\widehat{B}_n,\widehat{C}_n) \geq \beta_{\min}$ by Lemma 2.13. Then, using that $\mathbb{P}(I_n = i)$ can be expressed as (2.17), we have

$$\mathbb{P}_{|n}(I_n = \widehat{B}_n) \geq \beta_n(\widehat{B}_n,\widehat{C}_n)\mathbb{P}_{|n}(B_n = \widehat{B}_n)\mathbb{P}_{|n}(C_n = \widehat{C}_n \mid B_n = \widehat{B}_n) \geq \frac{\beta_{\min}}{K(K-1)}\,,$$

where the last inequality is obtained by using the definition of $(\widehat{B}_n,\widehat{C}_n)$ in (2.23) as maximizer of probabilities that sum to one. The case $N_{n,\widehat{B}_n} > N_{n,\widehat{C}_n}$ can be done similarly. ■

Let $\beta_{\min}$, L_0, L_1 as in Lemma 2.13, Properties 2.15 and 2.16. Let L_2 such that $\lfloor L\rfloor \geq KL^{3/4}$ for all $L \geq L_2$. Let $L \geq \max\{L_0^{4/3}, L_1, L_2\}$.

Suppose towards contradiction that $U^L_{\lfloor KL\rfloor} \neq \emptyset$. For any $1 \leq n \leq \lfloor KL\rfloor$, we have $U^L_n \neq \emptyset$, hence $V^L_n \cap \{\widehat{B}_n,\widehat{C}_n\} \neq \emptyset$ by using Assumption 2.14 with Properties 2.15 and 2.16. Using

the pigeonhole principle, there exists some $i \in [K]$ such that $N_{\lfloor L \rfloor, i} \geq L^{3/4}$. Thus, we have $\left|V^L_{\lfloor L \rfloor}\right| \leq K-1$. Our goal is to show that $\left|V^L_{\lfloor 2L \rfloor}\right| \leq K-2$. A sufficient condition is that one arm in $V^L_{\lfloor L \rfloor}$ is pulled at least $L^{3/4}$ times between $\lfloor L \rfloor$ and $\lfloor 2L \rfloor - 1$. Using Lemma 2.10 and $V^L_n \cap \{\widehat{B}_n, \widehat{C}_n\} \subseteq V^L_{\lfloor L \rfloor}$ and , we obtain

$$\begin{aligned}\sum_{n=\lfloor L \rfloor}^{\lfloor 2L \rfloor - 1} \mathbb{1}\left(I_n \in V^L_{\lfloor L \rfloor}\right) &\geq \sum_{n=\lfloor L \rfloor}^{\lfloor 2L \rfloor - 1} \mathbb{P}_{|n}\left(I_n \in V^L_n \cap \{\widehat{B}_n, \widehat{C}_n\}\right) - 2W_K\sqrt{(\lfloor 2L \rfloor + 1)\log(e + \lfloor 2L \rfloor + 1)} \\ &\geq \beta_{\min}(\lfloor 2L \rfloor - \lfloor L \rfloor)/K^2 - 2W_K\sqrt{(\lfloor 2L \rfloor + 1)\log(e + \lfloor 2L \rfloor + 1)} \geq KL^{3/4}\,,\end{aligned}$$

where the second inequality uses that $\{I_n \in \arg\min_{i \in \{\widehat{B}_n, \widehat{C}_n\}} N_{n,i}\} \subseteq \{I_n \in V^L_n \cap \{\widehat{B}_n, \widehat{C}_n\}\}$ and Lemma B.3. The last inequality is obtained for $L \geq L_3 + 1$ with

$$L_3 = \sup\left\{L \in \mathbb{N} \mid \beta_{\min}(\lfloor 2L \rfloor - \lfloor L \rfloor)/K^2 - 2W_K\sqrt{(\lfloor 2L \rfloor + 1)\log(e + \lfloor 2L \rfloor + 1)} < KL^{3/4}\right\}\,,$$

which satisfies that $\mathbb{E}_\nu[L_3] < +\infty$ by Lemma 2.10 since it can be upper bounded by a polynomial function of W_K . Therefore, we have shown that $\left|V^L_{\lfloor 2L \rfloor}\right| \leq K-2$.

By induction, there exists L_4 with $\mathbb{E}_\nu[L_4] < +\infty$ such that, for all $L \geq L_4$, we have $\left|V^L_{\lfloor kL \rfloor}\right| \leq K-k$ for any $1 \leq k \leq K$, hence $U^L_{\lfloor KL \rfloor} = \emptyset$ for all $L \geq L_5 = \max\{L_0^{4/3}, L_1, L_2, L_3, L_4\}$. Defining $N_1 = KL_5$, we have $\mathbb{E}_\nu[N_1] < +\infty$. For all $n \geq N_1$, taking $L = n/K$ yields $U^{n/K}_n = U^L_{\lfloor KL \rfloor} = \emptyset$, hence $\min_{i \in [K]} N_{n,i} \geq \sqrt{n/K}$.

Combining Lemma 2.10 with the above result, we have

$$|\mu_{n,i} - \mu_i| \leq W_\mu(K/n)^{1/4}\sqrt{\log(e + \sqrt{n/K})} \leq \min_{i \neq i^\star}(\mu_{i^\star} - \mu_i)/4\,.$$

The last inequality holds for $n \geq N_2$ which also satisfies $\mathbb{E}_\nu[N_2] < +\infty$ since it is defined as

$$N_2 = \sup\{n > 1 \mid W_\mu(K/n)^{1/4}\sqrt{\log(e + \sqrt{n/K})} > \min_{i \neq i^\star}(\mu_{i^\star} - \mu_i)/4\}\,.$$

Taking $N_0 = \max\{N_1, N_2\}$ concludes the proof for randomization.

Deterministic When the algorithm is fully deterministic, we have $(B_n, C_n) = (\widehat{B}_n, \widehat{C}_n)$. Suppose that we have $\sum_{t=\lfloor L \rfloor}^{\lfloor 2L \rfloor - 1} \mathbb{1}\left(\{B_t, C_t\} \subseteq V^L_t\right) \geq KL^{3/4}$. Then, $\sum_{t=\lfloor L \rfloor}^{\lfloor 2L \rfloor - 1} \mathbb{1}\left(I_t \in V^L_{\lfloor L \rfloor}\right) \geq KL^{3/4}$. In the following, we consider that $\sum_{t=\lfloor L \rfloor}^{\lfloor 2L \rfloor - 1} \mathbb{1}\left(\{B_t, C_t\} \subseteq V^L_t\right) < KL^{3/4}$ does not hold, hence using that $B_t \in V^L_t$ or $C_t \in V^L_t$

$$\sum_{t=\lfloor L \rfloor}^{\lfloor 2L \rfloor - 1}\left(\mathbb{1}\left(B_t \in V^L_t, C_t \notin V^L_t\right) + \mathbb{1}\left(B_t \notin V^L_t, C_t \in V^L_t\right)\right) > \lfloor 2L \rfloor - \lfloor L \rfloor - KL^{3/4}\,.$$

We distinguish between two cases.

$$\textbf{Case 1:} \quad \sum_{t=\lfloor L \rfloor}^{\lfloor 2L \rfloor - 1} \mathbb{1}\left(B_t \in V_t^L, C_t \notin V_t^L\right) > \left(\lfloor 2L \rfloor - \lfloor L \rfloor - KL^{3/4}\right)/2\,,$$

$$\textbf{Case 2:} \quad \sum_{t=\lfloor L \rfloor}^{\lfloor 2L \rfloor - 1} \mathbb{1}\left(B_t \notin V_t^L, C_t \in V_t^L\right) > \left(\lfloor 2L \rfloor - \lfloor L \rfloor - KL^{3/4}\right)/2\,.$$

Using Lemma 2.13, we obtain $\beta_t(B_t, C_t) \geq \beta_{\min}$ when $B_t \in V_t^L, C_t \notin V_t^L$, and $1 - \beta_t(B_t, C_t) \geq \beta_{\min}$ when $B_t \notin V_t^L, C_t \in V_t^L$.

Case 1. Using Lemma 2.7 and the above, we obtain

$$\begin{aligned}
\sum_{t=\lfloor L \rfloor}^{\lfloor 2L \rfloor - 1} \mathbb{1}\left(I_t \in V_{\lfloor L \rfloor}^L\right) &\geq \sum_{i \in V_{\lfloor L \rfloor}^L} \sum_{j \neq i} \sum_{t=\lfloor L \rfloor}^{\lfloor 2L \rfloor - 1} \mathbb{1}\left(I_t = i,\ (B_t, C_t) = (i,j)\right) \\
&\geq \sum_{i \in V_{\lfloor L \rfloor}^L} \sum_{j \neq i} \sum_{t=\lfloor L \rfloor}^{\lfloor 2L \rfloor - 1} \beta_t(i,j) \mathbb{1}\left((B_t, C_t) = (i,j)\right) - K^2 \\
&\geq \sum_{t=\lfloor L \rfloor}^{\lfloor 2L \rfloor - 1} \beta_t(B_t, C_t) \mathbb{1}\left(B_t \in V_t^L,\ C_t \notin V_t^L\right) - K^2 \\
&\geq \beta_{\min}\left(\lfloor 2L \rfloor - \lfloor L \rfloor - KL^{3/4}\right)/2 - K^2 \geq KL^{3/4}\,,
\end{aligned}$$

where the last inequality is obtained for $L \geq L_6 + 1$ with

$$L_6 = \sup\left\{L \in \mathbb{N} \mid \beta_{\min}\left(\lfloor 2L \rfloor - \lfloor L \rfloor - KL^{3/4}\right)/2 - K^2 < KL^{3/4}\right\}\,.$$

Therefore, there exists $i \in V_{\lfloor L \rfloor}^L$ which is sampled $L^{3/4}$ times between $\lfloor L \rfloor$ and $\lfloor 2L \rfloor - 1$.

Case 2. Using Lemma 2.7 and the same argument as above, we obtain

$$\begin{aligned}
\sum_{t=\lfloor L \rfloor}^{\lfloor 2L \rfloor - 1} \mathbb{1}\left(I_t \in V_{\lfloor L \rfloor}^L\right) &\geq \sum_{t=\lfloor L \rfloor}^{\lfloor 2L \rfloor - 1} (1 - \beta_t(B_t, C_t)) \mathbb{1}\left(B_t \notin V_t^L,\ C_t \in V_t^L\right) - K^2 \\
&\geq \beta_{\min}\left(\lfloor 2L \rfloor - \lfloor L \rfloor - KL^{3/4}\right)/2 - K^2 \geq KL^{3/4}\,,
\end{aligned}$$

where the last inequality is obtained for $L \geq L_6 + 1$. Therefore, there exists $i \in V_{\lfloor L \rfloor}^L$ which is sampled $L^{3/4}$ times between $\lfloor L \rfloor$ and $\lfloor 2L \rfloor - 1$.

Therefore, we have shown that $\left|V_{\lfloor 2L \rfloor}^L\right| \leq K - 2$, and we conclude the proof similarly.

B.7 Proof of Lemma 2.19

We start by presenting the result when using randomization to reach the target allocation, then sketch the one when using tracking for a fully deterministic algorithm.

Randomized optimal design IDS Let N_0 and (g_1, N_1) as in (2.22) and Property 2.18. Let $M \geq \max\{N_0, N_1\}$ and $n > M$. Using Lemma 2.10, $\max_{i\in[K]} \mathbb{E}[N_{M,i}] \leq M$ and $\mathbb{P}_{|t}(B_t \neq i^\star) \leq g_1(t)$, we can show that

$$\left| N_{n,i} - \sum_{t=M}^{n-1} (1 - \beta_t(i^\star, i))\mathbb{P}_{|t}((B_t, C_t) = (i^\star, i)) \right| \leq M + \sum_{t=M}^{n-1} g_1(t) + W_K\sqrt{(n+1)\log(e+n)}$$
$$\left| N_{n,i^\star} - \sum_{t=M}^{n-1} \sum_{j\neq i^\star} \beta_t(i^\star, j)\mathbb{P}_{|t}((B_t, C_t) = (i^\star, j)) \right| \leq M + \sum_{t=M}^{n-1} g_1(t) + W_K\sqrt{(n+1)\log(e+n)} .$$

For all $n > M$, let $H_n = N_{n,i^\star}^2 - \sum_{j\neq i^\star} N_{n,j}^2$ and $G_n =$

$$\left(\sum_{t=M}^{n-1} \sum_{j\neq i^\star} \beta_t(i^\star, j)\mathbb{P}_{|t}((B_t, C_t) = (i^\star, j)) \right)^2 - \sum_{j\neq i^\star} \left(\sum_{t=M}^{n-1} (1 - \beta_t(i^\star, j))\mathbb{P}_{|t}((B_t, C_t) = (i^\star, j)) \right)^2$$

Since $|a^2 - b^2| \leq 2\max\{|a|, |b|\}|a - b|$, we obtain

$$|H_n| \leq |H_n - G_n| + |G_n| \leq |G_n| + 2K(n-1)\left(M + \sum_{t=M}^{n} g_1(t) + W_K\sqrt{(n+1)\log(e+n)} \right) .$$

Since $|G_n| \leq \sum_{t=M+1}^{n-1} |G_{t+1} - G_t| + |G_{M+1}|$ and $|G_{M+1}| \leq K$, we study the increments,

$$\begin{aligned}
\frac{G_{n+1} - G_n}{2} &= \left(\sum_{j\neq i^\star} \beta_n(i^\star, j)\mathbb{P}_{|n}((B_n, C_n) = (i^\star, j)) \right) \left(\sum_{t=M}^{n-1} \sum_{j\neq i^\star} \beta_t(i^\star, j)\mathbb{P}_{|t}((B_t, C_t) = (i^\star, j)) \right) \\
&- \sum_{j\neq i^\star} (1 - \beta_n(i^\star, j))\mathbb{P}_{|n}((B_n, C_n) = (i^\star, j)) \left(\sum_{t=M}^{n-1} (1 - \beta_t(i^\star, j))\mathbb{P}_{|t}((B_t, C_t) = (i^\star, j)) \right) \\
&\left(\sum_{j\neq i^\star} \beta_n(i^\star, j)\mathbb{P}_{|n}((B_n, C_n) = (i^\star, j)) \right)^2 /2 - \sum_{j\neq i^\star} \left((1 - \beta_n(i^\star, j))\mathbb{P}_{|n}((B_n, C_n) = (i^\star, j)) \right)^2 /2 .
\end{aligned}$$

The optimal design IDS ensures that $N_{n,i^\star}\beta_n(i^\star, j) = (1 - \beta_n(i^\star, j))N_{n,j}$ since $\mu_{n,i^\star} > \mu_{n,j}$ due to (2.22). Leveraging the above results, a direct upper bound yields

$$|G_{n+1} - G_n| \leq 4\left(M + \sum_{t=M}^{n-1} g_1(t) + W_K\sqrt{(n+1)\log(e+n)} \right) + K .$$

Therefore, we have shown that

$$|H_n| \leq 2(K+2)(n-1)\left(M + \sum_{t\in[n-1]} g_1(t) + W_K\sqrt{(n+1)\log(e+n)}\right) + Kn\,.$$

Let $\gamma > 0$. Therefore, this concludes the proof by taking $N_2 = X_0 + 1$ with

$$X_0 = \sup\left\{n \mid \frac{\gamma(n-1)}{2(K+2)} < \max\{N_0, N_1\} + 1 + \sum_{t\in[n-1]} g_1(t) + W_K\sqrt{(n+1)\log(e+n)}\right\}$$

which satisfies $\mathbb{E}_\nu[X_0] < +\infty$ since it is at most a linear function of $\max\{N_0, N_1\}$ and polynomial function of W_K (Lemma 2.10).

For fixed design β , the result is a direct consequence of the first upper bound on $N_{n,i^\star}$ and $\beta_t(i^\star, j) = \beta$ for all $M \leq t \leq n-1$, and $\sum_{t=M}^{n-1} \mathbb{P}_{|t}(B_t = i^\star) \geq n - \sum_{t=M}^{n-1} \mathbb{P}_{|t}(B_t \neq i^\star) - M$ by Property 2.18.

Deterministic optimal design IDS Since the algorithm is fully deterministic, Property 2.18 can be rewritten: for n large enough, $B_n = i^\star$. Therefore, using Lemma 2.7, we obtain

$$\max\left\{\max_{i\neq i^\star}\left|N_{n,i} - \sum_{t=M}^{n-1}(1-\beta_t(i^\star,i))\mathbb{1}\left(C_t = i\right)\right|, \left|N_{n,i^\star} - \sum_{t=M}^{n-1}\beta_t(i^\star, C_t)\right|\right\} \leq 2M + K\,.$$

As above, we can show that $|H_n|$ is small if $|\tilde{G}_{n+1} - \tilde{G}_n|$ is small, where

$$\tilde{G}_n = \left(\sum_{t=M}^{n-1}\beta_t(i^\star, C_t)\right)^2 - \sum_{j\neq i^\star}\left(\sum_{t=M}^{n-1}(1-\beta_t(i^\star,i))\mathbb{1}\left(C_t = i\right)\right)^2\,.$$

It is direct to show that $|\tilde{G}_{n+1} - \tilde{G}_n| \leq 2(2M + K + 3/2)$ by using that $N_{n,i^\star}\beta_n(i^\star, j) = (1-\beta_n(i^\star,j))N_{n,j}$. This allows to conclude the proof similarly.

For fixed design β , the result is a direct consequence of the first upper bound on $N_{n,i^\star}$ and $\beta_t(i^\star, C_t) = \beta$ for all $M \leq t \leq n-1$.

B.8 Proof of Lemma 2.20

Using Lemma 2.19 for $\gamma \in \{1/2, (32(K-1))^{-1}\}$ yields the upper/lower bound.

$$1/2 \geq \left(\frac{N_{n,i^\star}}{n-1}\right)^2 - \sum_{i\neq i^\star}\left(\frac{N_{n,i}}{n-1}\right)^2 \geq \left(\frac{N_{n,i^\star}}{n-1}\right)^2 - \left(1 - \frac{N_{n,i^\star}}{n-1}\right)^2 = 2\frac{N_{n,i^\star}}{n-1} - 1\,,$$

$$\left(\frac{N_{n,i^\star}}{n-1}\right)^2 \geq -\frac{1}{32(K-1)} + \frac{1}{K-1}\left(1-\frac{N_{n,i^\star}}{n-1}\right)^2 \geq \frac{1}{32(K-1)} \,.$$

Then, using Lemma 2.19 for $\frac{\gamma}{32(K-1)}$ and $N_{n,i^\star}/(n-1) \geq (4\sqrt{2(K-1)})^{-1}$,

$$\left|1-\sum_{i\neq i^\star}\left(\frac{N_{n,i}}{N_{n,i^\star}}\right)^2\right| = \left(\frac{n-1}{N_{n,i^\star}}\right)^2 \left|\left(\frac{N_{n,i^\star}}{n-1}\right)^2 - \sum_{i\neq i^\star}\left(\frac{N_{n,i}}{n-1}\right)^2\right| \leq 32(K-1)\frac{\gamma}{32(K-1)} = \gamma \,.$$

B.9 Proof of Lemma 2.22

We only sketch the proof of the first result involving $T_\gamma(w^\star)$, since the same argument yields the one for $T_\gamma(w^\star_\beta)$. We start by presenting the result when using randomization to reach the target allocation, then sketch the one when using tracking for a fully deterministic algorithm.

Randomized optimal design IDS Let N_0 , (g_1, N_1) , N_2 , (γ_0, g_3, N_3) and N_4 as in (2.22), Property 2.18, Lemma 2.19, Property 2.21 and Lemma 2.20. Let $M \geq \max\{N_0, N_1, N_2, N_3, N_4\}$, $\gamma \in (0, \gamma_0]$, $n > M$

$$\forall i \neq i^\star, \quad t_{n,i}(\gamma) = \max\left\{M, \max\left\{t \in \{M, \cdots, n-1\} \mid N_{t,i}/N_{t,i^\star} < w^\star_i/w^\star_{i^\star} + \gamma/4\right\}\right\} \,.$$

Combining (2.17), Properties 2.18 and 2.21 and the definition of $t_{n,i}(\gamma)$, one can show that

$$\begin{aligned}
\mathbb{E}[N_{n,i}] &\leq \mathbb{E}[N_{t_{n,i}(\gamma),i}] + \sum_{t=t_{n,i}(\gamma)}^{n-1} \mathbb{P}_{|t}(B_t \neq i^\star) + \sum_{t=t_{n,i}(\gamma)}^{n-1} \mathbb{P}_{|t}(C_t = i \mid B_t = i^\star) \\
&\leq \mathbb{E}[N_{t_{n,i}(\gamma),i^\star}] \max\left\{M, \left(\frac{w^\star_i}{w^\star_{i^\star}} + \gamma/4\right)\right\} + \sum_{t\in[n-1]} g_1(t) + \sum_{t\in[n-1]} g_3(t) \,.
\end{aligned}$$

Using $\|\mathbb{E}[N_n] - N_n\|_\infty \leq W_K\sqrt{n+1\log(e+n)}$ (Lemma 2.10) and $N_{n,i^\star}/n \geq \left(4\sqrt{2(K-1)}\right)^{-1}$ (Lemma 2.20), we can show that

$$\exists N_5 \text{ s.t. } \mathbb{E}_\nu[N_5] < +\infty, \ \forall n \geq N_5, \quad N_{n,i}/N_{n,i^\star} \leq w^\star_i/w^\star_{i^\star} + \gamma/2 \,.$$

Using this result with $\left|1-\sum_{i\neq i^\star}\left(N_{n,i}/N_{n,i^\star}\right)^2\right| \leq \gamma/4$ (Lemma 2.20), we can show that

$$\exists N_6 \text{ s.t. } \mathbb{E}_\nu[N_6] < +\infty, \ \forall n \geq N_6, \quad N_{n,i}/N_{n,i^\star} \geq w^\star_i/w^\star_{i^\star} - \gamma \,.$$

Therefore, we have $T_\gamma(w^\star) \leq \max\{N_0, N_1, N_2, N_3, N_4, N_5, N_6\}$, hence $\mathbb{E}_\nu[T_\gamma(w^\star)] < +\infty$.

Deterministic optimal design IDS Since the algorithm is fully deterministic, Properties 2.18 and 2.21 can be rewritten: for n large enough, $B_n = i^\star$ and, for arm i which is overshooting the ratio optimal ratio, we have $C_n \neq i$. Therefore, we obtain directly that

$$N_{n,i} \leq N_{t_{n,i}(\gamma),i} \leq N_{t_{n,i}(\gamma),i^\star} \max\left\{M, \left(\frac{w_i^\star}{w_{i^\star}^\star} + \gamma/4\right)\right\} \leq N_{n,i^\star} \max\left\{M, \left(\frac{w_i^\star}{w_{i^\star}^\star} + \gamma/4\right)\right\} .$$

Therefore, we can conclude similarly.

B.10 Proof of Lemma 2.23

Let S_n^L and $\mathcal{I}_n^\star$ as in (2.24) such that $S_n^L \setminus \mathcal{I}_n^\star \neq \emptyset$. When $S_n^L \setminus \mathcal{I}_n^\star = \emptyset$, the statement is true. Suppose that $S_n^L \setminus \mathcal{I}_n^\star \neq \emptyset$. Let $\gamma > 0$. Using Lemma 2.10, there exists $L_4 = Poly(W_\mu)$ (hence $\mathbb{E}_\nu[(L_4)^\alpha] < +\infty$ for all $\alpha > 0$) such that for all $L \geq L_4$ and all $k \in S_n^L$, $|\mu_{n,k} - \mu_k| \leq \gamma$. Let $L \geq L_4$ and $(i,j) \in \mathcal{I}_n^\star \times \left(S_n^L \setminus \mathcal{I}_n^\star\right)$. By definition of $W_n(i,j)$ in (2.1), we obtain

$$W_n(i,j) \geq LC(i,j;\nu_n,1_K) \geq LC_{\nu,\gamma} \text{ with } C_{\nu,\gamma} = \min_{(i,j):\mu_i > \mu_j} \inf_{\kappa:\max_{k\in\{i,j\}} |m(\kappa)_k - \mu_k| \leq \gamma} C(i,j;\kappa,1_K) ,$$

where the last inequality takes the infimum. Since $\kappa \to C(i,j;\kappa,1_K)$ is continuous with strictly positive value for (i,j) such that $m(\kappa)_i > m(\kappa)_j$, there exists γ such that $C_{\nu,\gamma} > 0$.

B.11 Proof of Lemma 2.24

Let $(i,j) \in S_n^L \times \overline{S_n^L}$. By definition and taking $u = \mu_{n,i}$ yields

$$W_n(i,j) \leq N_{n,j} \mathcal{K}_{\inf}^+(\nu_{n,j}, \mu_{n,i}) \leq L\mathcal{K}_{\inf}^+(\nu_{n,j}, \mu_{n,i}) .$$

For Gaussian distributions with unit variance, we have

$$\sqrt{2\mathcal{K}_{\inf}^+(\nu_{n,j}, \mu_{n,i})} = \mu_{n,i} - \mu_{n,j} \leq \mu_i - \mu_j + 2W_\mu \sqrt{\log(e+1)} ,$$

where the inequality uses Lemma 2.10 and $x \to \sqrt{\log(e+x)/x}$ is decreasing.

B.12 EB Leader

Lemma B.4. *The EB leader satisfies Properties 2.15 and 2.18.*

Proof. Let L_4 as in Lemma 2.23, and $L \geq L_4$. Let S_n^L and $\mathcal{I}_n^\star$ as in (2.24) such that $S_n^L \setminus \mathcal{I}_n^\star \neq \emptyset$. When $S_n^L \setminus \mathcal{I}_n^\star = \emptyset$, the statement is true. Suppose that $S_n^L \setminus \mathcal{I}_n^\star \neq \emptyset$ and $\widehat{B}_n \in S_n^L$. Then, $W_n(i,j) \geq LC_\nu$ for all $(i,j) \in \mathcal{I}_n^\star \times \left(S_n^L \setminus \mathcal{I}_n^\star\right)$. Suppose towards contradiction that $\widehat{B}_n^{\mathrm{EB}} \notin \mathcal{I}_n^\star$. Therefore, $W_n(i, \widehat{B}_n^{\mathrm{EB}}) \geq LC_\nu > 0$ for all $i \in \mathcal{I}_n^\star$. Since the leader is deterministic, we have $B_n^{\mathrm{EB}} = \widehat{B}_n^{\mathrm{EB}}$. Since $B_n^{\mathrm{EB}} \in \arg\max_{i \in [K]} \mu_{n,i}$, we have $W_n(i, \widehat{B}_n^{\mathrm{EB}}) = 0$. This is a contradiction, hence $\widehat{B}_n^{\mathrm{EB}} \in \mathcal{I}_n^\star$.

Let N_0 as in (2.22), and $n \geq N_0$. Then, $\mathbb{P}_{|n}(B_n^{\mathrm{EB}} \neq i^\star) \leq \mathbb{P}_{|n}(i^\star(\mu_n) \neq \{i^\star\}) = 0$. ■

B.13 TC Challenger

Lemma B.5. *The TC challenger satisfies Property 2.16.*

Proof. Let $\widehat{B}_n$ be an effective leader satisfying Property 2.15. Let L_0, L_4 and L_5 as in Property 2.15, Lemmas 2.23 and 2.24, and $L \geq \max\{L_0^{4/3}, L_4^{4/3}, L_5^2\}$. Let n such that $U_n^L \neq \emptyset$ and $\widehat{B}_n \in V_n^L$, hence $\widehat{B}_n \in \mathcal{J}_n^\star = \arg\max_{i \in \overline{V_n^L}} \mu_i$. When $\widehat{C}_n^{\mathrm{TC}} \in \mathcal{J}_n^\star \setminus \left\{\widehat{B}_n\right\}$, the statement is true. Suppose that $\widehat{C}_n^{\mathrm{TC}} \notin \mathcal{J}_n^\star \setminus \left\{\widehat{B}_n\right\}$. Then,

$$\begin{aligned} \forall (i,j) \in \mathcal{J}_n^\star \times \left(\overline{V_n^L} \setminus \mathcal{J}_n^\star\right), \quad & W_n(i,j) \geq L^{3/4} C_\nu \,, \\ \forall (i,j) \in \overline{U_n^L} \times U_n^L, \quad & W_n(i,j) \leq \sqrt{L}(D_\nu + D_0 W_\mu)^2 \,. \end{aligned}$$

Let $L_6 = C_\nu^{-4}(D_\nu + D_0 W_\mu)^8 + 1$ which satisfies $\mathbb{E}_\nu[L_6] < +\infty$ by Lemma 2.10. Let $L_7 = \max\{L_0^{4/3}, L_4^{4/3}, L_5^2, L_6\}$, which satisfies $\mathbb{E}_\nu[L_7] < +\infty$. Then, for all $L \geq L_7$, we have

$$\forall (i,k,j) \in \mathcal{J}_n^\star \times U_n^L \times \left(\overline{V_n^L} \setminus \mathcal{J}_n^\star\right), \quad W_n(i,j) > W_n(i,k) \,.$$

As $\widehat{B}_n \in \mathcal{J}_n^\star$ and $\widehat{C}_n^{\mathrm{TC}} \notin \mathcal{J}_n^\star \setminus \left\{\widehat{B}_n\right\}$, the definition $\widehat{C}_n^{\mathrm{TC}} \in \arg\min_{j \neq \widehat{B}_n} W_n(\widehat{B}_n, j)$ yields that $\widehat{C}_n^{\mathrm{TC}} \in V_n^L$. Otherwise the strict inequality wields a contradiction. This concludes the proof. ■

Lemma B.6. *The TC challenger satisfies Property 2.21.*

Proof. We only sketch the proof of the optimal design IDS, since the same argument yields the one for fixed β. Let $\gamma > 0$. Let N_0, (g_1, N_1), N_2 and N_4 as in (2.22), Property 2.18, Lemmas 2.19 and 2.20. Let $n \geq \max\{N_0, N_1, N_2, N_4\}$. Let $i \neq i^\star$ such that $N_{n,i}/N_{n,i^\star} \geq w_i^\star/w_{i^\star}^\star + \gamma$.

Suppose towards contradiction that $N_{n,j}/N_{n,i^\star} > w_j^\star/w_{i^\star}^\star$ for all $j \neq i^\star$. Then, we have

$$\tilde{\gamma} \geq \sum_{j \neq i^\star} \left(\frac{N_{n,j}}{N_{n,i^\star}} \right)^2 - 1 \geq \left(\frac{w_i^\star}{w_{i^\star}^\star} + \gamma \right)^2 - \left(\frac{w_i^\star}{w_{i^\star}^\star} \right)^2 = \gamma \left(\gamma + 2 \frac{w_i^\star}{w_{i^\star}^\star} \right) .$$

Taking $\tilde{\gamma}$ small enough, *e.g.* $\tilde{\gamma} < \gamma^2$, yields a contradiction. Therefore, there exists $j \notin \{i, i^\star\}$ such that $N_{n,j}/N_{n,i^\star} \leq w_j^\star/w_{i^\star}^\star$. Since $C_n^{\mathrm{TC}} \in \arg\min_{j \neq B_n} W_n(B_n, j)$ and $W_n(i,j)$ as in (2.1),

$$\mathbb{P}_{|n}(C_n^{\mathrm{TC}} = i \mid B_n = i^\star) = 0 \quad \Longleftarrow \quad \frac{C(i^\star, i; \nu_n, N_n/N_{n,i^\star})}{C(i^\star, j; \nu_n, N_n/N_{n,i^\star})} > 1 .$$

Using that $w \to C(i^\star, i; \kappa, w)$ is increasing and the equality at equilibrium (Lemma 2.11),

$$\begin{aligned} \frac{C(i^\star, i; \nu_n, N_n/N_{n,i^\star})}{C(i^\star, j; \nu_n, N_n/N_{n,i^\star})} &\geq \frac{C(i^\star, i; \nu_n, w^\star/w_{i^\star}^\star + \gamma 1_i)}{C(i^\star, i; \nu, w^\star/w_{i^\star}^\star)} \frac{C(i^\star, j; \nu, w^\star/w_{i^\star}^\star)}{C(i^\star, j; \nu_n, w^\star/w_{i^\star}^\star)} \\ &= \left(\frac{\mu_{n,i^\star} - \mu_{n,i}}{\mu_{i^\star} - \mu_i} \frac{\mu_{i^\star} - \mu_j}{\mu_{n,i^\star} - \mu_{n,j}} \right)^2 \frac{1 + w_{i^\star}^\star/w_i^\star}{1 + (w_i^\star/w_{i^\star}^\star + \gamma)^{-1}} . \end{aligned}$$

The equality holds for Gaussian distribution with unit variance. Using Lemma 2.10 and (2.22), $\|\mu_n - \mu\|_\infty \leq W_\mu (K/n)^{1/4} \sqrt{\log\left(e + \sqrt{n/K}\right)}$. Therefore, there exists $N_5 = \mathrm{Poly}(W_\mu)$ (hence $\mathbb{E}_\nu[N_5] < +\infty$) such that, for all $n \geq N_5$,

$$\frac{\mu_{n,i^\star} - \mu_{n,i}}{\mu_{i^\star} - \mu_i} \frac{\mu_{i^\star} - \mu_j}{\mu_{n,i^\star} - \mu_{n,j}} \geq \left(\frac{1 + w_{i^\star}^\star/w_i^\star}{1 + (w_i^\star/w_{i^\star}^\star + \gamma)^{-1}} \right)^{-1/4} .$$

This concludes the proof since

$$\frac{C(i^\star, i; \nu_n, N_n/N_{n,i^\star})}{C(i^\star, j; \nu_n, N_n/N_{n,i^\star})} \geq \sqrt{\frac{1 + w_{i^\star}^\star/w_i^\star}{1 + (w_i^\star/w_{i^\star}^\star + \gamma)^{-1}}} > 1 .$$

■

B.14 Proof of Lemma 2.28

Let $t \in [n^{5/6}, n]$ such that $B_t = k$ with $k \neq i^\star$. Then,

$$\mu_{i^\star} \leq \mu_{t,i^\star} + \sqrt{3\log(t)/N_{t,i^\star}} \leq \mu_{t,k} + \sqrt{3\log(t)/N_{t,k}} \leq \mu_k + \sqrt{12\log(t)/N_{t,k}} ,$$

hence $N_{t,k} \leq 12\log(n)(\mu_{i^\star} - \mu_k)^{-2}$. Using Lemma 2.8 with $\beta = 1/2$, the leader is sampled half the time. Since $N_{t,k}$ is bounded and incremented by one half of the time, this event can not occur too often. Summing over $k \neq i^\star$ concludes the proof.

Appendix C

Complements on Chapter 3

C.1 Proof of Lemma 3.5

As in Section 1.4.2, for Gaussian with unknown variance, we have

$$\{\tau_\delta < +\infty\} \cap \mathcal{E}_\mu^{\mathrm{err}}(\tau_\delta) \subseteq \bigcup_{n\in\mathbb{N}} \bigcup_{i\neq i^\star} \left\{ \sum_{k\in\{i,i^\star\}} \frac{N_{n,k}}{2} \log\left(1 + (\mu_{n,k} - \mu_k)^2/\sigma_{n,k}^2\right) \leq c_{i,i^\star}(N_n, \delta) \right\} .$$

The family of thresholds in (3.7) is obtained by solving an optimization problem at each time n and for all $i \neq i^\star$ solution. Let $C, D \in (\mathbb{R}_+^\star)^2$ and $N \in (\mathbb{N})^2$,

$$\begin{aligned} &\text{maximize} \sum_{k\in\{1,2\}} \frac{N_k}{2} \log(1 + y_k) \\ &\text{such that } \forall k \in \{1,2\}, \quad y_k \geq 0,\ x_k y_k \leq C_k,\ x_k \geq D_k . \end{aligned}$$

Since $y \mapsto \log(1+y)$ is concave and increasing, $y \mapsto \sum_{k\in\{1,2\}} N_k \log(1+y_b)$ is concave and increasing in each of its coordinates. Since the constraints and the objective are separate between each coordinate, the maximum is achieved at C_k/D_k and has value

$$\sum_{k\in\{1,2\}} \frac{N_k}{2} \log\left(1 + \frac{C_k}{D_k}\right) .$$

Let n and $i \neq i^\star$. The variables y_k and x_k replace $(\mu_{n,i} - \mu_i)^2/\sigma_{n,i}^2$ and $\sigma_{n,i}^2/\sigma_i^2$. Then, we need to specify the constraint imposed by concentration, *i.e.* specify (D_k, C_k) . Combining the lower tail concentration on the empirical variance (Corollary C.8) and the upper and lower tail concentration of the empirical mean (Lemma C.9), we obtain that with probability greater

than $1 - \frac{\delta}{K-1}$, for all $k \in \{i, i^\star\}$ and all $n \geq \max_{k \in \{i,i^\star\}} t_k(\delta)$,

$$(\mu_{n,k} - \mu_k)^2 \leq \sigma_k^2 \varepsilon_\mu(N_{n,k}, \delta) \,,$$
$$\sigma_{n,k}^2 \geq \sigma_k^2 (1 - \varepsilon_{-,\sigma}(N_{n,k} - 1, \delta)) \,.$$

where ε_μ , $\varepsilon_{-,\sigma}$ and t_k are defined as in Lemma 3.5. Using Lemma A.1, this initial time condition ensures that $1 - \varepsilon_{-,\sigma}(N_{n,k} - 1, \delta) > 0$. Since $\overline{W}_0$ has values in $(0, 1)$, we obtain that $\varepsilon_{-,\sigma}(n, \delta) \in (0, 1)$. Taking a union bound over $i \neq i^\star$ concludes the proof.

C.1.1 Sub-Exponential Processes

We prove time-uniform and fixed-time concentration results for 1-sub-$\psi_{E,c}$ process with variance process $V_t = ct$ and 1-sub-$\psi_{E,-c}$ process with variance process $V_t = ct$. The concept of sub-ψ process (Definition C.1) was introduced in Howard et al. [2020]. This concept is particularly useful to derive time-uniform concentration results.

Definition C.1. *Let $(S_t)_{t \in \mathcal{T} \cup \{0\}}$ and $(V_t)_{t \in \mathcal{T} \cup \{0\}}$ be two real-valued processes adapted to an underlying filtration $(\mathcal{F}_t)_{t \in \mathcal{T} \cup \{0\}}$ with $S_0 = 0$ and $V_0 = 0$ a.s. and $V_t \geq 0$ a.s. for all $t \in \mathcal{T}$. For a function $\psi : [0, \lambda_{\max}) \mapsto \mathbb{R}$ and a scalar $l_0 \in [1, +\infty)$, we say that (S_t) is l_0-sub-ψ with variance process (V_t) if, for each $\lambda \in [0, \lambda_{\max})$, there exists a supermartingale $(L_t(\lambda))_{t \in \mathcal{T} \cup \{0\}}$ with respect to $(\mathcal{F}_t)$ such that $L_0(\lambda) \leq l_0$ a.s. and*

$$\exp\{\lambda S_t - \psi(\lambda) V_t\} \leq L_t(\lambda) \qquad \text{a.s. for all } t \in \mathcal{T} \,.$$

Lemma C.2 (Ville's inequality)**.** *Let $\mathbb{P}_0[\cdot] = \mathbb{P}_0[\cdot \mid \mathcal{F}_0]$. Let $\mathcal{T} \subseteq \mathbb{N}$, such that $|\mathcal{T}| = \infty$. If $(L_t)_{t \in \mathcal{T} \cup \{0\}}$ is a non-negative supermartingale with respect to the filtration $(\mathcal{F}_t)_{t \in \mathcal{T} \cup \{0\}}$, then*

$$\forall a > 0, \quad \mathbb{P}_0\left(\exists t \in \mathcal{T} : L_t \geq a\right) \leq L_0 / a \,.$$

Since we aim at deriving one-sided bounds on scalar martingales, we have $l_0 = 1$. Using Ville's inequality (Lemma C.2) on a sub-ψ process yields time-uniform concentration results. Let (S_t) be a 1-sub-ψ with variance process (V_t) , then for all $\lambda \in [0, \lambda_{\max})$, with probability greater than $1 - \delta$,

$$\forall t \in \mathcal{T}, \quad \lambda S_t - \psi(\lambda) V_t < \log(1/\delta) \,.$$

Let $\lambda \in [0, \lambda_{\max})$. Direct manipulations show the above result,

$$\mathbb{P}\left(\exists t \in \mathcal{T} :\ \lambda S_t - \psi(\lambda) V_t \geq \log\left(1/\delta\right)\right) \leq \mathbb{P}\left(\exists t \in \mathcal{T} :\ L_t(\lambda) \geq 1/\delta\right) \leq \delta\,.$$

In the following, we are interested by 1-sub-$\psi_{E,c}$ processes for $c \in \mathbb{R}$, where $\psi_{E,c}$ is defined as

$$\forall \lambda \in [0, 1/(c \vee 0))\,, \quad \psi_{E,c}(\lambda) = \frac{-\log(1-c\lambda) - c\lambda}{c^2}\,. \tag{C.1}$$

The derived upper and lower tails concentrations involve the positive ($i = 0$) and negative ($i = -1$) Lambert's branches W_i solutions of $W(x)e^{W(x)} = x$. We refer the reader to Appendix A for mode details and corresponding technical results.

Upper tail concentration We derive time-uniform and fixed-time upper tail concentration for 1-sub-$\psi_{E,c}$ process with variance process $V_t = ct$. While the time-uniform result require using the peeling method, the proof of the fixed-time concentration is simpler. To use the peeling method, we need to control the deviation of the process on slices of time (Lemma C.3).

Lemma C.3. *Let $c > 0$ and S_t a 1-sub-$\psi_{E,c}$ process with variance process $V_t = ct$. Let $N > 0$. For all $x > 1$, there exists $\lambda = \lambda(x)$ such that for all $t \geq N$,*

$$\{S_t + t \geq tx\} \subseteq \left\{\lambda S_t - ct\psi_{E,c}(\lambda) \geq \frac{N}{c}\left(h\left(x\right) - 1\right)\right\}$$

where $\lambda(x) = \arg\max_{\lambda \in [0,1/c)} \left(x\lambda + \frac{\log(1-c\lambda)}{c}\right)$ and $h(x) = x - \log(x)$ for $x > 1$.

Proof. Defining $\psi_U(\lambda) = \lambda + c\psi_{E,c}(\lambda) = -\frac{\log(1-c\lambda)}{c}$ and $\lambda(x) = \arg\max_{\lambda \in [0,1/c)} x\lambda - \psi_U(\lambda)$, we have $x\lambda(x) - \psi_U(\lambda(x)) = \psi_U^*\left(x\right)$ where ψ_U^* is the convex conjugate of ψ_U . Note that $\psi_U^*\left(x\right) \geq 0$ (see below), hence $t\psi_U^*\left(x\right) \geq N\psi_U^*\left(x\right)$ for $t \geq N$. Direct computations yield

$$\begin{aligned}
S_t + t \geq tx &\iff \lambda S_t - ct\psi_{E,c}(\lambda) \geq tx\lambda - t(\lambda + c\psi_{E,c}(\lambda)) \\
&\implies \lambda S_t - ct\psi_{E,c}(\lambda) \geq t\left(x\lambda - \psi_U(\lambda)\right) = t\psi_U^*\left(x\right) \\
&\implies \lambda S_t - ct\psi_{E,c}(\lambda) \geq N_i\psi_U^*\left(x\right) = \frac{N}{c}\left(h\left(x\right) - 1\right)
\end{aligned}$$

Note that for $f(\lambda) = \lambda x + \frac{\log(1-c\lambda)}{c}$, we have $f'(\lambda) = x - \frac{1}{1-c\lambda} = 0 \iff \lambda = \frac{1}{c}\left(1 - \frac{1}{x}\right)$ and $\frac{1}{c}\left(1 - \frac{1}{x}\right) \in [0, \frac{1}{c}) \iff x > 1$. Since $f''(\lambda) = -\frac{c}{(1-c\lambda)^2} \leq 0$, the function is concave hence this is a maximum. This yields that for all $x > 1$, $\psi_U^*(x) = f(\frac{1}{c}\left(1 - \frac{1}{x}\right)) = \frac{1}{c}\left(x - 1 - \log(x)\right) = \frac{1}{c}(h(x) - 1) \geq 0$ where $h(x) = x - \log(x)$. ∎

Let $\eta > 0$. Applying Lemma C.3 on slices of time with geometric growth rate $(N_i)_{i\in\mathbb{N}}$ with $N_i = (1+\eta)^{i-1}$, we obtain Lemma C.4.

Lemma C.4. *Let $\overline{W}_{-1}(x) = -W_{-1}(-e^{-x})$ for $x \geq 1$, $\delta \in (0,1)$, $\eta > 0$, $s > 1$, $c > 0$, and ζ be the Riemann ζ function. Let S_t a 1-sub-$\psi_{E,c}$ process with variance process $V_t = ct$. Then, with probability greater than $1-\delta$, for all $t \in \mathbb{N}$,*

$$S_t + t \leq t\overline{W}_{-1}\left(1 + \frac{c(1+\eta)}{t}\left(\log\left(\frac{\zeta(s)}{\delta}\right) + s\log\left(1 + \frac{\log(t)}{\log(1+\eta)}\right)\right)\right) .$$

Proof. Let $g(t,\delta)$ such that $g(t,\delta) \geq x_i(\delta)$ for $t \in [N_i, N_{i+1})$ and $x_i(\delta) > 1$. Using Lemma C.3 with $x_i(\delta) > 1$ and $g(t,\delta) \geq x_i(\delta)$ on $[N_i, N_{i+1})$, we obtain

$$\begin{aligned}
\mathbb{P}\left(\exists t \in \mathbb{N} : S_t + t \geq t g(t,\delta)\right) &\leq \sum_{i\in\mathbb{N}} \mathbb{P}\left(\exists t \in [N_i, N_{i+1}) : S_t + t \geq t x_i(\delta)\right) \\
&\leq \sum_{i\in\mathbb{N}} \mathbb{P}\left(\exists t \in [N_i, N_{i+1}) : \lambda S_t - ct\psi_{E,c}(\lambda) \geq \frac{N_i}{c}\left(h\left(x_i(\delta)\right) - 1\right)\right) \\
&\leq \sum_{i\in\mathbb{N}} e^{-\frac{N_i}{c}(h(x_i(\delta))-1)} ,
\end{aligned}$$

where the last inequality uses that S_t a 1-sub-$\psi_{E,c}$ process with variance process $V_t = ct$. Taking

$$g(t,\delta) = \overline{W}_{-1}\left(1 + \frac{c(1+\eta)}{t}\left(\log\left(\frac{\zeta(s)}{\delta}\right) + s\log\left(1 + \frac{\log(t)}{\log(1+\eta)}\right)\right)\right)$$

and $x_i(\delta) = \overline{W}_{-1}\left(1 + \frac{c}{N_i}\log\left(\frac{i^s\zeta(s)}{\delta}\right)\right)$ satisfies the required properties. First, we have $x_i(\delta) > 1$ (Lemma A.1). Second, since $\overline{W}_{-1}$ is increasing on $(1, +\infty)$ (Lemma A.1), $t \in [N_i, N_{i+1})$ and $i = 1 + \frac{\log(N_i)}{\log(1+\eta)}$, we obtain

$$g(t,\delta) \geq \overline{W}_{-1}\left(1 + \frac{c\log\left(\frac{\zeta(s)}{\delta}\right) + cs\log\left(1 + \frac{\log(t)}{\log(1+\eta)}\right)}{N_i}\right) \geq \overline{W}_{-1}\left(1 + \frac{c}{N_i}\log\left(\frac{i^s\zeta(s)}{\delta}\right)\right)$$

Using Lemma A.1 for each $i \in \mathbb{N}$ yields

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t + t \geq t g(t,\delta)\right) \leq \sum_{i\in\mathbb{N}} e^{-\frac{N_i}{c}(h(x_i(\delta))-1)} \leq \frac{\delta}{\zeta(s)}\sum_{i\in\mathbb{N}}\frac{1}{i^s} = \delta$$

■

Lower tail concentration We derive time-uniform and fixed-time lower tail concentration for 1-sub-$\psi_{E,-c}$ process with variance process $V_t = ct$. Likewise, we use the peeling method and control the deviation of the process on slices of time (Lemma C.5).

Lemma C.5. *Let $c > 0$ and $-S_t$ a 1-sub-$\psi_{E,-c}$ process with variance process $V_t = ct$. Let $N > 0$. For all $x \in (0,1)$, there exists $\lambda = \lambda(x)$ such that for all $t \geq N$,*

$$\{-S_t - t \geq -tx\} \subseteq \left\{\lambda(-S_t) - ct\psi_{E,-c}(\lambda) \geq \frac{N}{c}\left(h\left(x\right) - 1\right)\right\}$$

where $\lambda(x) = \arg\max_{\lambda \in [0,+\infty)} \left(-x\lambda + \frac{\log(1+c\lambda)}{c}\right)$ and $h(x) = x - \log(x)$ for $x \in (0,1)$.

Proof. Defining $\psi_L(\lambda) = -\lambda + c\psi_{E,-c}(\lambda) = -\frac{\log(1+c\lambda)}{c}$ and $\lambda(x) = \arg\max_{\lambda \in [0,+\infty)} -x\lambda - \psi_L(\lambda)$, we have $-x\lambda(x) - \psi_L(\lambda(x)) = \psi_L^*\left(-x\right) \geq 0$ (see below), hence $t\psi_L^*\left(-x\right) \geq N\psi_L^*\left(-x\right)$ for $t \geq N$. Direct computations yield

$$\begin{aligned}
-S_t - t \geq -tx &\iff \lambda(-S_t) - ct\psi_{E,-c}(\lambda) \geq -tx\lambda - t(-\lambda + c\psi_{E,-c}(\lambda)) \\
&\implies \lambda(-S_t) - ct\psi_{E,-c}(\lambda) \geq t\left(-x\lambda - \psi_L(\lambda)\right) = t\psi_L^*\left(-x\right) \\
&\implies \lambda(-S_t) - ct\psi_{E,-c}(\lambda) \geq N\psi_L^*\left(-x\right) = \frac{N}{c}\left(h\left(x\right) - 1\right)
\end{aligned}$$

Note that for $f(\lambda) = -\lambda x + \frac{\log(1+c\lambda)}{c}$, we have $f'(\lambda) = -x + \frac{1}{1+c\lambda} = 0 \iff \lambda = \frac{1}{c}\left(\frac{1}{x} - 1\right)$ and $\frac{1}{c}\left(\frac{1}{x} - 1\right) \in [0, +\infty) \iff x \in (0,1)$. Since $f''(\lambda) = -\frac{c}{(1+c\lambda)^2} \leq 0$, the function is concave hence this is a maximum. This yields that for all $x \in (0,1)$, $\psi_L^*(-x) = f(\frac{1}{c}\left(\frac{1}{x} - 1\right)) = \frac{1}{c}\left(x - 1 - \log(x)\right) = \frac{1}{c}\left(h(x) - 1\right) \geq 0$ where $h(x) = x - \log(x)$ for $x \in (0,1)$. ∎

Let $\eta > 0$. Applying Lemma C.5 on slices of time with geometric growth rate $(N_i)_{i \in \mathbb{N}}$ with $N_i = (1+\eta)^{i-1}$, we obtain Lemma C.6.

Lemma C.6. *Let $\overline{W}_0(x) = -W_0(-e^{-x})$ for $x \geq 1$, $\delta \in (0,1)$, $\eta > 0$, $s > 1$, $c > 0$, and ζ be the Riemann ζ function. Let S_t a 1-sub-$\psi_{E,-c}$ process with variance process $V_t = ct$. Then, with probability greater than $1 - \delta$, for all $t \in \mathbb{N}$,*

$$S_t + t \geq t\overline{W}_0\left(1 + \frac{c(1+\eta)}{t}\left(\log\left(\frac{\zeta(s)}{\delta}\right) + s\log\left(1 + \frac{\log(t)}{\log(1+\eta)}\right)\right)\right) .$$

Proof. Let $g(t,\delta)$ positive such that $g(t,\delta) \leq x_i(\delta)$ for $t \in [N_i, N_{i+1})$ and $x_i(\delta) \in (0,1)$. Using Lemma C.5 with $x_i(\delta) < 1$ and $g(t,\delta) \leq x_i(\delta)$ for $t \in [N_i, N_{i+1})$, we obtain

$$\begin{aligned}
\mathbb{P}\left(\exists t \in \mathbb{N} : S_t + t \leq t g(t,\delta)\right) &= \mathbb{P}\left(\exists t \in \mathbb{N} : -S_t - t \geq -t g(t,\delta)\right) \\
&\leq \sum_{i \in \mathbb{N}} \mathbb{P}\left(\exists t \in T_i : -S_t - t \geq -t x_i(\delta)\right) \\
&\leq \sum_{i \in \mathbb{N}} \mathbb{P}\left(\exists t \in T_i : \lambda(-S_t) - ct\psi_{E,-c}(\lambda) \geq \frac{N_i}{c}\left(h\left(x_i(\delta)\right) - 1\right)\right) \\
&\leq \sum_{i \in \mathbb{N}} e^{-\frac{N_i}{c}\left(h(x_i(\delta)) - 1\right)}
\end{aligned}$$

where the last inequality uses that $-S_t$ a 1-sub-$\psi_{E,-c}$ process with variance process $V_t = ct$. Taking

$$g(t,\delta) = \overline{W}_0\left(1 + \frac{c(1+\eta)}{t}\left(\log\left(\frac{\zeta(s)}{\delta}\right) + s\log\left(1 + \frac{\log(t)}{\log(1+\eta)}\right)\right)\right)$$

and $x_i(\delta) = \overline{W}_0\left(1 + \frac{c}{N_i}\log\left(\frac{i^s\zeta(s)}{\delta}\right)\right)$ satisfies the required properties. First, we have $x_i(\delta) \in (0,1)$ and $g(t,\delta) > 0$ (Lemma A.1). Second, since $\overline{W}_0$ is decreasing on $(1,+\infty)$ (Lemma A.1), $t \in [N_i, N_{i+1})$ and $i = 1 + \frac{\log(N_i)}{\log(1+\eta)}$, we obtain

$$g(t,\delta) \leq \overline{W}_0\left(1 + \frac{c\log\left(\frac{\zeta(s)}{\delta}\right) + cs\log\left(1 + \frac{\log(t)}{\log(1+\eta)}\right)}{N_i}\right) \leq \overline{W}_0\left(1 + \frac{c}{N_i}\log\left(\frac{i^s\zeta(s)}{\delta}\right)\right)$$

Using Lemma A.1 for each $i \in \mathbb{N}$ yields

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t + t \leq t g(t,\delta)\right) \leq \sum_{i \in \mathbb{N}} e^{-\frac{N_i}{c}\left(h(x_i(\delta)) - 1\right)} \leq \frac{\delta}{\zeta(s)}\sum_{i \in \mathbb{N}} \frac{1}{i^s} = \delta\,.$$

■

C.1.2 Univariate Gaussian

Empirical variance The empirical variance can be expressed as a function of a sub-exponential process (Lemma C.7). This is obtained with manipulations derived in the Appendix H of Howard et al. [2021], in which they consider martingales with χ^2 increments.

Lemma C.7. *Let σ_t^2 be the empirical variance of t* i.i.d. *samples from a Gaussian distribution with variance σ^2 . Then, $\frac{\sigma_t^2}{\sigma^2} = \frac{S_{t-1}-1}{t} + 1$ with $S_{t-1} + t - 1 = \sum_{i=1}^{t-1} Y_i^2$ where (Y_i) are* i.i.d. *with distributions $\mathcal{N}(0,1)$. In particular, S_t is a 1-sub-$\psi_{E,2}$ process and $-S_t$ is a 1-sub-$\psi_{E,-2}$ process, both with variance process $V_t = 2t$.*

Proof. Let $(X_i)_{i\in[n]}$ the n samples from a Gaussian distribution with parameters (μ, σ^2) . Let $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ be the empirical mean and variance. Let $Z_i = \frac{X_i-\mu}{\sigma}$ for all $i \in [n]$, $\hat{Z}_n = \frac{1}{n}\sum_{i\in[n]} Z_i$ and $S_{n-1} = \sum_{i=1}^n (Z_i - \hat{Z}_n)^2 - (n-1)$. Then, $S_0 = 0$ and for all $n \geq 2$

$$S_{n-1} = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \hat{\mu}_n)^2 - (n-1) = n\frac{\sigma_n^2}{\sigma^2} - (n-1)\,.$$

Rewriting the increment of S_n , we obtain for all $n \geq 2$

$$\begin{aligned} S_{n-1} - S_{n-2} &= (Z_n - \hat{Z}_n)^2 + \sum_{i=1}^{n-1}((Z_i - \hat{Z}_n)^2 - (Z_i - \hat{Z}_{n-1})^2) - 1 \\ &= Z_n^2 + \hat{Z}_n^2 - 2Z_n\hat{Z}_n + \sum_{i=1}^{n-1}(-2Z_i\hat{Z}_n + 2Z_i\hat{Z}_{n-1}) + (n-1)(\hat{Z}_n^2 - \hat{Z}_{n-1}^2) - 1 \\ &= Z_n^2 - n\hat{Z}_n^2 + (n-1)\hat{Z}_{n-1}^2 - 1 = \frac{n-1}{n}(Z_n - \hat{Z}_{n-1})^2 - 1\,. \end{aligned}$$

Since $S_{n-1} = \sum_{s=1}^{n-1}(S_s - S_{s-1})$ and $S_0 = 0$ a.s., we obtain $S_{n-1} = \sum_{i=1}^{n-1}(Y_i^2 - 1)$ where $Y_{n-1} = \sqrt{\frac{n-1}{n}}(Z_n - \hat{Z}_{n-1})$. The (Y_i) are *i.i.d.* with distribution $\mathcal{N}(0,1)$ and the CGF of $(Y_i^2 - 1)$ is

$$\log \mathbb{E}e^{\lambda(Y_i^2-1)} = -\frac{\log(1-2\lambda)}{2} - \lambda = 2\psi_{E,2}(\lambda) \quad \text{for } \lambda \in (-\infty, 1/2)\,.$$

By Definition C.1, we have that S_n is 1-sub-$\psi_{E,2}$ and that $-S_n$ is 1-sub-$\psi_{E,-2}$, both with variance process $V_n = 2(n-1)$. ■

Thanks to Lemmas C.4-C.6-C.7, Corollary C.8 gives time-uniform upper and lower tails concentrations on the empirical variance of Gaussian observation.

Corollary C.8. *For $i \in \{0, -1\}$, let $\overline{W}_i(x) = -W_i(-e^{-x})$ for $x \geq 1$, $\delta \in (0,1)$, $\eta_0, \eta_1 > 0$, $s > 1$ and ζ be the Riemann ζ function. Let σ_{t+1}^2 be the empirical variance of $t+1$ i.i.d. samples from a Gaussian distribution with variance σ^2 . Then, with probability greater than $1-\delta$, for all $t \in \mathbb{N}^\star$,*

$$\sigma_{t+1}^2 \leq \sigma^2\left(\overline{W}_{-1}\left(1 + \frac{2(1+\eta_1)}{t}\left(\log\left(\frac{\zeta(s)}{\delta}\right) + s\log\left(1 + \frac{\log(t)}{\log(1+\eta_1)}\right)\right)\right) - \frac{1}{t}\right)\,.$$

Moreover, with probability $1-\delta$, for all $t \geq t_0(\delta)$,

$$\sigma_{t+1}^2 \geq \sigma^2\left(\overline{W}_{0}\left(1 + \frac{2(1+\eta_0)}{t}\left(\log\left(\frac{\zeta(s)}{\delta}\right) + s\log\left(1 + \frac{\log(t)}{\log(1+\eta_0)}\right)\right)\right) - \frac{1}{t}\right)\,,$$

where the initial time condition, which ensures the lower bound is positive, is

$$t_0(\delta) = \inf\left\{ t \mid t > e^{1+W_0\left(\frac{2(1+\eta_0)}{e}\left(\log\left(\frac{\zeta(s)}{\delta}\right)+s\log\left(1+\frac{\log(t)}{\log(1+\eta_0)}\right)\right)-e^{-1}\right)} \right\} .$$

Proof. Combining Lemmas C.4-C.6-C.7 yields the desired result. Using Lemma A.1, we know that the upper bound is always positive. The initial time condition, after which the lower bound is positive, is obtained by Lemma A.1

$$\overline{W}_0\left(1+\frac{2(1+\eta_0)}{t}\left(\log\left(\frac{\zeta(s)}{\delta}\right)+s\log\left(1+\frac{\log(t)}{\log(1+\eta_0)}\right)\right)\right) > \frac{1}{t} \quad \Longleftrightarrow \quad t \geq t_0(\delta) .$$

■

Empirical mean While time-uniform concentration results for the empirical mean of Gaussian observations already exist in the literature, *e.g.* Kaufmann and Koolen [2021], Lemma C.9 is given to present unified concentration results involving $\overline{W}_{-1}$. For the sake of space, we omit the proof and refer the reader to Appendix E in Jourdan et al. [2023a] for more details.

Lemma C.9. *Let $\overline{W}_{-1}(x) = -W_{-1}(-e^{-x})$ for $x \geq 1$, $\delta \in (0,1)$, $s > 1$ and ζ be the Riemann ζ function. Let μ_t be the empirical mean of t* i.i.d. *samples from a Gaussian distribution with parameter (μ, σ^2) . Then, with probability greater than $1-\delta$, for all $t \in \mathbb{N}$,*

$$|\mu_t - \mu| \leq \sqrt{\frac{\sigma^2}{t}\overline{W}_{-1}\left(1+2\log\left(\frac{1}{\delta}\right)+2g(s)+2s\log\left(2s+\log t\right)\right)} ,$$

where $g(s) = \log(\zeta(s)) + s(1-\log(2s))$.

Appendix D

Complements on Chapter 4

D.1 Proof of Lemma 4.3

As in Section 1.4.2, for bounded distribution, we have

$$\{\tau_\delta < +\infty\} \cap \mathcal{E}_\mu^{\mathrm{err}}(\tau_\delta) \subseteq \bigcup_{n\in\mathbb{N}} \bigcup_{i\neq i^\star} \{N_{n,i}\mathcal{K}_{\inf}^-(\nu_{n,i},\mu_i) + N_{n,i^\star}\mathcal{K}_{\inf}^+(\nu_{n,i^\star},\mu_{i^\star}) > c(n-1,\delta)\}\,.$$

The key technical result from Agrawal et al. [2021b] is stated in Lemma D.1 without proof.

Lemma D.1 (Lemma E.1 in Agrawal et al. [2021b]). *Let a compact and convex set $\Lambda \subseteq \mathbb{R}^d$, and q be the uniform distribution on Λ . Let $g_t : \Lambda \mapsto \mathbb{R}$ be any series of exp-concave functions. Then,*

$$\max_{\boldsymbol{\lambda}\in\Lambda} \sum_{k=1}^{n} g_k(\boldsymbol{\lambda}) \leq \log \mathbb{E}_{\boldsymbol{\lambda}\sim q}\left[e^{\sum_{k=1}^n g_t(\boldsymbol{\lambda})}\right] + d\log(n+1) + 1$$

For all $(n,i) \in \mathbb{N}\times[K]$, we denote by $(X_{k,i})_{k\in[N_{n,i}]}$ the samples collected on arm i . Let $i^\star = i^\star(\boldsymbol{F})$ and $i \in [K]\setminus\{i^\star\}$. Using the dual formulation obtained by Honda and Takemura [2010], we have

$$N_{n,i^\star}\mathcal{K}_{\inf}^+(F_{n,i^\star},\mu_{i^\star})) = \max_{\lambda\in\left[0,\frac{1}{B-\mu_{i^\star}}\right]} \sum_{k\in[N_{n,i^\star}]} \log(1-\lambda(X_{k,i^\star}-\mu_{i^\star}))\,,$$

$$N_{n,i}\mathcal{K}_{\inf}^-(F_{n,i},\mu_i)) = \max_{\lambda\in\left[0,\frac{1}{\mu_i}\right]} \sum_{k\in[N_{n,i}]} \log(1+\lambda(X_{k,i}-\mu_i))\,.$$

Let q_i^+ and q_i^- be the uniform distributions over $\left[0, \frac{1}{B-\mu_i}\right]$ and $\left[0, \frac{1}{\mu_i}\right]$, which are compact and convex sets of $\mathbb{R}$. Define

$$\begin{aligned}
L_{n,i} &= \mathbb{E}_{\lambda\sim q_i^-}\left[\prod_{k\in[N_{n,i}]}(1+\lambda(X_{k,i}-\mu_i)) \mid X_{1,i},\cdots,X_{N_{n,i},i}\right], \\
U_{n,i} &= \mathbb{E}_{\lambda\sim q_i^+}\left[\prod_{k\in[N_{n,i}]}(1-\lambda(X_{i,k}-\mu_i)) \mid X_{1,i},\cdots,X_{N_{n,i},i}\right], \\
Y_{n,i}^- &= N_{n,i}\mathcal{K}_{\inf}^-(F_{n,i},\mu_i) - \log(N_{n,i}+1) - 1\,, \\
Y_{n,i}^+ &= N_{n,i}\mathcal{K}_{\inf}^+(F_{n,i},\mu_i) - \log(N_{n,i}+1) - 1\,.
\end{aligned}$$

With $d=1$, using Lemma D.1 with the exp-concave functions $g_{k,i}^+(\lambda) = \log(1-\lambda(X_{k,i}-\mu_i))$ for $k\in[N_{n,i}]$, and $g_{k,i}^-(\lambda) = \log(1+\lambda(X_{k,i}-\mu_i))$ for $k\in[N_{n,i}]$, yields

$$e^{Y_{n,i}^-} \le L_{n,i} \qquad \text{and} \qquad e^{Y_{n,i}^+} \le U_{n,i} \qquad \text{a.s.}$$

Furthermore, it is easy to verify that for each arm $i\in[K]$, $L_{n,i}$ and $U_{n,i}$ are non-negative martingales with unit initial value $L_{0,i}=1$ and $U_{0,i}=1$ almost surely. The martingale property is shown directly by the tower rule (conditioned on the arm sampled at time n) and $\mathbb{E}\left[1\pm\lambda(X_{N_{n,i},i}-\mu_i)\right]=1$. Furthermore, they satisfy $\mathbb{E}[U_{n,i}]\le 1$ and $\mathbb{E}[L_{n,i}]\le 1$. Thus, $U_{n,i^\star}L_{n,i}$ is a non-negative martingale with unit initial value.

By concavity of $\log$ and using $\sum_{j\in\{i,i^\star\}} N_{n,j} \le n-1$, we have

$$c(n-1,\delta) \ge \log\left(\frac{K-1}{\delta}\right) + 2 + \sum_{j\in\{i,i^\star\}} \log\left(N_{n,j}+1\right)\,.$$

Taking a union bound over $i\neq i^\star$ and using Ville's inequality, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\exists n\in\mathbb{N},\ \exists i\neq i^\star,\ N_{n,i}\mathcal{K}_{\inf}^-(F_{n,i},\mu_i) + N_{n,i^\star}\mathcal{K}_{\inf}^+(F_{n,i^\star},\mu_{i^\star}) > c(n-1,\delta)\right) \\
&\le \sum_{i\neq i^\star}\mathbb{P}\left(\exists n\in\mathbb{N},\ Y_{n,i}^- + Y_{n,i^\star}^+ > \log\left(\frac{K-1}{\delta}\right)\right) \\
&\le \sum_{i\neq i^\star}\mathbb{P}\left(\exists n\in\mathbb{N},\ U_{n,i^\star}L_{n,i} > \frac{K-1}{\delta}\right) \le \delta\,.
\end{aligned}$$

This concludes the proof.

D.2 Proof of Lemma 4.6

Using (4.5), $u \mapsto \mathcal{K}_{\text{inf}}^{+}(\kappa, u)$ is strictly convex on $(m(\kappa), B]$ if and only if $\lambda_\star^+(\kappa, u)$ is increasing for $u > m(\kappa)$. Using Lemma 4.5, we obtain that $\lambda_\star^+(\kappa, u)$ is increasing on $u \in [u^+(\kappa), B)$ (since $u \to (B-u)^{-1}$ is), and null on $[0, m(\kappa)]$.

Suppose towards contradiction that $u \mapsto \lambda_\star^+(\kappa, u)$ is not increasing for $(m(\kappa), u^+(\kappa))$. Therefore, there exists an open $\mathcal{O} \subseteq (m(\kappa), u^+(\kappa))$, such that $u \mapsto \lambda_\star^+(\kappa, u)$ is constant on $\mathcal{O}$, *i.e.* there exists $c_{\mathcal{O}} \in [0, (B - u_{\mathcal{O}})^{-1}]$ such that $\lambda_\star^+(\kappa, u) = c_{\mathcal{O}}$ and $u_{\mathcal{O}} = \inf_{u \in \mathcal{O}} u$. Using Lemma 4.5, we know that $c_{\mathcal{O}} \in (0, (B - u_{\mathcal{O}})^{-1})$. On $\mathcal{O}$, $u \mapsto \lambda_\star^+(\kappa, u)$ is constant, hence it is continuously differentiable with null derivative. Since $\mathcal{O} \subseteq (m(\kappa), u^+(\kappa))$, we have $\lambda_\star^+(\kappa, u) \in (0, (B - u_{\mathcal{O}})^{-1})$ and $\mathbb{E}_F\left[(1 - \lambda_\star^+(\kappa, u)(X - u))^{-1}\right] = 1$ for all $u \in \mathcal{O}$. Therefore, the function $(x, u) \mapsto (1 - \lambda_\star^+(\kappa, u)(x - u))^{-1}$ is bounded on $[0, B] \times \mathcal{O}$, hence integrable, and the function $u \mapsto (1 - \lambda_\star^+(\kappa, u)(x - u))^{-1}$ is continuously differentiable. Moreover, the function $x \mapsto (1 - \lambda_\star^+(\kappa, u)(x - u))^{-2}$ is strictly positive and bounded on $[0, B]$, hence integrable with strictly positive integrable. Having checked all the conditions to interchange the derivative with the expectation, differentiating the above yields

$$0 = \mathbb{E}_F\left[-\frac{\lambda_\star^+(\kappa, u) + (u - X)\frac{\partial \lambda_\star^+(\kappa, u)}{\partial u}}{\left(1 - (X - u)\lambda_\star^+(\kappa, u)\right)^2}\right] = -c_{\mathcal{O}}\mathbb{E}_F\left[(1 - (X - u)c_{\mathcal{O}})^{-2}\right] < 0\,,$$

where the strict inequality is obtained since we show that $c_{\mathcal{O}} > 0$ and $\mathbb{E}_F\left[(1 - (X - u)c_{\mathcal{O}})^{-2}\right] > 0$. This is a contradiction, hence such $\mathcal{O} \subset (m(\kappa), u^+(\kappa))$ doesn't exist. Therefore, $u \mapsto \lambda_\star^+(\kappa, u)$ is increasing on $(m(\kappa), u^+(\kappa))$.

By convexity, we already knew that $u \mapsto \lambda_\star^+(\kappa, u)$ is non-decreasing on $(m(\kappa), B]$. Combining the above, we have shown that $u \mapsto \lambda_\star^+(\kappa, u)$ is increasing on $(m(\kappa), B]$, hence $u \to \mathcal{K}_{\text{inf}}^{+}(\kappa, u)$ is strictly convex on $(m(\kappa), B]$. The fact that $u \mapsto \lambda_\star^+(\kappa, u)$ is increasing on $(m(\kappa), u^+(\kappa))$ and on $[u^+(\kappa), B]$, yields that $u \mapsto \lambda_\star^+(\kappa, u)$ is increasing on $(m(\kappa), B]$.

Since $\mathcal{K}_{\text{inf}}^{+}(\kappa, u) = 0$ for all $u \in [0, m(\kappa)]$ and $\mathcal{K}_{\text{inf}}^{+}$ is nonnegative and strictly convex for $u > m(\kappa)$, we obtain that $u \mapsto \mathcal{K}_{\text{inf}}^{+}(\kappa, u)$ is increasing on $(m(\kappa), B]$.

Appendix E

Complements on Chapter 5

E.1 Reduction of an ε-BAI Problem to a BAI Problem

As detailed in Chapter 2, the (β-)characteristic time for the fixed-confidence BAI setting with Gaussian bandits $\mathcal{N}(\mu, 1)$ is defined as

$$T^\star(\nu) = \min_{\beta \in (0,1)} T^\star_\beta(\nu) \quad \text{with} \quad 2T^\star_\beta(\nu)^{-1} = \max_{w \in \triangle_K, w_{i^\star} = \beta} \min_{i^\star \neq i} \frac{(\mu_{i^\star} - \mu_i)^2}{1/\beta + 1/w_i} \,. \tag{E.1}$$

It satisfies $H(\mu) \leq T^\star(\nu) \leq 2H(\mu)$ where $H(\mu) = 2\sum_{i \in [K]} \Delta_i^{-2}$ where $\Delta_{i^\star} = \Delta_{\min}$. Using the equality at equilibrium (2.20) (Lemma 2.11), one can show that

$$\beta w^\star_\beta(\nu)_i^{-1} = \beta T^\star_\beta(\nu)(\mu_{i^\star} - \mu_i)^2/2 - 1 \,. \tag{E.2}$$

Lemma E.1 gives a reduction of a ε-BAI problem to a BAI one on a modified instance, which is easier. Thanks to Lemma E.1, it is possible to leverage existing results on $T^\star(\nu)$, $T^\star_\beta(\nu)$, $w^\star(\nu)$, $w^\star_\beta(\nu)$ in order to study $T_\varepsilon(\nu)$, $T_{\varepsilon,\beta}(\nu)$, $T_{\varepsilon,\beta}(\nu, i)$, $w_\varepsilon(\nu)$, $w_{\varepsilon,\beta}(\nu)$ and $w_{\varepsilon,\beta}(\nu, i)$.

Lemma E.1. *Let $\mu \in \mathbb{R}^K$, $\varepsilon \geq 0$ and $\tilde{\varepsilon} \in [0, \varepsilon]$ and $\beta \in (0, 1)$. For all $i \in \mathcal{I}_{\tilde{\varepsilon}}(\mu)$, let $\nu_\varepsilon(i)$ be the instance with mean $\mu_\varepsilon(i)$ as $\mu_\varepsilon(i)_j = \mu_j - \varepsilon$ for all $j \neq i$ and $\mu_\varepsilon(i)_i = \mu_i$. Then, for all $i \in \mathcal{I}_{\tilde{\varepsilon}}(\mu)$, $T_{\varepsilon,\beta}(\nu, i) = T^\star_\beta(\nu_\varepsilon(i))$ and $w_{\varepsilon,\beta}(\nu, i) = w^\star_\beta(\nu_\varepsilon(i))$. Moreover, for all $i^\star \in i^\star(\mu)$,*

$$T_\varepsilon(\nu) = T^\star(\nu_\varepsilon(i^\star)) \quad \text{and} \quad T_{\varepsilon,\beta}(\nu) = T^\star_\beta(\nu_\varepsilon(i^\star)) \,.$$

Moreover, we have $w_\varepsilon(\nu) = \bigcup_{i^\star \in i^\star(\mu)} w^\star(\nu_\varepsilon(i^\star))$ and $w_{\varepsilon,\beta}(\nu) = \bigcup_{i^\star \in i^\star(\mu)} w^\star_\beta(\nu_\varepsilon(i^\star))$.

Proof. For $\varepsilon = 0$, the result is direct by definition. Let $\varepsilon > 0$ and $\tilde{\varepsilon} \in [0, \varepsilon]$. The first part is obtained by definition of $T^\star_{\varepsilon,\beta}(\nu, i)$, $w^\star_{\varepsilon,\beta}(\nu, i)$. Let $i^\star \in i^\star(\mu)$ and $\mu_\varepsilon(i^\star)$ defined as above, hence $i^\star(\nu_\varepsilon(i^\star)) = \{i^\star\}$. For all $i \in \mathcal{I}_\varepsilon(\mu) \setminus i^\star(\mu)$, let us denote by $\kappa_\varepsilon^{(i^\star,i)}$ be an instance with mean $\lambda_\varepsilon^{(i^\star,i)}$ such that $\lambda_{\varepsilon,i}^{(i^\star,i)} = \mu_i$ and $\lambda_{\varepsilon,j}^{(i^\star,i)} = \mu_j - \varepsilon$ for all $j \neq i$. If $\mu_{i^\star} - \mu_i = \varepsilon$ then $i^\star(\lambda_\varepsilon^{(i^\star,i)}) = \{i\} \cup i^\star(\mu) \setminus \{i^\star\}$, otherwise $i^\star(\lambda_\varepsilon^{(i^\star,i)}) = \{i\}$. We consider the permutation σ that swaps arm i with arm $i^\star$. By symmetry, we have $T^\star(\kappa_\varepsilon^{(i^\star,i)}) = T^\star(\sigma(\kappa_\varepsilon^{(i^\star,i)}))$. Moreover, we have that the gaps of $\sigma(\kappa_\varepsilon^{(i^\star,i)})$ are all strictly smaller than the gaps of μ_ε since $\varepsilon \geq \Delta_i > 0$. Therefore, Lemma 11 of Barrier et al (2022) yields that $T^\star(\sigma(\kappa_\varepsilon^{(i^\star,i)})) > T^\star(\nu_\varepsilon)$. We have proved that

$$\forall i^\star \in i^\star(\mu), \quad T^\star(\nu_\varepsilon(i^\star)) < \min_{i \in \mathcal{I}_\varepsilon(\mu) \setminus i^\star(\mu)} T^\star(\kappa_\varepsilon^{(i^\star,i)}) .$$

By symmetry $T^\star(\nu_\varepsilon(i^\star))$ is constant for all $i^\star \in i^\star(\mu)$, hence we have shown that

$$\forall i^\star \in i^\star(\mu), \quad T_\varepsilon(\nu) = T^\star(\nu_\varepsilon(i^\star)) .$$

It also shows that $w_\varepsilon(\nu) = \bigcup_{i^\star \in i^\star(\mu)} w^\star(\nu_\varepsilon(i^\star))$. The same reasoning yields the result for $T_{\varepsilon,\beta}(\nu)$ and $w_{\varepsilon,\beta}(\nu)$. ■

Lemma E.2 links the characteristic times for ε-BAI where $\varepsilon \in \{\varepsilon_0, \varepsilon_1\}$.

Lemma E.2. *Let $\mu \in \mathbb{R}^K$ such that $|i^\star(\mu)| = 1$. Let $\varepsilon_0 > \varepsilon_1$. Then, for all $\beta \in (0,1)$, we have*

$$T_{\varepsilon_0}(\nu)(\Delta_{\min}+\varepsilon_0)^2 \geq T_{\varepsilon_1}(\nu)(\Delta_{\min}+\varepsilon_1)^2 \quad \textit{and} \quad T_{\varepsilon_0,\beta}(\nu)(\Delta_{\min}+\varepsilon_0)^2 \geq T_{\varepsilon_1,\beta}(\nu)(\Delta_{\min}+\varepsilon_1)^2.$$

Let $\varepsilon_0 < \varepsilon_1$. Then,

$$T_{\varepsilon_0}(\nu)(\Delta_{\max}+\varepsilon_0)^2 \geq T_{\varepsilon_1}(\nu)(\Delta_{\max}+\varepsilon_1)^2 \quad \textit{and} \quad T_{\varepsilon_0,\beta}(\nu)(\Delta_{\max}+\varepsilon_0)^2 \geq T_{\varepsilon_1,\beta}(\nu)(\Delta_{\max}+\varepsilon_1)^2.$$

Proof. Let $i^\star(\mu) = \{i^\star\}$. Let $\varepsilon_0 > \varepsilon_1$. Using Lemma E.1, we have

$$2T_\varepsilon(\nu)^{-1}(\Delta_{\min}+\varepsilon)^{-2} = \max_{w \in \Sigma_K} \min_{j \neq i^\star} \frac{\tilde{\Delta}_j(\varepsilon)^2}{1/w_{i^\star} + 1/w_j} \quad \text{with} \quad \tilde{\Delta}_j(\varepsilon) = \frac{\mu_\star - \mu_j + \varepsilon}{\Delta_{\min} + \varepsilon} .$$

To conclude the first part of the first result, a sufficient condition is to show that $\tilde{\Delta}_j(\varepsilon_1) \geq \tilde{\Delta}_j(\varepsilon_0)$ for all $j \neq i^\star$. Direct manipulations show that, for all $j \neq i^\star$,

$$\tilde{\Delta}_j(\varepsilon_1) \geq \tilde{\Delta}_j(\varepsilon_0) \iff 1 - \frac{\varepsilon_0 - \varepsilon_1}{\mu_\star - \mu_j + \varepsilon_0} \geq 1 - \frac{\varepsilon_0 - \varepsilon_1}{\Delta_{\min} + \varepsilon_0} \iff \mu_\star - \mu_j \geq \Delta_{\min} ,$$

hence the result holds. The same proof can be used to obtain the second part of the first result.

Let $\varepsilon_0 < \varepsilon_1$. Using Lemma E.1, we have

$$2T_\varepsilon(\nu)^{-1}(\Delta_{\max}+\varepsilon)^{-2} = \max_{w\in\Sigma_K}\min_{j\neq i^\star}\frac{\bar\Delta_j(\varepsilon)^2}{1/w_{i^\star}+1/w_j} \quad \text{with} \quad \bar\Delta_j(\varepsilon) = \frac{\mu_\star-\mu_j+\varepsilon}{\Delta_{\max}+\varepsilon}\,.$$

To conclude the first part of the second result, a sufficient condition is to show that $\bar\Delta_j(\varepsilon_1) \geq \bar\Delta_j(\varepsilon_0)$ for all $j \neq i^\star$. Direct manipulations show that, for all $j \neq i^\star$,

$$\bar\Delta_j(\varepsilon_1) \geq \bar\Delta_j(\varepsilon_0) \iff 1+\frac{\varepsilon_1-\varepsilon_0}{\mu_\star-\mu_j+\varepsilon_0} \geq 1+\frac{\varepsilon_1-\varepsilon_0}{\Delta_{\max}+\varepsilon_0} \iff \mu_\star-\mu_j \leq \Delta_{\max}\,,$$

hence the result holds. The same proof can be used to obtain the second part of the second result. ■

E.2 Proof of Lemma 5.7

Let $i \in \mathcal{I}_{\varepsilon/2}(\mu)$, hence $\mu_i \geq \mu_\star - \varepsilon/2$. Let $\mu_\varepsilon(i)$ as in Lemma E.1 which satisfies that $i^\star(\nu_\varepsilon(i)) = \{i\}$. Let $\Delta_{i,j} = \mu_i - \mu_j$ and $\Delta_{i,i} = \mu_i - \max_{j\neq i}\mu_j$. Then, we have

$$T_{\varepsilon,1/2}(\nu,i) = T^\star_{1/2}(\nu_\varepsilon(i)) \leq 2T^\star(\nu_\varepsilon(i)) \leq 8\sum_{j\in[K]}(\Delta_{i,j}+\varepsilon)^{-2} \leq 8\sum_{j\in[K]}(\Delta_j+\varepsilon/2)^{-2}\,,$$

where we used $i \in \mathcal{I}_{\varepsilon/2}(\mu)$ for the last inequality. Since $\Delta_j \geq 0$, we conclude that $T_{\varepsilon,\beta}(\nu,i) \leq 32K/\varepsilon^2$. Then, we have

$$T_{\varepsilon,1/2}(\nu,i) = T^\star_{1/2}(\nu_\varepsilon(i)) \geq T^\star(\nu_\varepsilon(i)) \geq 2\sum_{j\in[K]}(\Delta_{i,j}+\varepsilon)^{-2}\,.$$

Likewise, using Lemma E.1 and (E.2), we obtain that, for all $j \neq i$,

$$\begin{aligned} w_{\varepsilon,1/2}(\nu)_j^{-1}/2 = w^\star_{1/2}(\nu_\varepsilon(i))_j^{-1}/2 &= T^\star_{1/2}(\nu_\varepsilon(i))(\mu_i-\mu_j+\varepsilon)^2/4 - 1 \\ &\leq 2\sum_{k\notin\{i,j\}}\left(\frac{\mu_i-\mu_j+\varepsilon}{\mu_i-\mu_k+\varepsilon}\right)^2 + 1 \end{aligned}$$

where the last inequality uses what we proved above. When $\mu_k \leq \mu_j$, the ratio is smaller than one. When $\mu_k > \mu_j$, we have $\mu_\star \geq \mu_k > \mu_j \geq \mu_\star - \varepsilon/2$, hence $\mu_k - \mu_j \leq \varepsilon/2$ and

$$\frac{\mu_i-\mu_j+\varepsilon}{\mu_i-\mu_k+\varepsilon} \leq 1+\frac{\varepsilon/2}{\mu_\star-\mu_k+\varepsilon/2} \leq 2\,.$$

Therefore, we obtain $w_{\varepsilon,1/2}(\nu)_j^{-1}/2 \leq 8(K-2)+1$, which concludes the result.

E.3 Proof of Lemma 5.8

Using the inclusion of events given by the assumption on $(A_{t,\delta}(n,\delta))_{n\geq t>K}$, we obtain

$$\begin{aligned}\sum_{t=K+1}^{n} \mathbb{1}\left(A_{t,\delta}(n,\delta)\right) &\leq \sum_{t=K+1}^{n} \mathbb{1}\left(\exists k_t \in [K],\, T_t(k_t) \leq D_{k_t}(n,\delta),\, T_{t+1}(k_t) = T_t(k_t)+1\right) \\ &\leq \sum_{i\in[K]} \sum_{t=K+1}^{n} \mathbb{1}\left(T_t(i) \leq D_i(n,\delta),\, T_{t+1}(i) = T_t(i)+1\right) \leq \sum_{i\in[K]} D_i(n,\delta)\,.\end{aligned}$$

The second inequality is obtained by union bound. The third inequality is direct since the number of times one can increase by one a quantity that is positive and bounded by $D_i(n,\delta)$ is at most $D_i(n,\delta)$.

E.4 Proof of Lemma 5.9

Let $s \geq 0$. For all $n > K$ and $\delta \in (0,1]$, let $\mathcal{E}_{n,\delta} = \mathcal{E}^1_{n,\delta} \cap \mathcal{E}^2_{n,\delta}$ with $(\mathcal{E}^1_{n,\delta})_{n>K}$ and $(\mathcal{E}^2_{n,\delta})_{n>K}$ as

$$\begin{aligned}\mathcal{E}^1_{n,\delta} &= \left\{\forall k \in [K], \forall t \leq n,\ |\mu_{t,k} - \mu_k| < \sqrt{\frac{2f_1(n,\delta)}{N_{t,k}}}\right\}\,, \\ \mathcal{E}^2_{n,\delta} &= \left\{\forall (i,k) \in [K]^2 \text{ s.t. } i \neq k,\ \forall t \leq n,\ \frac{|(\mu_{t,i} - \mu_{t,k}) - (\mu_i - \mu_k)|}{\sqrt{1/N_{t,i} + 1/N_{t,k}}} < \sqrt{2f_2(n,\delta)}\right\}\,,\end{aligned}$$

where $f_2(x,\delta) = \log(1/\delta) + (1+s)\log(x)$ and $f_2(x,\delta) = \log(1/\delta) + (2+s)\log(x)$.

Lemma E.3 shows that when there are arms with strictly higher true mean than the one of the leader, then at least one of those arms is undersampled.

Lemma E.3. *Under $\mathcal{E}^1_{n,\delta}$, for all $t \in [n] \setminus [K]$ let $B_t^{EB} = k$. Then,*

$$\forall i \neq k, \quad \mathbb{1}\left(\mu_i > \mu_k\right) \min\{N_{t,k}, N_{t,i}\} \leq \frac{8f_1(n,\delta)}{(\mu_i - \mu_k)^2}\,.$$

Proof. Under $\mathcal{E}^1_{n,\delta}$, for all $t \in [n] \setminus [K]$, let $B_t^{\mathrm{EB}} = k$. Then, for all $i \neq k$, we have

$$\begin{aligned}\mu_i - \sqrt{\frac{2f_1(n,\delta)}{\min\{N_{t,k}, N_{t,i}\}}} \leq \mu_i - \sqrt{\frac{2f_1(n,\delta)}{N_{t,i}}} \leq \mu_{t,i} \leq \mu_{t,k} &\leq \mu_k + \sqrt{\frac{2f_1(n,\delta)}{N_{t,k}}} \\ &\leq \mu_k + \sqrt{\frac{2f_1(n,\delta)}{\min\{N_{t,k}, N_{t,i}\}}}\,.\end{aligned}$$

Re-ordering the above equations for i such that $\mu_i > \mu_k$ yields the result. ■

Lemma E.4. *Let $\varepsilon \geq 0$, $\Delta_\mu(\varepsilon) = \min_{k \notin \mathcal{I}_\varepsilon(\mu)} \Delta_k$ and $C_{\mu,\varepsilon_0}(\varepsilon) = \max\{2\Delta_\mu(\varepsilon)^{-1} - \varepsilon_0^{-1}, \varepsilon_0^{-1}\}^2$. Let $A_{\varepsilon_0,\varepsilon,i} = 2/\Delta_\mu(\varepsilon)^2$ for all $i \in i^\star(\mu)$, $A_{\varepsilon_0,\varepsilon,i} = C_{\mu,\varepsilon_0}(\varepsilon)$ for all $i \in \mathcal{I}_\varepsilon(\mu) \setminus i^\star(\mu)$, otherwise $A_{\varepsilon_0,\varepsilon,i} = \max\{C_{\mu,\varepsilon_0}(\varepsilon), 2/\Delta_i^2\}$. For all $n > K$, under event $\mathcal{E}_{n,\delta}$, for all $t \in [n] \setminus [K]$ such that $B_t^{EB} \notin \mathcal{I}_\varepsilon(\mu)$, there exists $i_t \in [K]$ such that*

$$T_t(i_t) \leq \frac{4f_2(n,\delta)}{\min\{\beta, 1-\beta\}} A_{\varepsilon_0,\varepsilon,i_t} + 3(K-1)/2 \quad \textit{and} \quad T_{t+1}(i_t) = T_t(i_t) + 1 \,.$$

Proof. Let $\Delta_i = \mu_\star - \mu_i$ and $\Delta_{\max} = \max_{i \in [K]} \Delta_i$. When $\varepsilon \geq \Delta_{\max}$, we have $\mathcal{I}_\varepsilon(\mu)^\complement = \emptyset$, hence the above result holds trivially since the event $B_t^{\mathrm{EB}} \notin \mathcal{I}_\varepsilon(\mu)$ cannot happen. Let $\varepsilon \in [0, \Delta_{\max})$, *i.e.* $\mathcal{I}_\varepsilon(\mu)^\complement \neq \emptyset$. We will consider in two distinct cases since

$$\{B_t \notin \mathcal{I}_\varepsilon(\mu)\} = \{B_t \notin \mathcal{I}_\varepsilon(\mu), C_t \in i^\star(\mu)\} \cup \{B_t \notin \mathcal{I}_\varepsilon(\mu), C_t \notin i^\star(\mu)\} \,.$$

Case 1. Let $t \in [n] \setminus [K]$ such that $(B_t^{\mathrm{EB}}, C_t^{\mathrm{TC}\varepsilon_0}) = (i,j)$ with $i \notin \mathcal{I}_\varepsilon(\mu)$ and $j \in i^\star(\mu)$. Using Lemmas 5.3 and E.3, we obtain

$$\min\{\beta, 1-\beta\} \left(\min\{T_t(i), T_t(j)\} - 3(K-1)/2\right) \leq \min\{N_{t,i}, N_{t,j}\} \leq \frac{8f_1(n,\delta)}{\Delta_i^2} \leq \frac{8f_2(n,\delta)}{\Delta_i^2} \,,$$

which can be rewritten as

$$\min\{T_t(i), T_t(j)\} \leq \frac{8f_2(n,\delta)}{\min\{\beta, 1-\beta\}\Delta_i^2} + 3(K-1)/2 \,.$$

Let us define $\Delta_\mu(\varepsilon) = \min_{k \notin \mathcal{I}_\varepsilon(\mu)} \Delta_k$, and

$$\forall i \notin \mathcal{I}_\varepsilon(\mu), \quad D_{\varepsilon,i} = 2/\Delta_i^2 \quad \text{and} \quad \forall i \in i^\star(\mu), \quad D_{\varepsilon,i} = 2/\Delta_\mu(\varepsilon)^2 \,.$$

The above shows that there exists $k_t \in \mathcal{I}_\varepsilon(\mu)^\complement \cup i^\star(\mu)$ such that

$$T_t(k_t) \leq \frac{4f_2(n,\delta)}{\min\{\beta, 1-\beta\}} D_{\varepsilon,k_t} + 3(K-1)/2 \quad \text{and} \quad T_{t+1}(k_t) = T_t(k_t) + 1 \,.$$

Case 2. Let $t \in [n] \setminus [K]$ such that $(B_t^{\mathrm{EB}}, C_t^{\mathrm{TC}\varepsilon_0}) = (i,j)$ with $i \notin \mathcal{I}_\varepsilon(\mu)$ and $j \notin i^\star(\mu)$. Let $i_0 \in i^\star(\mu)$. Using the TC challenger, we obtain

$$\frac{\varepsilon_0 - \Delta_i}{\sqrt{1/N_{t,i} + 1/N_{t,i_0}}} + \sqrt{2f_2(n,\delta)} \geq \frac{\mu_{t,i} - \mu_{t,i_0} + \varepsilon_0}{\sqrt{1/N_{t,i} + 1/N_{t,i_0}}} \geq \frac{\mu_{t,i} - \mu_{t,j} + \varepsilon_0}{\sqrt{1/N_{t,i} + 1/N_{t,j}}}$$

$$\geq \varepsilon_0 \sqrt{\min\{N_{t,i}, N_{t,j}\}/2}\,.$$

Using Lemma E.3, we obtain

$$\frac{1}{1/N_{t,i} + 1/N_{t,i_0}} \leq \min\{N_{t,i}, N_{t,i_0}\} \leq \frac{8 f_1(n,\delta)}{\Delta_i^2} \leq \frac{8 f_2(n,\delta)}{\Delta_i^2}\,.$$

By distinguishing between $\varepsilon_0 > \Delta_i$ and $\varepsilon_0 \leq \Delta_i$ and using that $\Delta_i > 0$, we have

$$\frac{\varepsilon_0 - \Delta_i}{\sqrt{1/N_{t,i} + 1/N_{t,i_0}}} + \sqrt{2 f_2(n,\delta)} \leq \max\{2\varepsilon_0/\Delta_i - 1, 1\}\sqrt{2 f_2(n,\delta)}\,.$$

Using Lemma 5.3 to lower bound $\min\{N_{t,i}, N_{t,j}\}$ and reordering, we have shown that

$$\min\{T_t(i), T_t(j)\} \leq \max\left\{\left(\frac{2}{\Delta_i} - \frac{1}{\varepsilon_0}\right)^2, \frac{1}{\varepsilon_0^2}\right\} \frac{4 f_2(n,\delta)}{\min\{\beta, 1-\beta\}} + 3(K-1)/2\,.$$

Let us define $C_{\mu,\varepsilon_0}(\varepsilon) = \max\{2\Delta_\mu(\varepsilon)^{-1} - \varepsilon_0^{-1}, \varepsilon_0^{-1}\}^2$. The above shows that, there exists $k_t \notin i^\star(\mu)$ such that

$$T_t(k_t) \leq \frac{4 f_2(n,\delta)}{\min\{\beta, 1-\beta\}} C_{\mu,\varepsilon_0}(\varepsilon) + 3(K-1)/2 \quad \text{and} \quad T_{t+1}(k_t) = T_t(k_t) + 1\,.$$

Summary. Let us define $(A_{\varepsilon_0,\varepsilon,i})_{i\in[K]}$ as in the statement of Lemma E.4. Under $\mathcal{E}_{n,\delta}$, we have show that, when $B_t \notin \mathcal{I}_\varepsilon(\mu)$, there exists $i_t \in [K]$ such that

$$T_t(k_t) \leq \frac{4 f_2(n,\delta)}{\min\{\beta, 1-\beta\}} A_{\varepsilon,\varepsilon_0,i_t} + 3(K-1)/2 \quad \text{and} \quad T_{t+1}(k_t) = T_t(k_t) + 1\}\,.$$

■

For all $n > K$, under event $\mathcal{E}_{n,\delta}$, combining Lemma 5.8 and E.4 for $A_t(n,\delta) = \{B_t^{\mathrm{EB}} \notin \mathcal{I}_\varepsilon(\mu)\}$ and $D_i(n,\delta) = \frac{4 f_2(n,\delta)}{\min\{\beta,1-\beta\}} A_{\varepsilon_0,\varepsilon,i} + 3(K-1)/2$ yields that

$$\sum_{t=K+1}^{n} \mathbb{1}\left(B_t^{\mathrm{EB}} \notin \mathcal{I}_\varepsilon(\mu)\right) \leq \frac{4 f_2(n,\delta)}{\min\{\beta, 1-\beta\}} H_{\mu,\varepsilon_0}(\varepsilon) + 3K(K-1)/2\,.$$

where we used that $\sum_{i\in[K]} A_{\varepsilon_0,\varepsilon,i} = H_{\mu,\varepsilon_0}(\varepsilon)$ where $H_{\mu,\varepsilon_0}(\varepsilon)$ is defined in (5.5). To conclude the proof of Lemma 5.9, we use that

$$\sum_{t=K+1}^{n} \mathbb{1}\left(B_t^{\mathrm{EB}} \notin \mathcal{I}_\varepsilon(\mu)\right) = n - 1 - \sum_{i\in\mathcal{I}_\varepsilon(\mu)} \sum_{j} T_n(i,j)\,.$$

E.5 Proof of Lemma 5.13

Let $|\mathcal{I}_\varepsilon(\mu)| = I_\varepsilon$ and $g_{\varepsilon,\varepsilon_0}(n,\delta) = \frac{4H_{\mu,\varepsilon_0}(\varepsilon)}{\min\{\beta,1-\beta\}}\tilde{f}_2(n,\delta) + 3K(K-1)/2$. Using Lemma 5.9 and the pigeonhole principle, there exists $i_0 \in \mathcal{I}_\varepsilon(\mu)$ such that

$$\sum_j T_n(i_0,j) \geq \left(n - 1 - g_{\varepsilon,\varepsilon_0}(n,\delta)\right)/I_\varepsilon \,.$$

Therefore, we have

$$N_{n,i_0} \geq \beta \sum_j T_n(i_0,j) - (K-1) \geq \beta\left(n - 1 - g_{\varepsilon,\varepsilon_0}(n,\delta)\right)/I_\varepsilon - (K-1) \,.$$

Let $S_{\varepsilon,\varepsilon_1,\varepsilon_0,\mu}(\delta)$ defined as in the statement of Lemma 5.13. Direct manipulations shows that

$$S_{\varepsilon,\varepsilon_1,\varepsilon_0,\mu}(\delta) \geq \sup\left\{ n \mid n - 1 \leq g_{\varepsilon,\varepsilon_0}(n,\delta) + \frac{I_\varepsilon}{\beta}\left(4\tilde{f}_2(n,\delta)\max_{i\notin\mathcal{I}_{\varepsilon_1}(\mu)} C_{\varepsilon,i} + K - 1\right)\right\} \,.$$

Therefore, we have $N_{n,i_0} > 4\tilde{f}_2(n,\delta)\max_{i\notin\mathcal{I}_{\varepsilon_1}(\mu)} C_{\varepsilon,i}$ for all $n > S_{\varepsilon,\varepsilon_1,\varepsilon_0,\mu}(\delta)$.

Let $n > S_{\varepsilon,\varepsilon_1,\varepsilon_0,\mu}(\delta)$. Suppose that $B_n^{\mathrm{EB}} = i \notin \mathcal{I}_{\varepsilon_1}(\mu)$. Using Lemma E.3, under the event $\tilde{\mathcal{E}}_{n,\delta}$, we obtain that

$$\min\{N_{n,i}, N_{n,i_0}\} \leq \frac{8\tilde{f}_1(n,\delta)}{(\mu_{i_0} - \mu_i)^2} \leq 4C_{\varepsilon,i}\tilde{f}_2(n,\delta) \,.$$

Suppose towards contradiction that $\min\{N_{n,i}, N_{n,i_0}\} = N_{n,i_0}$, then $N_{n,i_0} \leq 4C_{\varepsilon,i}\tilde{f}_2(n,\delta)$. This is a direct contradiction with $N_{n,i_0} > 4\tilde{f}_2(n,\delta)\max_{i\notin\mathcal{I}_{\varepsilon_1}(\mu)} C_{\varepsilon,i}$ since $n > S_{\varepsilon,\varepsilon_1,\varepsilon_0,\mu}(\delta)$. Therefore, we have shown that $\min\{N_{n,i}, N_{n,i_0}\} = N_{n,i}$, hence $i \in U_{\varepsilon,\varepsilon_1,n}(n,\delta)$. This concludes the proof.

E.6 Proof of Lemma 5.14

We will be interested in three distinct cases since

$$\begin{aligned}
\{U_{\varepsilon,\varepsilon_1,t}(n,\delta) \neq \emptyset\} &= \left\{U_{\varepsilon,\varepsilon_1,t}(n,\delta) \cap \{B_t^{\mathrm{EB}}, C_t^{\mathrm{TC}}\} \neq \emptyset\right\} \\
&\cup \left\{U_{\varepsilon,\varepsilon_1,t}(n,\delta) \neq \emptyset,\ U_{\varepsilon,\varepsilon_1,t}(n,\delta) \cap \{B_t^{\mathrm{EB}}, C_t^{\mathrm{TC}}\} = \emptyset,\ B_t^{\mathrm{EB}} \notin \mathcal{I}_\varepsilon(\mu)\right\} \\
&\cup \left\{U_{\varepsilon,\varepsilon_1,t}(n,\delta) \neq \emptyset,\ U_{\varepsilon,\varepsilon_1,t}(n,\delta) \cap \{B_t^{\mathrm{EB}}, C_t^{\mathrm{TC}}\} = \emptyset,\ B_t^{\mathrm{EB}} \in \mathcal{I}_\varepsilon(\mu)\right\} \,,
\end{aligned}$$

Case 1. Let $t \in [n] \setminus [K]$ such that $\{B_t^{\mathrm{EB}}, C_t^{\mathrm{TC}}\} \cap U_{\varepsilon,\varepsilon_1,t}(n,\delta) \neq \emptyset$. Let $k_t \in \{B_t^{\mathrm{EB}}, C_t^{\mathrm{TC}}\} \cap U_{\varepsilon,\varepsilon_1,t}(n,\delta)$. For this $k_t \notin \mathcal{I}_{\varepsilon_1}(\mu)$, we have $T_{t+1}(k_t) = T_t(k_t) + 1$ and, by combining Lemma E.3

and $N_{t,k_t} \leq 4C_{\varepsilon,k_t}\tilde{f}_2(n,\delta)$, we obtain that

$$T_t(k_t) \leq \frac{N_{t,k_t}}{\min\{\beta, 1-\beta\}} + \frac{3(K-1)}{2} \leq \frac{4\tilde{f}_2(n,\delta)}{\min\{\beta, 1-\beta\}} C_{\varepsilon,k_t} + 3(K-1)/2 \,.$$

Case 2. Let $t \in [n] \setminus [K]$ such that $U_{\varepsilon,\varepsilon_1,t}(n,\delta) \neq \emptyset$, $U_{\varepsilon,\varepsilon_1,t}(n,\delta) \cap \{B_t^{\mathrm{EB}}, C_t^{\mathrm{TC}}\} = \emptyset$ and $B_t^{\mathrm{EB}} \notin \mathcal{I}_\varepsilon(\mu)$. Let $\Delta_\mu(\varepsilon)$ and $C_{\mu,\varepsilon_0}(\varepsilon)$ defined as in the statement of Lemma 5.14. Let $D_{\varepsilon_0,\varepsilon,i} = 2/\Delta_\mu(\varepsilon)^2$ for all $i \in i^\star(\mu)$, $D_{\varepsilon_0,\varepsilon,i} = C_{\mu,\varepsilon_0}(\varepsilon)$ for all $i \in \mathcal{I}_\varepsilon(\mu) \setminus i^\star(\mu)$, otherwise $D_{\varepsilon_0,\varepsilon,i} = \max\{C_{\mu,\varepsilon_0}(\varepsilon), 2/\Delta_i^2\}$. Using Lemma E.4, there exists $k_t \in [K]$ such that

$$T_t(k_t) \leq \frac{4\tilde{f}_2(n,\delta)}{\min\{\beta, 1-\beta\}} D_{\varepsilon,\varepsilon_0,k_t} + 3(K-1)/2 \quad \text{and} \quad T_{t+1}(k_t) = T_t(k_t) + 1 \,.$$

Case 3. Let $t \in [n] \setminus [K]$ such that $U_{\varepsilon,\varepsilon_1,t}(n,\delta) \neq \emptyset$, $U_{\varepsilon,\varepsilon_1,t}(n,\delta) \cap \{B_t^{\mathrm{EB}}, C_t^{\mathrm{TC}}\} = \emptyset$ and $B_t^{\mathrm{EB}} \in \mathcal{I}_\varepsilon(\mu)$. Let $j_0 \in U_{\varepsilon,\varepsilon_1,t}(n,\delta) \setminus \{B_t^{\mathrm{EB}}, C_t^{\mathrm{TC}}\}$, which is possible since $U_{\varepsilon,\varepsilon_1,t}(n,\delta) \cap \{B_t^{\mathrm{EB}}, C_t^{\mathrm{TC}}\} = \emptyset$ and $U_{\varepsilon,\varepsilon_1,t}(n,\delta) \neq \emptyset$. Let us denote by $(B_t^{\mathrm{EB}}, C_t^{\mathrm{TC}}) = (i,j)$ with $i \in \mathcal{I}_\varepsilon(\mu)$ and $j \neq j_0$. Using the TC challenger, under the event $\mathcal{E}_{n,\delta}$, we obtain

$$\begin{aligned}\frac{\mu_i - \mu_{j_0} + \varepsilon_0}{\sqrt{1/N_{t,i} + 1/N_{t,j_0}}} + \sqrt{2\tilde{f}_2(n,\delta)} \geq \frac{\mu_{t,i} - \mu_{t,j_0} + \varepsilon_0}{\sqrt{1/N_{t,i} + 1/N_{t,j_0}}} \geq \frac{\mu_{t,i} - \mu_{t,j} + \varepsilon_0}{\sqrt{1/N_{t,i} + 1/N_{t,j}}} \\ \geq \varepsilon_0 \sqrt{\min\{N_{t,i}, N_{t,j}\}/2} \,.\end{aligned}$$

Since $\mu_i - \mu_{j_0} + \varepsilon_0 \geq \varepsilon_0 > 0$, we have

$$\begin{aligned}\frac{\mu_i - \mu_{j_0} + \varepsilon_0}{\sqrt{1/N_{t,i} + 1/N_{t,j_0}}} &\leq \sqrt{N_{t,j_0}}(\mu_i - \mu_{j_0} + \varepsilon_0) + \sqrt{2\tilde{f}_2(n,\delta)} \\ &\leq \left(2\frac{\mu_i - \mu_{j_0} + \varepsilon_0}{\min_{k \in \mathcal{I}_\varepsilon(\mu)}(\mu_k - \mu_{j_0})} + 1\right)\sqrt{2\tilde{f}_2(n,\delta)} \,.\end{aligned}$$

Using Lemma E.3 to lower bound $\min\{N_{t,i}, N_{t,j}\}$ and $\mu_i - \mu_{j_0} \leq \Delta_{j_0}$ for all $i \in \mathcal{I}_\varepsilon(\mu)$, direct manipulation yields that

$$\min\{T_t(i), T_t(j)\} \leq \frac{4\tilde{f}_2(n,\delta)}{\min\{\beta, 1-\beta\}} C_{\mu,\varepsilon_0}(\varepsilon,\varepsilon_1)^2 + 3(K-1)/2 \,.$$

where $C_{\mu,\varepsilon_0}(\varepsilon,\varepsilon_1)$ is defined as in the statement of Lemma 5.14. For all $i \notin \mathcal{I}_{\varepsilon_1}(\mu)$, it is direct to see that $C_{\mu,\varepsilon_0}(\varepsilon,\varepsilon_1)^2 \geq \max\{1/\varepsilon_0^2, \max_{i \notin \mathcal{I}_{\varepsilon_1}(\mu)} C_{\varepsilon,i}\}$ and $C_{\varepsilon,i} \geq 2/\Delta_i^2$. The above shows that there exists $k_t \in [K]$ such that

$$T_t(k_t) \leq \frac{4\tilde{f}_2(n,\delta)}{\min\{\beta, 1-\beta\}} C_{\mu,\varepsilon_0}(\varepsilon,\varepsilon_1)^2 + 3(K-1)/2 \quad \text{and} \quad T_{t+1}(k_t) = T_t(k_t) + 1 \,.$$

Summary. Let us define $(A_{\varepsilon,\varepsilon_1,\varepsilon_0,i})_{i\in[K]}$ as in the statement of Lemma 5.14. Under $\tilde{\mathcal{E}}_{n,\delta}$, we have show that, when $U_{\varepsilon,\varepsilon_1,t}(n,\delta) \neq \emptyset$, there exists $i_t \in [K]$ such that

$$T_t(i_t) \leq \frac{4\tilde{f}_2(n,\delta)}{\min\{\beta, 1-\beta\}} A_{\varepsilon,\varepsilon_1,\varepsilon_0,i_t} + 3(K-1)/2 \quad \text{and} \quad T_{t+1}(i_t) = T_t(i_t) + 1 \,.$$

E.7 Proof of Lemma 5.15

For all $n > K$, under event $\tilde{\mathcal{E}}_{n,\delta}$, combining Lemma 5.8 and 5.14 for $A_t(n,\delta) = \{U_{\varepsilon,\varepsilon_1,t}(n,\delta) \neq \emptyset\}$ and $D_i(n,\delta) = \frac{4\tilde{f}_2(n,\delta)}{\min\{\beta,1-\beta\}} A_{\varepsilon,\varepsilon_1,\varepsilon_0,i} + 3(K-1)/2$ yields that

$$\sum_{t=K+1}^{n} \mathbb{1}\left(U_{\varepsilon,\varepsilon_1,t}(n,\delta) \neq \emptyset\right) \leq \frac{4\tilde{f}_2(n,\delta)}{\min\{\beta, 1-\beta\}} \bar{H}_{\varepsilon,\varepsilon_1}(\mu,\varepsilon_0) + 3K(K-1)/2 \,.$$

where we used that $\sum_{i\in[K]} A_{\varepsilon,\varepsilon_1,\varepsilon_0,i} = \bar{H}_{\varepsilon,\varepsilon_1}(\mu,\varepsilon_0)$ where $\bar{H}_{\varepsilon,\varepsilon_1}(\mu,\varepsilon_0)$ is defined in (5.8).

For all $i \notin \mathcal{I}_{\varepsilon_1}(\mu)$, let us define

$$t_i(n,\delta) = \max\left\{t \in [n] \setminus [K] \mid \; i \in U_{\varepsilon,\varepsilon_1,t}(n,\delta)\right\} \,.$$

By definition, for all $i \notin \mathcal{I}_{\varepsilon_1}(\mu)$, we have $i \in U_{\varepsilon,\varepsilon_1,t}(n,\delta)$ for all $t \in (K, t_i(n,\delta)]$ and $i \notin U_{\varepsilon,\varepsilon_1,t}(n,\delta)$ for all $t \in (t_i(n,\delta), n]$. Therefore, for all $t \in (K, \max_{i\notin\mathcal{I}_{\varepsilon_1}(\mu)} t_i(n,\delta)]$, we have $U_{\varepsilon,\varepsilon_1,t}(n,\delta) \neq \emptyset$ and $U_{\varepsilon,\varepsilon_1,t}(n,\delta) = \emptyset$ for all $t > \max_{i\notin\mathcal{I}_{\varepsilon_1}(\mu)} t_i(n,\delta)$, hence

$$\begin{aligned}
\max_{i\notin\mathcal{I}_{\varepsilon_1}(\mu)} (t_i(n,\delta) - K) &= \sum_{t=K+1}^{n} \mathbb{1}\left(U_{\varepsilon,\varepsilon_1,t}(n,\delta) \neq \emptyset\right) \\
&\leq \frac{4\tilde{f}_2(n,\delta)}{\min\{\beta, 1-\beta\}} \bar{H}_{\varepsilon,\varepsilon_1}(\mu,\varepsilon_0) + 3K(K-1)/2 \,.
\end{aligned}$$

Let $T_{\varepsilon,\varepsilon_1,\varepsilon_0,\mu}(\delta)$ defined as in the statement of Lemma 5.15. Direct manipulations show that

$$T_{\varepsilon,\varepsilon_1,\varepsilon_0,\mu}(\delta) \geq \sup\left\{n \mid n - K \leq \frac{4\tilde{f}_2(n,\delta)}{\min\{\beta, 1-\beta\}} \bar{H}_{\varepsilon,\varepsilon_1}(\mu,\varepsilon_0) + 3K(K-1)/2\right\} \,.$$

Let $n > T_{\varepsilon,\varepsilon_1,\varepsilon_0,\mu}(\delta)$. Then, we have

$$n - K > \frac{4\tilde{f}_2(n,\delta)}{\min\{\beta, 1-\beta\}} \bar{H}_{\varepsilon,\varepsilon_1}(\mu,\varepsilon_0) + 3K(K-1)/2 \geq \max_{i\notin\mathcal{I}_{\varepsilon_1}(\mu)} (t_i(n,\delta) - K) \,,$$

hence $n > \max_{i\notin\mathcal{I}_{\varepsilon_1}(\mu)} t_i(n,\delta)$. This conclude the proof that $U_{\varepsilon,\varepsilon_1,n}(n,\delta) = \emptyset$.

E.8 Inversion Result

Lemma E.5 is an inversion result to upper bound a probability which is implicitly defined based on times that are implicitly defined.

Lemma E.5. *Let $\overline{W}_{-1}$ as in Appendix A. Let $A > 0$, $B > 0$, $C > 0$, $E > 0$, $\alpha > 0$, $\beta > 0$ and*

$$D_{A,B,C}(\delta) = \sup\{x \mid x \leq A(\log(1/\delta) + C\log x) + B\} ,$$
$$D_{A,B,C,E,\alpha,\beta}(\delta) = \sup\left\{x \mid x \leq \frac{A}{\alpha}\overline{W}_{-1}\left(\alpha\left(\log(1/\delta) + C\log(\beta + \log x) + E\right)\right) + B\right\} .$$

Then,

$$\inf\{\delta \mid x > D_{A,B,C}(\delta)\} \leq x^C \exp\left(-\frac{x-B}{A}\right) ,$$
$$\inf\{\delta \mid x > D_{A,B,C,E,\alpha,\beta}(\delta)\} \leq e^E \left(\alpha\frac{x-B}{A}\right)^{1/\alpha} (\beta + \log x)^C \exp\left(-\frac{x-B}{A}\right) .$$

Suppose that $B/A + \log A > 1$ and $C(A,B) = \sup\{x \mid x < A\log x + B\}$. Then, $C(A,B) < h_1(A,B)$ with $h_1(z,y) = z\overline{W}_{-1}(y/z + \log z)$.

Proof. Direct manipulations yield that

$$x > D_{A,B,C}(\delta) \iff x > A(\log(1/\delta) + C\log x) + B \iff \delta < x^C \exp\left(-\frac{x-B}{A}\right) .$$

Likewise, using Lemma A.1, we obtain

$$\begin{aligned} x > D_{A,B,C,E,\alpha,\beta}(\delta) &\iff \alpha\frac{x-B}{A} > \overline{W}_{-1}\left(\alpha\left(\log(1/\delta) + C\log(\beta + \log x) + E\right)\right) \\ &\iff \frac{x-B}{A} - \frac{1}{\alpha}\log\left(\alpha\frac{x-B}{A}\right) > \log(1/\delta) + C\log(\beta + \log x) + E \\ &\iff \delta < e^E \left(\alpha\frac{x-B}{A}\right)^{1/\alpha} (\beta + \log x)^C \exp\left(-\frac{x-B}{A}\right) . \end{aligned}$$

Since $B/A + \log A > 1$, we have $C(A,B) \geq A$, hence

$$C(A,B) = \sup\{x \mid x < A\log(x) + B\} = \sup\{x \geq A \mid x < A\log(x) + B\} .$$

Using Lemma A.1 yields that

$$x \geq A\log x + B \iff \frac{x}{A} - \log\left(\frac{x}{A}\right) \geq \frac{B}{A} + \log A \iff x \geq A\overline{W}_{-1}\left(\frac{B}{A} + \log A\right) .$$

■

Appendix F

Complements on Chapter 6

F.1 Proof of Lemma 6.4

Using the inclusion of events given by the assumption on $(A_t(n,\delta))_{K<t\leq n}$, we obtain

$$\begin{aligned}\sum_{t=K+1}^{T} \mathbb{1}\left(A_t(n,\delta)\right) &\leq \sum_{t=K+1}^{T} \mathbb{1}\left(N_{t,i_t} \leq D_{i_t}(n,\delta),\ N_{t+1,i_t} = N_{t,i_t}+1\right) \\ &\leq \sum_{i\in[K]} \sum_{t=K+1}^{T} \mathbb{1}\left(N_{t,i} \leq D_i(n,\delta),\ N_{t+1,i} = N_{t,i}+1\right) \leq \sum_{i\in[K]} D_i(n,\delta)\,.\end{aligned}$$

The second inequality is obtained by union bound. The third inequality is direct since the number of times one can increment by one a quantity that is positive and bounded by $D_i(n,\delta)$ is at most $D_i(n,\delta)$.

F.2 Proof of Lemma 6.6

We will be interested in three distinct cases since

$$\begin{aligned}\{U_t(n,\delta)=\emptyset\} \quad = \quad & \underbrace{\{U_t(n,\delta)=\emptyset,\ \max_{i\in[K]}\mu_{t,i} > \gamma\}}_{\text{Case 1}} \cup \underbrace{\{U_t(n,\delta)=\emptyset,\ \max_{i\in[K]}\mu_{t,i} < \gamma\}}_{\text{Case 2}} \\ & \cup \underbrace{\{U_t(n,\delta)=\emptyset,\ \max_{i\in[K]}\mu_{t,i} = \gamma\}}_{\text{Case 3}}\end{aligned}$$

Case 1. Let $k = \arg\max_{i\in[K]} \mu_{t,k}$. Since $I_t \in \arg\max_{i\in[K]} \sqrt{N_{t,i}}(\mu_{t,i}-\gamma)_+$ and $\sqrt{N_{t,k}}(\mu_{t,k}-\gamma)_+ > 0$, we obtain $\mu_{t,I_t} > \gamma$. Under $\tilde{\mathcal{E}}_{n,\delta}$, we have

$$0 < \sqrt{N_{t,I_t}}(\mu_{t,I_t}-\gamma)_+ = \sqrt{N_{t,I_t}}(\mu_{t,I_t}-\gamma) \leq \sqrt{N_{t,I_t}}(\mu_{I_t}-\gamma) + \sqrt{2\tilde{f}_1(n,\delta)}\,,$$

hence $N_{t,I_t} \leq 2\tilde{f}_1(n,\delta)\Delta_{\gamma,I_t}^{-2}$ and $N_{t+1,I_t} = N_{t,I_t}+1$.

Case 2. Let $I_t \in \arg\min_{i\in[K]} \sqrt{N_{t,i}}(\gamma-\mu_{t,i})_+$ and $i \in U_t(n,\delta)$. Under $\tilde{\mathcal{E}}_{n,\delta}$, we have

$$\begin{aligned}
&\sqrt{N_{t,I_t}}(\gamma-\mu_{I_t}) - \sqrt{2\tilde{f}_1(n,\delta)} \leq \sqrt{N_{t,I_t}}(\gamma-\mu_{t,I_t}) = \sqrt{N_{t,I_t}}(\gamma-\mu_{t,I_t})_+\,,\\
&\sqrt{N_{t,i}}(\gamma-\mu_{t,i})_+ = \sqrt{N_{t,i}}(\gamma-\mu_{t,i}) \leq \sqrt{N_{t,i}}(\gamma-\mu_i) + \sqrt{2\tilde{f}_1(n,\delta)} \leq 2\sqrt{2\tilde{f}_1(n,\delta)}\,.
\end{aligned}$$

Using that $\sqrt{N_{t,I_t}}(\gamma-\mu_{t,I_t})_+ \leq \sqrt{N_{t,i}}(\gamma-\mu_{t,i})_+$, we have proven that $N_{t,I_t} \leq 18\tilde{f}_1(n,\delta)\Delta_{\gamma,I_t}^{-2}$ and $N_{t+1,I_t} = N_{t,I_t}+1$.

Case 3. Then, $\arg\min_{i\in[K]} \sqrt{N_{t,i}}(\gamma-\mu_{t,i})_+ = \{i\in[K] \mid \mu_{t,i}=\gamma\}$. Therefore, we have $\mu_{t,I_t} = \gamma$ hence $\gamma = \mu_{t,I_t} \leq \mu_{I_t} + \sqrt{2\tilde{f}_1(n,\delta)/N_{t,I_t}}$. Therefore, we have proven that $N_{t,I_t} \leq 2\tilde{f}_1(n,\delta)\Delta_{\gamma,I_t}^{-2}$ and $N_{t+1,I_t} = N_{t,I_t}+1$.

Summary. Combing the three above cases yields the result.

F.3 Proof of Lemma 6.7

Combining Lemmas 6.6 and 5.8, we obtain $\sum_{t=K+1}^{n} \mathbb{1}\left(U_t(n,\delta)\neq\emptyset\right) \leq 18H_1(\mu)\tilde{f}_1(n,\delta)$. For all $i\in[K]$, let us define $t_i(n,\delta) = \max\{t\in(K,n]\cap\mathbb{N} \mid i\in U_t(n,\delta)\}$. By definition, we have $i\in U_t(n,\delta)$ for all $t\in(K,t_i(n,\delta)]$ and $i\notin U_t(n,\delta)$ for all $t\in(t_i(n,\delta),n]$. Therefore, for all $t\in(K,\max_{i\in[K]} t_i(n,\delta)]$, we have $U_t(n,\delta)\neq\emptyset$ and $U_t(n,\delta)=\emptyset$ for all $t > \max_{i\in[K]} t_i(n,\delta)$, hence $\max_{i\in[K]}(t_i(n,\delta)-K) = \sum_{t=K+1}^{n}\mathbb{1}\left(U_t(n,\delta)\neq\emptyset\right) \leq 18H_1(\mu)\tilde{f}_1(n,\delta)$. Let $T_\mu(\delta)$ defined as in the statement of Lemma 6.7 and $n > T_\mu(\delta)$. Then, we have

$$n-K > 18H_1(\mu)\tilde{f}_1(n,\delta) \geq \max_{i\in[K]}(t_i(n,\delta)-K)\,,$$

hence $n > \max_{i\in[K]} t_i(n,\delta)$. This concludes the proof that $U_n(n,\delta)=\emptyset$.

F.4 Proof of Lemma 6.9

Let $t\leq n$ such that $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)\subseteq U_t(n,\delta)$. When $I_t\in\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$, we have directly that $N_{t,I_t} \leq \left(\sqrt{2\tilde{f}_1(n,\delta)\Delta_{\gamma,I_t}^{-2}}+1\right)^2$ and $N_{t+1,I_t} = N_{t,I_t}+1$. In the following, we consider $I_t\notin\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$. We

will be interested in three cases since

$$\{\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \subseteq U_t(n,\delta),\ I_t \notin \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)\} = \underbrace{\{\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \subseteq U_t(n,\delta),\ I_t \notin \mathcal{I}_\gamma^{\mathrm{thr}}(\mu),\ \max_{i\in[K]} \mu_{t,i} > \gamma\}}_{\text{Case 1}}$$
$$\cup \underbrace{\{\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \subseteq U_t(n,\delta),\ I_t \notin \mathcal{I}_\gamma^{\mathrm{thr}}(\mu),\ \max_{i\in[K]} \mu_{t,i} < \gamma\}}_{\text{Case 2}}$$
$$\cup \underbrace{\{\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \subseteq U_t(n,\delta),\ I_t \notin \mathcal{I}_\gamma^{\mathrm{thr}}(\mu),\ \max_{i\in[K]} \mu_{t,i} = \gamma\}}_{\text{Case 3}} .$$

Case 1. Let $k = \arg\max_{i\in[K]} \mu_{t,i}$. Since $I_t \in \arg\max_{i\in[K]} \sqrt{N_{t,i}}(\mu_{t,i}-\gamma)_+$ and $\sqrt{N_{t,k}}(\mu_{t,k}-\gamma)_+ > 0$, we have $\mu_{t,I_t} > \gamma$. Since $I_t \notin \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$, under $\tilde{\mathcal{E}}_{n,\delta}$, we have

$$\sqrt{N_{t,I_t}}(\mu_{t,I_t}-\gamma)_+ = \sqrt{N_{t,I_t}}(\mu_{t,I_t}-\gamma) \le \sqrt{N_{t,I_t}}(\mu_{I_t}-\gamma) + \sqrt{2\tilde{f}_1(n,\delta)} .$$

Using that $\sqrt{N_{t,I_t}}(\mu_{t,I_t}-\gamma)_+ > 0$, we obtain $N_{t,I_t} \le 2\tilde{f}_1(n,\delta)\Delta_{\gamma,I_t}^{-2}$ and $N_{t+1,I_t} = N_{t,I_t}+1$.

Case 2. Let $I_t \in \arg\min_{i\in[K]} \sqrt{N_{t,i}}(\gamma-\mu_{t,i})_+$. Since $I_t \notin \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$, under $\tilde{\mathcal{E}}_{n,\delta}$, for all $i \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$, we have

$$\sqrt{N_{t,I_t}}(\gamma-\mu_{I_t}) - \sqrt{2\tilde{f}_1(n,\delta)} \le \sqrt{N_{t,I_t}}(\gamma-\mu_{t,I_t}) = \sqrt{N_{t,I_t}}(\gamma-\mu_{t,I_t})_+$$
$$\sqrt{N_{t,i}}(\gamma-\mu_{t,i})_+ = \sqrt{N_{t,i}}(\gamma-\mu_{t,i}) \le \sqrt{N_{t,i}}(\gamma-\mu_i) + \sqrt{2\tilde{f}_1(n,\delta)} \le \sqrt{2\tilde{f}_1(n,\delta)} .$$

Combining both inequality by using that $\sqrt{N_{t,I_t}}(\gamma-\mu_{t,I_t})_+ \le \sqrt{N_{t,i}}(\gamma-\mu_{t,i})_+$ yields $\sqrt{N_{n,I_n}}(\gamma-\mu_{I_t}) \le 2\sqrt{2\tilde{f}_1(n,\delta)}$, hence $N_{t,I_t} \le 8\tilde{f}_1(n,\delta)\Delta_{\gamma,I_t}^{-2}$ and $N_{t+1,I_t} = N_{t,I_t}+1$.

Case 3. Then, $I_t \in \arg\min_{i\in[K]} \sqrt{N_{t,i}}(\gamma-\mu_{t,i})_+ = \{i \in [K] \mid \mu_{n,i} = \gamma\}$. Therefore, we have $\gamma = \mu_{n,I_n} \le \mu_{I_t} + \sqrt{2\tilde{f}_1(n,\delta)/N_{n,I_n}}$. Since $I_t \notin \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$, we obtain $N_{t,I_t} \le 2\tilde{f}_1(n,\delta)\Delta_{\gamma,I_t}^{-2}$ and $N_{t+1,I_t} = N_{t,I_t}+1$.

Summary. Combing the three above cases yields the result.

F.5 Proof of Lemma 6.10

Let $i \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \cap U_n(n,\delta)^{\complement}$. Then, $N_{n,i} > \left(\sqrt{2\tilde{f}_1(n,\delta)\Delta_{\gamma,i}^{-2}}+1\right)^2 > 2\tilde{f}_1(n,\delta)\Delta_{\gamma,i}^{-2}$. Under $\tilde{\mathcal{E}}_{n,\delta}$, we have $\max_{j\in[K]} \mu_{n,j} \ge \mu_{n,i} \ge \mu_i - \sqrt{2\tilde{f}_1(n,\delta)/N_{n,i}} > \gamma$, hence

$$\hat{\imath}_n = I_n \in \arg\max_{i\in[K]} \sqrt{N_{n,i}}(\mu_{n,i}-\gamma)_+ .$$

Suppose towards contradiction that $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)^\complement \cap \arg\max_{i\in[K]} \sqrt{N_{n,i}}(\mu_{n,i}-\gamma)_+ \neq \emptyset$. Let $i \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)^\complement \cap \arg\max_{i\in[K]} \sqrt{N_{n,i}}(\mu_{n,i}-\gamma)_+ \neq \emptyset$. It is direct to see that $\mu_{n,i} > \gamma$, otherwise there is a contradiction. Then, using that $i \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)^\complement$ (*i.e.* $\mu_i \leq \gamma$), we have for all $j \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \cap U_n(n,\delta)^\complement$

$$\begin{aligned}
&\sqrt{2\tilde{f}_1(n,\delta)} \geq \sqrt{N_{n,i}}(\mu_i-\gamma) + \sqrt{2\tilde{f}_1(n,\delta)} \geq \sqrt{N_{n,i}}(\mu_{n,i}-\gamma) = \sqrt{N_{n,i}}(\mu_{n,i}-\gamma)_+ \,, \\
&\sqrt{N_{n,j}}(\mu_{n,j}-\gamma)_+ = \sqrt{N_{n,j}}(\mu_{t,j}-\gamma) \geq \sqrt{N_{n,j}}(\mu_j-\gamma) - \sqrt{2\tilde{f}_1(n,\delta)/N_{n,j}} \\
&\qquad > \left(\sqrt{N_{n,j}}-1\right)(\mu_j-\gamma) > \sqrt{2\tilde{f}_1(n,\delta)} \,.
\end{aligned}$$

Since $i \neq j$ and $\sqrt{N_{n,i}}(\mu_{n,i}-\gamma)_+ \geq \sqrt{N_{n,j}}(\mu_{n,j}-\gamma)_+$, combining the above yields $\sqrt{2\tilde{f}_1(n,\delta)} > \sqrt{2\tilde{f}_1(n,\delta)}$ which is a contradiction. Therefore, $\hat{\imath}_n \in \arg\max_{i\in[K]} \sqrt{N_{n,i}}(\mu_{n,i}-\gamma)_+ \subseteq \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$.

F.6 Proof of Lemma 6.11

Let $(D_i(n,\delta))_{i\in[K]}$ as in Lemma 6.9. Combining Lemmas 6.9 and 5.8, we obtain

$$\sum_{t=K+1}^{n} \mathbb{1}\left(\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \subseteq U_t(n,\delta)\right) \leq \sum_{i\in[K]} D_i(n,\delta) \,.$$

For all $i \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$, let us define $t_i(n,\delta) = \max\{t \in (K,n] \cap \mathbb{N} \mid i \in U_t(n,\delta)\}$. By definition, we have $i \in U_t(n,\delta)$ for all $t \in (K, t_i(n,\delta)]$ and $i \notin U_t(n,\delta)$ for all $t \in (t_i(n,\delta), n]$. Therefore, for all $t \in (K, \min_{i\in\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)} t_i(n,\delta)]$, we have $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \subseteq U_t(n,\delta)$ and $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \cap U_t(n,\delta)^\complement \neq \emptyset$ for all $t > \max_{i\in[K]} t_i(n,\delta)$, hence

$$\min_{i\in\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)} (t_i(n,\delta)-K) = \sum_{t=K+1}^{n} \mathbb{1}\left(\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \subseteq U_t(n,\delta)\right) \leq \sum_{i\in[K]} D_i(n,\delta) \,.$$

Let $S_\mu(\delta)$ defined as in the statement of Lemma 6.11 and $n > S_\mu(\delta)$. Using that $(a+1)^2 \leq 2a^2+2$, we have $S_\mu(\delta) \geq \sup\left\{n \mid n \leq \sum_{i\in[K]} D_i(n,\delta) + K\right\}$. Then, we have $n-K > \sum_{i\in[K]} D_i(n,\delta) \geq \min_{i\in\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)}(t_i(n,\delta)-K)$, hence $n > \min_{i\in\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)} t_i(n,\delta)$ and $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \cap U_n(n,\delta)^\complement \neq \emptyset$. Using Lemma 6.10, we obtain that $\hat{\imath}_n \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$. This concludes the proof.

F.7 Proof of Lemma 6.12

Using Lemma C.9 for $s = 2$, we can show that

$$\mathbb{P}\left(\exists n \in \mathbb{N},\ \exists i \in [K],\ \sqrt{N_{n,i}}|\mu_{n,i}-\mu_i| > \sqrt{2c(n-1,\delta)}\right) \leq \delta \,,$$

where $c(n,\delta)$ as in (6.4) since $2\log(\pi^2/6)+5-4\log(4) \leq 1/2$.

For GAI, we have $\mathcal{E}_\mu^{\mathrm{err}}(n) = \{\hat{\imath}_n \in \{\emptyset\} \cup ([K] \setminus \mathcal{I}_\gamma^{\mathrm{thr}}(\mu))\}$ when $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \neq \emptyset$, otherwise $\mathcal{E}_\mu^{\mathrm{err}}(n) = \{\hat{\imath}_n \neq \emptyset\}$ when $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) = \emptyset$. Using the stopping time $\tau_{\gamma,\delta}^{\mathrm{thr}} := \min(\tau_{>,\delta}, \tau_{<,\delta})$ where $\tau_{>,\delta}, \tau_{<,\delta}$ are defined in (F.7), it is direct to show that

$$\{\tau_{\gamma,\delta}^{\mathrm{thr}} < +\infty\} \cap \mathcal{E}_\mu^{\mathrm{err}}(\tau_{\gamma,\delta}^{\mathrm{thr}}) \subseteq \bigcup_{n\in\mathbb{N}} \bigcup_{i\in[K]} \left\{ \sqrt{N_{n,i}}|\mu_{n,i} - \mu_i| > \sqrt{2c(n-1,\delta)} \right\} .$$

This concludes the proof.

F.8 Proof of Theorem 6.13

When combined with the GLR stopping (6.3) using threshold (6.4), APGAI becomes dependent of a confidence $\delta \in (0,1)$. Let $s > 1$ and

$$\mathcal{E}_n = \left\{ \forall i \in [K], \forall t \leq n,\ |\mu_{t,i} - \mu_i| < \sqrt{2f_1(n)/N_{t,i}} \right\} , \tag{F.1}$$

with $f_1(n) = (1+s)\log n$. Using concentration arguments, we have $\sum_n \mathbb{P}_\nu(\mathcal{E}_n^\complement) \leq K\zeta(s)$ where ζ is the Riemann ζ function. Using Lemma 2.26 in Chapter 2, the proof boils down to construct a time $T_\mu(\delta) > K$ such that $\mathcal{E}_n \subseteq \{\tau_{\gamma,\delta}^{\mathrm{thr}} \leq n\}$ for $n \geq T_\mu(\delta)$ since it yields that $\mathbb{E}_\nu[\tau_{\gamma,\delta}^{\mathrm{thr}}] \leq T_\mu(\delta) + K\zeta(s)$. Taking $s = 2$ yields the result.

As for the proof of Theorem 6.3, our main technical tool is Lemma 6.4, and we distinguish between instances μ such that $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) = \emptyset$ (Appendix F.8.1) and instances μ such that $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \neq \emptyset$ (Appendix F.8.2). It is direct to see that Lemmas 6.7 and 6.11 can be adapted to hold for $\mathcal{E}_n$ and $f_1(n) = (1+s)\log n$. Using an inversion result (Lemma E.5 in Appendix E.8), we state those results in a more explicit form, and omit the proof for the sake of space.

Lemma F.1. *Let $\mu \in \mathbb{R}^K$ such that $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) = \emptyset$ and $\mu_i \neq \gamma$ for all $i \in [K]$. Let $s > 1$. Let $T_\mu = h_1(18(1+s)H_1(\mu), K)$ where h_1 is defined in Lemma E.5. For all $n > T_\mu$, under the event $\mathcal{E}_n$ as in (F.1), we have $N_{n,i} > 2f_1(n)\Delta_{\gamma,i}^{-2}$ for all $i \in [K]$.*

Lemma F.2. *Let $\mu \in \mathbb{R}^K$ such that $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \neq \emptyset$ and $\mu_i \neq \gamma$ for all $i \in [K]$. Let $s > 1$. Let $S_\mu = h_1(4(1+s)H_1(\mu), K + 2|\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)|)$ where h_1 is defined in Lemma E.5. For all $n > S_\mu$, under the event $\mathcal{E}_n$ as in (F.1), we have $\hat{\imath}_n \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$ and there exists $i \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$ such that $N_{t,i} > (\Delta_{\gamma,i}^{-1}\sqrt{2f_1(n)} + 1)^2$.*

Theorem 6.13 is obtained by combining Lemmas F.4 and F.6.

F.8.1 Instances Where $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) = \emptyset$

When $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) = \emptyset$, we will have $\tau_{\gamma,\delta}^{\mathrm{thr}} = \tau_{<,\delta}$ almost surely and, for T large enough, $\hat{\imath}_n = \emptyset$ and $I_n \in \arg\min_{i\in[K]} \sqrt{N_{n,i}}(\gamma - \mu_{n,i})_+$. Lemma F.3 formalizes this intuition.

Lemma F.3. *Let $s > 1$. Let $T_\mu = h_1(18(1+s)H_1(\mu), K)$ where h_1 is defined in Lemma E.5. For all $n > T_\mu$, $I_n \in \arg\min_{i\in[K]} \sqrt{N_{n,i}}(\gamma - \mu_{n,i})_+$ and $\hat{\imath}_n = \emptyset$. Moreover, we have $\tau_{\gamma,\delta}^{\mathrm{thr}} = \tau_{<,\delta}$ almost surely.*

Proof. Let T_μ as in Lemma 6.7. Let $n > T_\mu$. Using Lemma 6.7, we obtain that $N_{n,i} > 2f_1(n)\Delta_{\gamma,i}^{-2}$ for all $i \in [K]$. Then, under $\mathcal{E}_n$ as in (F.1), $\mu_{n,i} \le \mu_i + \sqrt{2f_1(n)/N_{n,i}} < \gamma$ for all $i \in [K]$, hence $\max_{i\in[K]} \mu_{n,i} < \gamma$. Using the definition of the sampling rule when $\max_{i\in[K]} \mu_{n,i} < \gamma$, for all $n > T_\mu$, we have $I_n \in \arg\min_{i\in[K]} \sqrt{N_{n,i}}(\gamma - \mu_{n,i})_+$ and $\hat{\imath}_n = \emptyset$. A direct consequence is that $\tau_{>,\delta} = +\infty$, hence $\tau_{<,\delta} = \tau_{\gamma,\delta}^{\mathrm{thr}}$ almost surely. ■

When coupled with the GLR stopping (6.3) using threshold (6.4), Lemma F.4 gives an upper bound on the expected sample complexity of APGAI when $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) = \emptyset$.

Lemma F.4. *Let $\delta \in (0,1)$. Combined with GLR stopping (6.3) using threshold (6.4), the APGAI algorithm is δ-correct and it satisfies that, for all $\nu \in \mathcal{D}^K$ with mean μ such that $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) = \emptyset$ and $\Delta_{\gamma,\min} > 0$, $\mathbb{E}_\nu[\tau_{\gamma,\delta}^{\mathrm{thr}}] \le C_\mu(\delta) + K\pi^2/6 + 1$ with $H_1(\mu)$ as in (6.1) and $T_\mu = h_1(54H_1(\mu), K)$ with h_1 is defined in Lemma E.5 and*

$$C_\mu(\delta) = \sup\left\{ n \mid \frac{n - T_\mu}{2H_1(\mu)} \le \left(\sqrt{c(n-1,\delta)} + \sqrt{3\log n}\right)^2 + \left(\gamma - \min_{i\in[K]} \mu_i\right)^2 - 3\log T_\mu \right\}$$
$$= \sup\{ n \mid n \le 2H_1(\mu)(\sqrt{c(n-1,\delta)} + \sqrt{3\log n})^2 + D_1(\mu)\} ,$$

where $D_1(\mu) = T_\mu + 2H_1(\mu)\left(\gamma - \min_{i\in[K]} \mu_i\right)^2 - 6H_1(\mu)\log T_\mu$. In particular, it satisfies $\limsup_{\delta\to 0} \mathbb{E}_\nu[\tau_{\gamma,\delta}^{\mathrm{thr}}]/\log(1/\delta) \le 2H_1(\mu)$.

Proof. Let T_μ as in Lemma F.3. Let $n > T_\mu$ such that $\mathcal{E}_n \cap \{\tau_{\gamma,\delta}^{\mathrm{thr}} > n\}$ holds true. Let $w \in \triangle_K$ such that $w_i = \Delta_{\gamma,i}^{-2} H_1(\mu)^{-1}$ for all $i \in [K]$. Using the pigeonhole principle, at time n there exists $i_1 \in [K]$ such that $N_{n,i_1} - N_{T_\mu,i_1} \ge (n - T_\mu)w_{i_1}$. Let $n \ge T_\mu + (\min_{i\in[K]} w_i)^{-1}$, hence we have $N_{n,i_1} - N_{T_\mu,i_1} \ge w_{i_1}/\min_{i\in[K]} w_i \ge 1$. Therefore, arm i_1 has been sampled at least once in (T_μ, n) . Let $t_{i_1} \in (T_\mu, n)$ be the last time at which arm i_1 was selected to be pulled next, *i.e.* $I_{t_{i_1}} = i_1$ and $N_{n,i_1} = N_{t_{i_1}+1,i_1} = N_{t_{i_1},i_1} + 1$. Since $t_{i_1} > T_\mu$, Lemma F.3 yields that

$i_1 = I_{t_{i_1}} \in \arg\min_{i\in[K]} \sqrt{N_{t_{i_1},i}}(\gamma - \mu_{t_{i_1},i})_+$. Moreover, we have

$$N_{t_{i_1},i_1} = N_{n,i_1} - 1 \geq (n - T_\mu)w_{i_1} + N_{T_\mu,i_1} - 1 \geq nw_{i_1} + \frac{2f_1(T_\mu) - T_\mu H_1(\mu)^{-1}}{\Delta^2_{\gamma,i_1}} - 2\,,$$

where we used that $N_{T_\mu,i_1} \geq N_{T_\mu+1,i_1} - 1 > 2f_1(T_\mu+1)\Delta_{i_1}^{-2}$ and f_1 is increasing. Under $\mathcal{E}_n$ as in (F.1), using that $i_1 = I_{t_{i_1}} \in \arg\min_{i\in[K]} \sqrt{N_{t_{i_1},i}}(\gamma - \mu_{t_{i_1},i})_+$, we obtain

$$\begin{aligned}
\sqrt{N_{t_{i_1},i_1}}(\gamma - \mu_{t_{i_1},i_1})_+ &= \sqrt{N_{t_{i_1},i_1}}(\gamma - \mu_{t_{i_1},i_1}) \geq \sqrt{N_{t_{i_1},i_1}}(\gamma - \mu_{i_1}) - \sqrt{2f_1(n)} \\
&\geq \sqrt{(nw_{i_1}(\gamma - \mu_{i_1})^2 + 2f_1(T_\mu) - T_\mu H_1(\mu)^{-1} - 2(\gamma - \mu_{i_1})^2)} - \sqrt{2f_1(n)} \\
&= \sqrt{(n - T_\mu)H_1(\mu)^{-1} + 2f_1(T_\mu) - 2(\gamma - \mu_{i_1})^2} - \sqrt{2f_1(n)}\,.
\end{aligned}$$

Since $i_1 = I_{t_{i_1}} \in \arg\min_{i\in[K]} \sqrt{N_{t_{i_1},i}}(\gamma - \mu_{t_{i_1},i})_+$, using that the condition of the stopping rule is not met at time t_{i_1} yields

$$\begin{aligned}
&\sqrt{2c(n-1,\delta)} \geq \sqrt{2c(t_{i_1}-1,\delta)} \geq \min_{j\in[K]} \sqrt{N_{t_{i_1},j}}(\gamma - \mu_{t_{i_1},j})_+ = \sqrt{N_{t_{i_1},i_1}}(\gamma - \mu_{t_{i_1},i_1})_+ \quad \text{hence} \\
&\sqrt{2c(n-1,\delta)} \geq \sqrt{(n - T_\mu)H_1(\mu)^{-1} + 2f_1(T_\mu) - 2(\gamma - \mu_{i_1})^2} - \sqrt{2f_1(n)}\,.
\end{aligned}$$

Using $\mu_{i_1} \geq \min_{i\in[K]} \mu_i$, the above inequality can be rewritten as

$$n - T_\mu \leq 2\left(\sqrt{c(n-1,\delta)} + \sqrt{f_1(n)}\right)^2 H_1(\mu) + 2H_1(\mu)\left((\gamma - \min_{i\in[K]} \mu_i)^2 - f_1(T_\mu)\right)\,.$$

Let us define

$$C_\mu(\delta) = \sup\left\{n \mid \frac{n - T_\mu}{2H_1(\mu)} \leq \left(\sqrt{c(n-1,\delta)} + \sqrt{f_1(n)}\right)^2 + (\gamma - \min_{i\in[K]} \mu_i)^2 - f_1(T_\mu)\right\}\,.$$

It is direct to notice that $T_\mu + (\min_{i\in[K]} w_i)^{-1} = T_\mu + (\gamma - \min_{i\in[K]} \mu_i)^2 H_1(\mu) \leq C_\mu(\delta)$. Therefore, we have shown that for $n \geq C_\mu(\delta) + 1$, we have $\mathcal{E}_n \subset \{\tau_{<,\delta} \leq n\} = \{\tau^{\mathrm{thr}}_{\gamma,\delta} \leq n\}$ (by using Lemma F.3). Using Lemma 2.26, we obtain $\mathbb{E}_\nu[\tau^{\mathrm{thr}}_{\gamma,\delta}] \leq C_\mu(\delta) + K\zeta(s) + 1$. Taking $s = 2$, using that $\zeta(2) = \pi^2/6$ and $f_1(n) = 3\log n$ yields the second part of the result. Direct manipulations show that

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau^{\mathrm{thr}}_{\gamma,\delta}]}{\log(1/\delta)} \leq \limsup_{\delta\to 0} \frac{C_\mu(\delta)}{\log(1/\delta)} \leq 2H_1(\mu)\,.$$

According to Lemma 6.1, we have proven asymptotic optimality and Lemma 2.3 gives the δ-correctness of the APGAI algorithm.

■

F.8.2 Instances Where $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \neq \emptyset$

When $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \neq \emptyset$, we will have $\tau_{\gamma,\delta}^{\mathrm{thr}} = \tau_{>,\delta}$ almost surely and, for T large enough, $\hat{\imath}_n = I_n \in \arg\max_{i \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)} \sqrt{N_{n,i}}(\mu_{n,i} - \gamma)_+$. Lemma F.5 formalizes this intuition.

Lemma F.5. *Let $s > 1$. Let $S_\mu = h_1(4(1+s)H_1(\mu), K + 2|\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)|)$ where h_1 is defined in Lemma E.5. For all $n > S_\mu$, $\hat{\imath}_n = I_n \in \arg\max_{i \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)} \sqrt{N_{n,i}}(\mu_{n,i} - \gamma)_+$. Moreover, we have $\tau_{\gamma,\delta}^{\mathrm{thr}} = \tau_{>,\delta}$ almost surely.*

Proof. Let S_μ as in Lemma F.2 Let $n > S_\mu$. Using Lemma 6.11, there exists $i \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$ such that $N_{n,i} > 2f_1(n)\Delta_{\gamma,i}^{-2}$. Then, we have $\mu_{n,i} \geq \mu_i - \sqrt{2f_1(n)/N_{n,i}} > \gamma$, hence $\max_{i \in [K]} \mu_{n,i} > \gamma$. Using Lemma 6.10 and the definition of the recommendation rule when $\max_{i \in [K]} \mu_{n,i} > \gamma$, we obtain that $\hat{\imath}_n = I_n \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$. Using the definition of the sampling rule when $\max_{i \in [K]} \mu_{n,i} > \gamma$, for all $n > S_\mu$, we have $\hat{\imath}_n = I_n \in \arg\max_{i \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)} \sqrt{N_{n,i}}(\mu_{n,i} - \gamma)_+$. A direct consequence is that $\tau_{<,\delta} = +\infty$, hence $\tau_{>,\delta} = \tau_{\gamma,\delta}^{\mathrm{thr}}$ almost surely. ■

When coupled with the GLR stopping (6.3) using threshold (6.4), Lemma F.6 gives an upper bound on the expected sample complexity of APGAI when $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \neq \emptyset$.

Lemma F.6. *Let $\delta \in (0,1)$. Combined with GLR stopping (6.3) using threshold (6.4), the APGAI algorithm is δ-correct and it satisfies that, for all $\nu \in \mathcal{D}^K$ with mean μ such that $\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \neq \emptyset$ and $\Delta_{\gamma,\min} > 0$, $\mathbb{E}_\nu[\tau_{\gamma,\delta}^{\mathrm{thr}}] \leq C_\mu(\delta) + K\pi^2/6 + 1$, where $H_\gamma(\mu)$ as in (6.1) and $S_\mu = h_1(12H_1(\mu), K + 2|\mathcal{I}_\gamma^{\mathrm{thr}}(\mu)|)$ with h_1 is defined in Lemma E.5 and*

$$C_\mu(\delta) = \sup\left\{ n \mid \frac{n - S_\mu - 1}{2H_\gamma(\mu)} \leq \left(\sqrt{c(n-1,\delta)} + \sqrt{3\log n}\right)^2 - \frac{3\log S_\mu}{H_\gamma(\mu)\max_{i \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)} \Delta_{\gamma,i}^2} \right\}$$
$$= \sup\{ n \mid n \leq 2H_\gamma(\mu)(\sqrt{c(n-1,\delta)} + \sqrt{3\log n})^2 + D_\gamma(\mu) \} ,$$

where $D_\gamma(\mu) = S_\mu + 1 - \frac{6\log S_\mu}{\max_{i \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)} \Delta_{\gamma,i}^2}$. In particular, it satisfies $\limsup_{\delta \to 0} \mathbb{E}_\nu[\tau_{\gamma,\delta}^{\mathrm{thr}}]/\log(1/\delta) \leq 2H_\gamma(\mu)$.

Proof. Let S_μ as in Lemma F.5. Let $n > S_\mu$ such that $\mathcal{E}_n \cap \{\tau_{\gamma,\delta}^{\mathrm{thr}} > n\}$ holds true. Using Lemma F.5, we know that $I_t \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)$ for all $t \in (S_\mu, n]$. Direct summation yields that

$$n - S_\mu = \sum_{i \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)} (N_{n,i} - N_{S_\mu,i}) + \sum_{t \in (S_\mu, n]} \mathbb{1}\left(I_t \notin \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)\right) = \sum_{i \in \mathcal{I}_\gamma^{\mathrm{thr}}(\mu)} (N_{n,i} - N_{S_\mu,i}) .$$

At time $S_\mu + 1$, let $i_1 \in \mathcal{I}^{\mathrm{thr}}_\gamma(\mu)$ as in Lemma F.5, *i.e.* such that $N_{S_\mu+1,i_1} > \frac{2f_1(S_\mu+1)}{(\mu_{i_1}-\gamma)^2}$. Using that f_1 is increasing, we obtain

$$\sum_{j\in\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)} N_{S_\mu,j} \geq N_{S_\mu+1,i_1} - 1 > \frac{2f_1(S_\mu+1)}{(\mu_{i_1}-\gamma)^2} - 1 \geq \frac{2f_1(S_\mu)}{\max_{i\in\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)}(\mu_i-\gamma)^2} - 1 \,.$$

Therefore, we have shown that $\sum_{i\in\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)} N_{n,i} \geq n - g(S_\mu)$ with $g(S_\mu) = S_\mu - \frac{2f_1(S_\mu)}{\max_{i\in\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)}\Delta^2_{\gamma,i}} + 1$. Let $A_\gamma = |\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)|$ and $w \in \triangle_{A_\gamma}$ such that $w_i = (\mu_i-\gamma)^{-2}H_\gamma(\mu)^{-1}$ with $H_\gamma(\mu)$ as in (6.1). Using the pigeonhole principle, there exists $i_0 \in \mathcal{I}^{\mathrm{thr}}_\gamma(\mu)$ such that $N_{n,i_0} \geq w_{i_0}(n-g(S_\mu)) = \Delta_{i_0}^{-2}H_\gamma(\mu)^{-1}(n-g(S_\mu))$. Let us define $E_\mu(\delta) = \sup\{n \mid n \leq g(S_\mu) + 2H_\gamma(\mu)f_1(n)\}$ and let $n > E_\mu(\delta)$. Then, we have $N_{n,i_0} \geq \Delta_{i_0}^{-2}H_\gamma(\mu)^{-1}(n-g(S_\mu)) > 2f_1(n)\Delta_{i_0}^{-2}$, hence $\mu_{n,i_0} > \gamma$. Using that the condition of the stopping rule is not met at time T , we obtain

$$\sqrt{2c(n-1,\delta)} \geq \max_{i\in[K]}\sqrt{N_{n,i}}(\mu_{n,i}-\gamma)_+ \geq \sqrt{N_{n,i_0}}(\mu_{n,i_0}-\gamma)_+ = \sqrt{N_{n,i_0}}(\mu_{n,i_0}-\gamma) \,.$$

Then, we obtain

$$\begin{aligned}\sqrt{2c(n-1,\delta)} \geq \sqrt{N_{n,i_0}}(\mu_{i_0}-\gamma) - \sqrt{2f_1(n)} &\geq \sqrt{n-g(S_\mu)}\sqrt{w_{i_0}(\mu_{i_0}-\gamma)^2} - \sqrt{2f_1(n)} \\ &= \sqrt{n-g(S_\mu)}H_\gamma(\mu)^{-1/2} - \sqrt{2f_1(n)} \,.\end{aligned}$$

The above can be rewritten as $n \leq 2\left(\sqrt{c(n-1,\delta)} + \sqrt{f_1(n)}\right)^2 H_\gamma(\mu) + g(S_\mu)$. Using that $g(S_\mu) = S_\mu - \frac{2f_1(S_\mu)}{\max_{i\in\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)}\Delta^2_{\gamma,i}} + 1$, let us define

$$D_\mu(\delta) = \sup\left\{n \mid \frac{n-S_\mu-1}{2H_\gamma(\mu)} \leq \left(\sqrt{c(n-1,\delta)} + \sqrt{f_1(n)}\right)^2 - \frac{f_1(S_\mu)}{H_\gamma(\mu)\max_{i\in\mathcal{I}^{\mathrm{thr}}_\gamma(\mu)}\Delta^2_{\gamma,i}}\right\} .$$

It is direct to see that $D_\mu(\delta) \geq E_\mu(\delta) \geq S_\mu$. Therefore, we have shown that for $n \geq D_\mu(\delta)+1$, we have $\mathcal{E}_n \subset \{\tau_{>,\delta} \leq n\} = \{\tau^{\mathrm{thr}}_{\gamma,\delta} \leq n\}$ (by using Lemma F.5). Using Lemma 2.26, we obtain $\mathbb{E}_\nu[\tau^{\mathrm{thr}}_{\gamma,\delta}] \leq D_\mu(\delta) + K\zeta(s) + 1$. Taking $s = 2$, using that $\zeta(2) = \pi^2/6$ and $f_1(n) = 3\log n$ yields the second part of the result. Direct manipulations show that

$$\limsup_{\delta\to 0}\frac{\mathbb{E}_\nu[\tau^{\mathrm{thr}}_{\gamma,\delta}]}{\log(1/\delta)} \leq \limsup_{\delta\to 0}\frac{D_\mu(\delta)}{\log(1/\delta)} \leq 2H_\gamma(\mu) \,.$$

■

Appendix G

Complements on Chapter 7

G.1 Proof of Lemma 7.3

As explained in Section 1.4.1, change of measures with a low-level form (involving probabilities and not expectation) is key to derive a lower bound on the sample complexity of ε-BAI [Garivier and Kaufmann, 2021, Degenne and Koolen, 2019]. We use Lemma 19 in Degenne and Koolen [2019], which is proven using (1.4).

Lemma G.1 (Lemma 19 in Degenne and Koolen [2019]). *Let $z \in \mathcal{Z}$. Let w and $\lambda_1, \cdots, \lambda_K$ be a minimax witness from Lemma G.2, and let us introduce the abbreviation $\alpha_a = \|\theta - \sum_{k\in[K]} w_k\lambda_k\|^2_{aa^\intercal}$ for all $a \in \mathcal{A}$. Fix a sample size n, and consider any event $A \in \mathcal{F}_n$. Then, for any $\beta > 0$, $\max_{k\in[K]} \mathbb{P}_{\lambda_k}(A) \geq e^{-nT_\varepsilon(\nu,z)^{-1}-\beta}\left(\mathbb{P}_\nu(A) - \exp\left(\frac{-\beta^2}{2n\max_{a\in\mathcal{A}}\alpha_a}\right)\right)$, where $T_\varepsilon(\nu,z)^{-1} = \sup_{w\in\Sigma_K}\inf_{\lambda\in\neg_\varepsilon z}\|\theta-\lambda\|^2_{V_w}$.*

We rewrite Lemma 2 in Degenne and Koolen [2019] in the setting of (ε,δ)-PAC BAI for transductive linear bandits with Gaussian distribution.

Lemma G.2 (Lemma 2 in Degenne and Koolen [2019]). *For any answer $z \in \mathcal{Z}$, the divergence from ν to $\neg_\varepsilon z$ equals $T_\varepsilon(\nu,z)^{-1} = \sup_{w\in\Sigma_K}\inf_{\lambda\in\neg_\varepsilon z}\|\theta-\lambda\|^2_{V_w} = \inf_{\mathbb{P}}\max_{a\in\mathcal{A}}\mathbb{E}_{\lambda\sim\mathbb{P}}\left[\|\theta-\lambda\|^2_{aa^\intercal}\right]$ where the infimum ranges over probability distributions on $\neg_\varepsilon z$ supported on (at most) K points.*

We will bound the expectation of the stopping time τ_δ through Markov's inequality. For all $T > 0$, $\mathbb{E}_\nu[\tau_\delta] \geq T(1 - \mathbb{P}_\nu(\tau_\delta \leq T))$. The event $\{\tau_\delta \leq T\}$ can be partitioned depending on the answer whether the answer is ε-optimal or not, and then whether it's $z^\star(\theta)$ or not. By

hypothesis,

$$\mathbb{P}_\nu\left(\tau_\delta \le T, \hat{z} \notin \mathcal{Z}_\varepsilon(\theta)\right) \le \mathbb{P}_\nu\left(\tau_\delta < +\infty, \hat{z} \notin \mathcal{Z}_\varepsilon(\theta)\right) \le \delta$$
$$0 \le \lim_{\delta\to 0} \mathbb{P}_\nu\left(\tau_\delta \le T, \hat{z} \in \mathcal{Z}_\varepsilon(\theta) \setminus z^\star(\theta)\right) \le \lim_{\delta\to 0} \mathbb{P}_\nu\left(\tau_\delta < +\infty, \hat{z} \in \mathcal{Z}_\varepsilon(\theta) \setminus z^\star(\theta)\right) = 0\,.$$

Then, it is direct to show that $\lim_{\delta\to 0} \mathbb{P}_\nu\left(\tau_\delta \le T\right) \le \sum_{z\in z^\star(\theta)} \lim_{\delta\to 0} \mathbb{P}_\nu\left(\tau_\delta \le T, \hat{z} = z\right)$. Let $z \in z^\star(\theta)$, w and $\lambda_1, \cdots, \lambda_K$ be a minimax witness from Lemma G.2. Then by Lemma G.1, for any $\beta > 0$

$$\begin{aligned}\mathbb{P}_\nu\left(\tau_\delta \le T, \hat{z} = z\right) &\le \exp\left(\frac{T}{T_\varepsilon(\nu, z)} + \beta\right) \max_{k\in[K]} \mathbb{P}_{\lambda_k}\left(\tau_\delta \le T, \hat{z} = z\right) + \exp\left(\frac{-\beta^2}{2T \max_{a\in\mathcal{A}} \alpha_a}\right) \\ &\le \delta \exp\left(\frac{T}{T_\varepsilon(\nu, z)} + \beta\right) + \exp\left(\frac{-\beta^2}{2T \max_{a\in\mathcal{A}} \alpha_a}\right)\end{aligned}$$

where the second inequality uses that $\lambda_k \in \neg_\varepsilon z$ for all $k \in [K]$, hence $z \in z^\star(\theta) \subseteq \mathcal{Z} \setminus \mathcal{Z}_\varepsilon(\lambda_k)$ and that the strategy satisfies $\mathbb{P}_\lambda\left(\tau_\delta < +\infty, \hat{z} \notin \mathcal{Z}_\varepsilon(\lambda)\right) \le \delta$ for all $\lambda \in \mathcal{M}$. Let $\alpha = \max_{a\in\mathcal{A}} \alpha_a$. For $\eta \in (0,1)$, $T = (1-\eta) \min_{z\in z^\star(\theta)} T_\varepsilon(\nu, z) \log(1/\delta)$, $\beta = \frac{\eta}{2\sqrt{1-\eta}} \sqrt{\frac{T}{\min_{z\in z^\star(\theta)} T_\varepsilon(\nu,z)} \log(1/\delta)}$, and all $z \in z^\star(\theta)$,

$$\begin{aligned}&\mathbb{P}_\nu\left(\tau_\delta \le T, \hat{z} = z\right) \\ &\le \delta \exp\left(\frac{T}{T_\varepsilon(\nu, z)} + \frac{\eta}{2\sqrt{1-\eta}} \sqrt{\frac{T \log(1/\delta)}{\min_{z\in z^\star(\theta)} T_\varepsilon(\nu, z)}}\right) + \exp\left(\frac{-\eta^2 \log(1/\delta)}{8(1-\eta) \min_{z\in z^\star(\theta)} T_\varepsilon(\nu, z)\alpha}\right) \\ &\le \delta \exp\left((1-\eta/2) \log \frac{1}{\delta}\right) + \exp\left(\frac{-\eta^2 \log(1/\delta)}{8(1-\eta) \min_{z\in z^\star(\theta)} T_\varepsilon(\nu, z)\alpha}\right) \\ &= \delta^{\eta/2} + \delta^{\eta^2/\left(8(1-\eta) \min_{z\in z^\star(\theta)} T_\varepsilon(\nu,z)\alpha\right)} \to_{\delta\to 0} 0\,,\end{aligned}$$

where we used that $\min_{z\in z^\star(\theta)} T_\varepsilon(\nu, z) \le T_\varepsilon(\nu, z)$. Since we have just shown $\lim_{\delta\to 0} \mathbb{P}_\nu\left(\tau_\delta \le T\right) = 0$ for $T = (1-\eta) \min_{z\in z^\star(\theta)} T_\varepsilon(\nu, z) \log(1/\delta)$, we obtain

$$\begin{aligned}\lim_{\delta\to 0} \frac{\mathbb{E}_\nu\left[\tau_\delta\right]}{\log(1/\delta)} \ge \lim_{\delta\to 0} \frac{T}{\log(1/\delta)} \left(1 - \mathbb{P}_\nu\left(\tau_\delta \le T\right)\right) &\ge (1-\eta) \min_{z\in z^\star(\theta)} T_\varepsilon(\nu, z) \left(1 - \lim_{\delta\to 0} \mathbb{P}_\nu\left(\tau_\delta \le T\right)\right) \\ &= (1-\eta) \min_{z\in z^\star(\theta)} T_\varepsilon(\nu, z)\,.\end{aligned}$$

Taking $\eta \to 0$ concludes the proof.

Appendix H

Complements on Chapter 8

H.1 Proof of Lemma 8.4

Since Slater's condition holds, the KKT conditions are necessary and sufficient for global optimality. Let $\lambda \geq 0$, $\alpha \in \mathbb{R}_+^K$ and $\gamma \in \mathbb{R}_+^{Z-1}$ be the dual variables for the Lagrangian

$$\mathcal{L}(\phi, w; \lambda, \alpha, \gamma) = \phi + \lambda \left(\sum_{a \in \mathcal{A}} w_a - 1 \right) - \sum_{a \in \mathcal{A}} \alpha_a w_a + \sum_{x \in \mathcal{Z} \setminus \{z\}} \gamma_x (\phi - C_\varepsilon(z, x; \nu, w)) \, .$$

Using the complementary slackness condition, we have $\gamma_x(\phi - C_\varepsilon(z, x; \nu, w)) = 0$ for all $x \neq z$. Combining it with the stationarity condition for arm a , we obtain

$$\lambda = \alpha_a + \sum_{x \in \mathcal{Z} \setminus \{z\}} \gamma_x \frac{\partial C_\varepsilon(z, x; \nu, w)}{\partial w_a} \quad \text{hence} \quad \frac{\lambda}{\phi} = \frac{\alpha_a}{\phi} + \sum_{x \in \mathcal{Z} \setminus \{z\}} \frac{\gamma_x}{C_\varepsilon(z, x; \nu, w)} \frac{\partial C_\varepsilon(z, x; \nu, w)}{\partial w_a} \, .$$

Multiplying by w_a and using that $\alpha_a w_a = 0$ yields that

$$w_a \frac{\lambda}{\phi} = \sum_{x \in \mathcal{Z} \setminus \{z\}} \gamma_x \frac{w_a}{C_\varepsilon(z, x; \nu, w)} \frac{\partial C_\varepsilon(z, x; \nu, w)}{\partial w_a} \, .$$

Using that $\sum_{a \in \mathcal{A}} w_a = 1$ (since $\lambda > 0$) and $\sum_{a \in \mathcal{A}} \frac{w_a}{C_\varepsilon(z, x; \nu, w)} \frac{\partial C_\varepsilon(z, x; \nu, w)}{\partial w_a}$, summing those equations yields that $\lambda = \phi \sum_{x \in \mathcal{Z} \setminus \{z\}} \gamma_x$. By scaling, we can consider $\tilde{\gamma}_x = \gamma_x (\sum_{x \in \mathcal{Z} \setminus \{z\}} \gamma_x)^{-1}$ for all $x \in \mathcal{Z} \setminus \{z\}$, *i.e.* $\tilde{\gamma} \in \Sigma_{Z-1}$.

This completes the proof for the necessity of the conditions in the theorem for optimality. Those conditions are also sufficient since it is direct to construct dual variables such that the KKT conditions hold.

H.2 Proof of Lemma 8.5

As in Lemma 2.6, the goal is to explicit the dual variable $\gamma \in \Sigma_{Z-1}$ from Lemma 8.4 by using the KKT conditions. The crucial difference lies in the fact that the optimal allocation has no reason to be densely supported. Since $\min_{a\in\mathcal{A}} w_a = 0$, some Lagrangian multipliers α_a might not be null while other Lagragian multipliers γ_x might be null.

Let $\mathcal{A}_1 = \{a \in \mathcal{A} \mid \alpha_a = 0\}$ and $\mathcal{Z}_1 = \{x \in \mathcal{Z} \setminus \{z\} \mid \gamma_x > 0\}$, hence we have $\phi = C_\varepsilon(z, x; \nu, w)$ for all $x \in \mathcal{Z}_1$ and, for all $a \notin \mathcal{A}_1$, $w_a = 0$. Then, we obtain $\lambda = \sum_{x\in\mathcal{Z}_1} \gamma_x \frac{\partial C_\varepsilon(z,x;\nu,w)}{\partial w_a}$ for all $a \in \mathcal{A}_1$. Multiplying by w_a and summing over $\mathcal{A}_1$ yields that

$$\lambda = \sum_{x\in\mathcal{Z}_1} \gamma_x \sum_{a\in\mathcal{A}_1} w_a \frac{\partial C_\varepsilon(z, x; \nu, w)}{\partial w_a} = \sum_{x\in\mathcal{Z}_1} \gamma_x C_\varepsilon(z, x; \nu, w) = \sum_{x\in\mathcal{Z}_1} \gamma_x \phi = \phi \,.$$

This completes the proof for the necessity of the conditions. Those conditions are also sufficient since it is direct to construct dual variables such that the KKT conditions hold.

List of Notation

Acronyms and Abbreviations

a.s.	almost surely
e.g.	exempli gratia, means "for example"
i.e.	id est, means "that is"
i.i.d.	independent and identically distributed
l.h.s.	left hand side
r.h.s.	right hand side
s.t.	such that
w.r.t.	with respect to
BAI	Best Arm Identification
GAI	Good Arm Identification
GLR	Generalized log-likelihood ratio
KL	Kullback-Leibler
MAB	Multi-Armed Bandits

General Notation

$[K]$	Set of integers $\{1, \cdots, K\}$
$X \sim \nu$	The random variable X has distribution ν
$\mathbb{P}_\nu(\mathcal{E})$	Probability of a random event $\mathcal{E}$ under distribution ν
$\mathbb{E}_\nu[X]$	Expectation of a random variable X under distribution ν
$\mathbb{1}(\mathcal{E})$	Indicator function of an event $\mathcal{E}$
$X^{\complement}$	Complement of a set X
$\mathring{X}$	Interior of a set X
o , $\mathcal{O}$, Ω and Θ ($\tilde{o}$, $\tilde{\mathcal{O}}$, $\tilde{\Omega}$ and $\tilde{\Theta}$)	Landau's notation (up to polylogarithmic terms)
ζ	Riemann ζ function, $\zeta(s) := \sum_{n=1}^{+\infty} n^{-s}$ for all $s > 1$
W_0 , W_{-1}	Positive and negative branches of the Lambert W function, $W(x)e^{W(x)} := x$
$\overline{W}_0$, $\overline{W}_{-1}$	Transform on the Lambert W function, $\overline{W}_i(x) := -W_i(-e^{-x})$ for all $x \geq 1$

$\mathcal{C}_G$	Function defined in (B.1)
$\langle x, y \rangle$	Cartesian product between vectors, $\langle x, y \rangle = \sum_{i \in [d]} x_i y_i$
$\overline{X}$	Closure of set X
∂X	Boundary of set X
$\|x\|_p$	ℓ_p-norm, *e.g.* $\|x\|_2 = \sqrt{\langle x, x \rangle}$, $\|x\|_1 = \sum_{i \in [d]} \|x_i\|$, $\|x\|_\infty = \max_{i \in [d]} \|x_i\|$
$(x)_+$	Positive part, $\max\{x, 0\}$
$\Pi_{i \in [K]} X_i$	Cartesian product between sets or distributions $(X_i)_{i \in [K]}$
V_w	Design matrix, $\sum_{a \in \mathcal{A}} w_a a a^\intercal$
1_A	Indicator vector for a set $A \subseteq [d]$, $1_A = (\mathbb{1}(i \in A))_{i \in [d]}$
$\{e_i\}_{i \in [d]}$	Canonical basis of $\mathbb{R}^d$, $e_i = (\mathbb{1}(j = i))_{j \in [d]}$
$\mathrm{diag}(x) \in \mathbb{R}^{d \times d}$	Diagonal matrix for a vector $x \in \mathbb{R}^d$
$\mathrm{Span}(\mathcal{A})$	Span of a set of vectors $\mathcal{A}$
$I_d \in \mathbb{R}^{d \times d}$	Identity matrix
$V^\dagger$	Moore-Penrose pseudo-inverse of matrix V

Distributions

Σ_K	$(K-1)$-dimensional probability simplex, $\Sigma_K := \left\{ w \in \mathbb{R}_+^K \mid w \geq 0,\ \sum_{i \in [K]} w_i = 1 \right\}$
$\mathcal{D}$	Set of distributions
$\mathcal{D}_{\mathcal{N}}$	Set of Gaussian distributions with unknown variance
$\mathcal{D}_{\mathcal{N}_\sigma}$	Set of Gaussian distributions with variance σ^2
$\mathcal{D}_\sigma$	Set of σ-sub-Gaussian distributions
$\mathcal{D}_{\mathcal{B}}$	Set of Bernoulli distributions
$\mathcal{P}(\mathbb{R})$	Probability distributions over $\mathbb{R}$
B	Upper bound for bounded distributions
$\mathcal{D}_{[0,B]}$	Set of bounded distributions on $[0, B]$
$\mathcal{D}_{\exp}$	One-parameter exponential family
δ_x	Dirac distribution at x
KL	Kullback-Leibler divergence
d_{KL}	Mean-parametrized Kullback-Leibler divergence
kl	Mean-parametrized Kullback-Leibler divergence of the Bernoulli family
m	Mean operator, $m(\nu) = \mathbb{E}_{X \sim \nu}[X]$
$\mathcal{K}_{\inf}^+$	Infimum of KL divergence *r.h.s.* , $\mathcal{K}_{\inf}^+(\nu, u) := \inf\{\mathrm{KL}(\nu, \kappa) \mid \kappa \in \mathcal{D},\ m(\kappa) > u\}$
$\mathcal{K}_{\inf}^-$	Infimum of KL divergence *l.h.s.* , $\mathcal{K}_{\inf}^-(\nu, u) := \inf\{\mathrm{KL}(\nu, \kappa) \mid \kappa \in \mathcal{D},\ m(\kappa) < u\}$
$\mathbb{P} \ll \mathbb{Q}$	$\mathbb{Q}$ absolutely dominates $\mathbb{P}$

$\mathbb{P}^X$	Push-forward measure of a random variable X under probability $\mathbb{P}$

Multi-Armed Bandits

$K \in \mathbb{N}$	Number of arms
$i \in [K]$	Arm
$\nu_i \in \mathcal{D}$	Distribution of arm i
$\nu \in \mathcal{D}^K$	Vector of distributions, $\nu := (\nu_i)_{i\in[K]}$
$\mathcal{I} \subseteq \mathbb{R}$	Set of possible means for the arms, $\mathcal{I} = \{m(\nu) \mid \nu \in \mathcal{D}\}$
$\mathcal{M} \subseteq \mathbb{R}^d$	Set of possible mean vectors
$\mu_i \in \mathcal{I}$	Mean of arm i , $\mu_i = \mathbb{E}_{X\sim\nu_i}[X]$
$\sigma_i^2 \in \mathbb{R}_+$	Variance of arm i , $\sigma_i^2 = \mathbb{E}_{X\sim\nu_i}[(X-\mu_i)^2]$
$\mu \in \mathcal{I}^K$	Vector of means, $\mu := (\mu_i)_{i\in[K]}$
$\sigma^2 \in \mathbb{R}_+^K$	Vector of variance, $\sigma^2 := (\sigma_i^2)_{i\in[K]}$
$\mu_\star \in \mathcal{I}$	Largest mean, $\mu_\star := \max_{i\in[K]} \mu_i$
$i^\star(\mu) \subseteq [K]$	Set of arms with highest mean, $i^\star(\mu) := \arg\max_{i\in[K]} \mu_i$
$i^\star \in [K]$	Arm with highest mean when unique (*i.e.* best arm), $i^\star(\mu) = \{i^\star\}$
$\Delta_i \in \mathbb{R}_+$	Gap of arm i , $\Delta_i := \mu_\star - \mu_i$
$\Delta_{\min}(\mu) \in \mathbb{R}_+^\star$	Smallest strictly positive gap, $\Delta_{\min}(\mu) := \min_{i\notin i^\star(\mu)}(\mu_\star - \mu_i)$
$\overline{\Delta}_{\min}(\mu) \in \mathbb{R}_+$	Minimum gap between any arm, $\overline{\Delta}_{\min}(\mu) := \min_{i\neq j} \lvert\mu_i - \mu_j\rvert$
$\Delta_{\max} \in \mathbb{R}_+$	Largest gap, $\Delta_{\max} := \max_{i\in[K]}(\mu_\star - \mu_i)$
$\neg i \subseteq \mathcal{I}^K$	Set of alternative parameters *s.t.* i is not a best arm, $\neg i := \{\lambda \in \mathcal{I}^K \mid i \notin i^\star(\lambda)\}$
$T^\star(\nu)$, $T_\beta^\star(\nu) \in \mathbb{R}_+^\star$	Asymptotic (β-)characteristic time for BAI
$w^\star(\nu)$, $w_\beta^\star(\nu) \in \Sigma_K$	Asymptotic (β-)optimal allocation for BAI
$\varepsilon \in \mathbb{R}_+^\star$	Slack parameter for ε-BAI (additive or multiplicative)
$\mathcal{I}_\varepsilon(\mu) \subseteq [K]$	ε-good arms (additive) for μ , $\mathcal{I}_\varepsilon(\mu) := \{i \in [K] \mid \Delta_i \leq \varepsilon\}$
$\mathcal{I}_\varepsilon^{\mathrm{mul}}(\mu) \subseteq [K]$	ε-good arms (multiplicative) for μ , $\mathcal{I}_\varepsilon^{\mathrm{mul}}(\mu) := \{i \in [K] \mid \Delta_i \leq \varepsilon\mu_\star\}$
$T_\varepsilon(\nu)$, $T_{\varepsilon,\beta}(\nu)$, $T_{\varepsilon,\beta}(\nu,i) \in \mathbb{R}_+^\star$	Asymptotic (β-)characteristic time for ε-BAI (*w.r.t.* arm $i \in \mathcal{I}_\varepsilon(\mu)$)
$w_\varepsilon(\nu)$, $w_{\varepsilon,\beta}(\nu)$, $w_{\varepsilon,\beta}(\nu,i) \in \Sigma_K$	Asymptotic (β-)optimal allocation for ε-BAI (*w.r.t.* arm $i \in \mathcal{I}_\varepsilon(\mu)$)
$\gamma \in \mathbb{R}$	Threshold parameter for GAI
$\mathcal{I}_\gamma^{\mathrm{thr}}(\mu) \subseteq [K]$	Good arms *w.r.t.* the threshold γ for the vector of means μ
$C(i,j;\nu,w)$, $C_\varepsilon(i,j;\nu,w)$	Transportation cost between arm i and arm j *w.r.t.* (ν,w)
$i_F(\nu)$, $i_F(\nu,w)$	Set of (instantaneous) furthest answers for bandit instance ν (resp. *w.r.t.* w)

Strategies

n , $T \in \mathbb{N}$	Time or budget
$\delta \in (0,1)$	Confidence at level $1-\delta$

τ_δ , $\tau_{\varepsilon,\delta}$, $\tau^{\mathrm{thr}}_{\gamma,\delta} \in \mathbb{N}$	Sample complexity, *i.e.* stopping time at confidence level $1-\delta$ for (ε-)BAI and GAI
$c(n,\delta) \in \mathbb{R}^\star_+$	Stopping threshold function at time n at confidence level $1-\delta$
$I_n \in [K]$	Arm sampled at time n
$N_{n,i} \in \mathbb{N}$	Number of pulls of arm i before time n , $N_{n,i} := \sum_{t\in[n-1]} \mathbb{1}(I_t = i)$
$\mu_{n,i} \in \mathbb{R}$	Empirical mean of arm i before time n , $\mu_{n,i} := N_{n,i}^{-1} \sum_{t\in[n-1]} \mathbb{1}(I_t = i) X_{n,i}$
$B_n \in [K]$	Leader answer at time n
$C_n \in [K]$	Challenger answer at time n
$X_{n,I_n} \in \mathbb{R}$	Sample observed at the end of time n , $X_{n,I_n} \sim \nu_{I_n}$
$\mathcal{F}_n$	History before time n , σ-algebra $\mathcal{F}_n := \sigma(U_1, I_1, X_{1,I_1}, \cdots, I_{n-1}, X_{n,I_{n-1}}, U_n)$
$\hat{\imath}_n \in [K]$	Answer recommended before time n
$T_n(i,j) \in \mathbb{N}$	Counts of $(B_t, C_t) = (i,j)$ before time n , $T_n(i,j) := \sum_{t\in[n-1]} \mathbb{1}((B_t, C_t) = (i,j))$
$T_n(i) \in \mathbb{N}$	Counts of $i \in \{B_t, C_t\}$ before time n , $T_n(i) := \sum_{t\in[n-1]} \mathbb{1}(i \in \{B_t, C_t\})$
$N^i_{n,j} \in \mathbb{N}$	Counts of $(B_t, C_t, I_t) = (i,j,j)$ before n , $N^i_{n,j} := \sum_{t\in[n-1]} \mathbb{1}((B_t, C_t, I_t) = (i,j,j))$
$\beta \in (0,1)$	Fixed proportion
$W_n(i,j)$, $W_{\varepsilon,n}(i,j)$	Empirical transportation cost for the pair of answers (i,j) before time n for (ε-)BAI
$\beta_n(i,j) \in (0,1)$	IDS proportion at time n for the leader/challenger pair (i,j)
ν_n	Empirical estimator of ν
$\bar{\beta}_n(i,j) \in (0,1)$	Averaged IDS proportion, $\bar{\beta}_n(i,j) := T_n(i,j)^{-1} \sum_{t\in[n-1]} \mathbb{1}((B_t, C_t) = (i,j))\, \beta_t(i,j)$
$\mathcal{E}^{\mathrm{err}}_\mu(n)$	error event at time n, *i.e.* $\hat{\imath}_n$ is not a correct answer for μ
Linear Bandits	
$d \in \mathbb{N}$	Dimension
$\theta \in \mathcal{M}$	Regression parameter
$\mathcal{A} \subseteq \mathbb{R}^d$	Set of arm vectors
$a \in \mathcal{A}$	Arm vector
$\mathcal{Z} \subseteq \mathbb{R}^d$	Set of answer vectors
$z \in \mathcal{Z}$	Answer vector
$Z \in \mathbb{N}$	Number of correct answers, $Z = \lvert\mathcal{Z}\rvert$
μ_a , $\mu_z \in \mathbb{R}$	Mean of an arm or an answer, $\mu_a := \langle\theta, a\rangle$ and $\mu_z := \langle\theta, z\rangle$
$z^\star(\theta) \subseteq \mathcal{Z}$	Set of answers with highest mean for the mean vector θ , $z^\star(\theta) := \arg\max_{z\in\mathcal{Z}} \langle\theta, z\rangle$
$\neg_\varepsilon z \subseteq \mathcal{M}$	Set of alternative *s.t.* z is not an ε-good answer, $\neg_\varepsilon z := \overline{\{\lambda \in \mathcal{M} \mid z \notin \mathcal{Z}_\varepsilon(\lambda)\}}$
$z_F(\theta) \subseteq \mathcal{Z}$	Set of furthest answers for the mean vector θ
$z_F(\theta, w) \subseteq \mathcal{Z}$	Set of furthest answers for the mean vector θ and the allocation w
$I_n \in \mathcal{A}$	Arm sampled at time n
$\hat{z}_n \in \mathcal{Z}$	Answer recommended before time n
$\theta_n \in \mathbb{R}^d$	Ordinary Least Square (OLS) estimator before time n , $\theta_n := V_{N_n}^{-1} \sum_{t\in[n-1]} X_{t,a_t} a_t$
$L_{\mathcal{X}} \in \mathbb{R}^\star_+$	Maximum ℓ_2-norm of an arm vector in $\mathcal{X}$, $L_{\mathcal{X}} := \max_{x\in\mathcal{X}} \lVert a\rVert_2$

List of References

Y. Abbasi-Yadkori, D.Pál, and C.Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems* (*NIPS*), 2011.

S. Agrawal, S. Juneja, and P. W. Glynn. Optimal δ-correct best-arm selection for heavy-tailed distributions. In *Algorithmic Learning Theory* (*ALT*), 2020.

S. Agrawal, S. K. Juneja, and W. M. Koolen. Regret minimization in heavy-tailed bandits. In *Conference on Learning Theory* (*COLT*), 2021a.

S. Agrawal, W. M. Koolen, and S. Juneja. Optimal best-arm identification methods for tail-risk measures. *Advances in Neural Information Processing Systems* (*NeurIPS*), 34, 2021b.

S. Agrawal, S. Juneja, K. Shanmugam, and A. S. Suggala. Optimal best-arm identification in bandits with access to offline data. *arXiv preprint arXiv:2306.09048*, 2023.

J.-Y. Audibert and S. Bubeck. Regret Bounds and Minimax Policies under Partial Monitoring. *Journal of Machine Learning Research*, 2010.

J.-Y. Audibert, S. Bubeck, and R. Munos. Best Arm Identification in Multi-armed Bandits. In *Conference on Learning Theory*, 2010.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

P. Auer, C.-K. Chiang, R. Ortner, and M. Drugan. Pareto front identification from stochastic bandit feedback. In *Artificial intelligence and statistics*, pages 939–947. PMLR, 2016.

O. Avner, S. Mannor, and O. Shamir. Decoupling exploration and exploitation in multi-armed bandits. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1107–1114, 2012.

A. Azize, M. Jourdan, A. Al Marjani, and D. Basu. On the complexity of differentially private best-arm identification with fixed confidence. *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023.

A. Bandyopadhyay, S. Juneja, and S. Agrawal. Optimal top-two method for best arm identification and fluid analysis. *arXiv preprint arXiv:2403.09123*, 2024.

A. Barrier, A. Garivier, and T. Kocák. A non-asymptotic approach to best-arm identification for gaussian bandits. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.

R. P. Brent. *Algorithms for minimization without derivatives*. Courier Corporation, 2013.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Fondations and Trends in Machine Learning*, 5(1):1–122, 2012.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory* (*ALT*), pages 23–37. Springer, 2009.

S. Bubeck, R. Munos, and G. Stoltz. Pure Exploration in Finitely Armed and Continuous Armed Bandits. *Theoretical Computer Science 412, 1832-1852*, 412:1832–1852, 2011.

A. Carpentier and A. Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Proceedings of the 29th Conference on Learning Theory* (*COLT*), 2016.

K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 1995.

L. Chen, A. Gupta, J. Li, M. Qiao, and R. Wang. Nearly optimal sampling algorithms for combinatorial pure exploration. In *Proceedings of the 30th Conference on Learning Theory* (*COLT*), 2017a.

L. Chen, J. Li, and M. Qiao. Towards instance optimal bounds for best arm identification. *Conference on Learning Theory*, 2017b.

L. Chen, J. Li, and M. Qiao. Nearly instance optimal sample complexity bounds for top-k arm selection. In *Artificial Intelligence and Statistics*, 2017c.

W. Chen, Y. Wang, and Y. Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, 2013.

Y. Chen and I. O. Ryzhov. Balancing optimal large deviations in sequential selection. *Management Science*, 69(6):3457–3473, 2023.

H. Chernoff. Sequential design of Experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959.

J. Cheshire, P. Menard, and A. Carpentier. Problem Dependent View on Structured Thresholding Bandit Problems. In *International Conference on Machine Learning* (*ICML*), 2021.

S. R. Chowdhury, P. Saux, O. Maillard, and A. Gopalan. Bregman deviations of generic exponential families. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 394–449. PMLR, 2023.

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

W. Cowan, J. Honda, and M. N. Katehakis. Normal bandits of unknown means and variances. *Journal of Machine Learning Research*, 18:154:1–154:28, 2017.

S. de Rooij, T. van Erven, P. D. Grünwald, and W. M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15(1):1281–1316, 2014.

R. Degenne. *Impact of structure on the design and analysis of bandit algorithms*. PhD thesis, Université de Paris, 2019.

R. Degenne. On the existence of a complexity in fixed budget bandit identification. *International Conference on Learning Theory* (*COLT*), 2023.

R. Degenne and W. M. Koolen. Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems* (*NeurIPS*), 2019.

R. Degenne, W. M. Koolen, and P. Ménard. Non-asymptotic pure exploration by solving games. In *Advances in Neural Information Processing Systems* (*NeurIPS*), 2019.

R. Degenne, P. Ménard, X. Shang, and M. Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning* (*ICML*), 2020a.

R. Degenne, H. Shao, and W. M. Koolen. Structure adaptive algorithms for stochastic bandits. In *International Conference on Machine Learning* (*ICML*), 2020b.

E. Even-Dar, S. Mannor, and Y. Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, 2002.

E. Even-Dar, S. Mannor, and Y. Mansour. Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

M. Faella, A. Finzi, and L. Sauro. Rapidly finding the best arm using variance. In *European Conference on Artificial Intelligence*, 2020.

W. Fan, L. J. Hong, and B. L. Nelson. Indifference-zone-free selection of the best. *Operations Research*, 64: 1499–1514, 2016.

L. Faury. *Variance-sensitive confidence intervals for parametric and offline bandits*. PhD thesis, Institut Polytechnique de Paris, 2021.

T. Fiez, L. Jain, K. G. Jamieson, and L. J. Ratliff. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems* (*NeurIPS*), 2019.

S. Filippi, O. Cappé, A. Garivier, and C. Szepesvári. Parametric Bandits : The Generalized Linear case. In *Advances in Neural Information Processing Systems*, 2010.

V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence. In *Advances in Neural Information Processing Systems*, 2012.

A. Garivier and E. Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference On Learning Theory*, 2016.

A. Garivier and E. Kaufmann. Non-asymptotic sequential tests for overlapping hypotheses and application to near optimal arm identification in bandit models. *Sequential Analysis*, 40(1):61–96, 2021.

A. Garivier, P. Ménard, and G. Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathemathics of Opereration Research*, 44(2):377–399, 2019.

J. Honda and A. Takemura. An Asymptotically Optimal Bandit Algorithm for Bounded Support Models. In *Proceedings of the 23rd Conference on Learning Theory*, 2010.

J. Honda and A. Takemura. Optimality of Thompson Sampling for Gaussian Bandits depends on priors. In *Proceedings of the 17th conference on Artificial Intelligence and Statistics*, 2014.

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

L. Hong, W. Fan, and J. Luo. Review on ranking and selection: A new perspective. *Frontiers of Engineering Management*, 8:321–343, 2021.

G. Hoogenboom, C. Porter, K. Boote, V. Shelia, P. Wilkens, U. Singh, J. White, S. Asseng, J. Lizaso, L. Moreno, et al. The dssat crop modeling ecosystem. *Advances in crop modelling for a sustainable agriculture*, pages 173–216, 2019.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49:1055 – 1080, 2021.

R. Huang, M. M. Ajallooeian, C. Szepesvári, and M. Müller. Structured best arm identification with fixed confidence. In *International Conference on Algorithmic Learning Theory* (*ALT*), 2017.

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'UCB: an Optimal Exploration Algorithm for Multi-Armed Bandits. In *Proceedings of the 27th Conference on Learning Theory*, 2014.

K. Jamieson, S. Katariya, A. Deshpande, and R. Nowak. Sparse Dueling Bandits. In *Proceedings of the 18th Conference on Artificial Intelligence and Statistics*, 2015.

K. G. Jamieson and R. D. Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *Conference on Information Sciences and Systems* (*CISS*), 2014.

Y. Jedra and A. Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017, 2020.

M. Jourdan and R. Degenne. Choosing answers in ε-best-answer identification for linear bandits. In *International Conference on Machine Learning* (*ICML*), 2022.

M. Jourdan and R. Degenne. Non-asymptotic analysis of a ucb-based top two algorithm. *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023.

M. Jourdan and C. Réda. An anytime algorithm for good arm identification. *arXiv preprint arXiv:2310.10359*, 2023.

M. Jourdan, M. Mutnỳ, J. Kirschner, and A. Krause. Efficient pure exploration for combinatorial bandits with semi-bandit feedback. In *Algorithmic Learning Theory* (*ALT*), 2021.

M. Jourdan, R. Degenne, D. Baudry, R. De Heide, and E. Kaufmann. Top two algorithms revisited. *Advances in Neural Information Processing Systems*, 2022.

M. Jourdan, R. Degenne, and E. Kaufmann. Dealing with unknown variances in best-arm identification. *International Conference on Algorithmic Learning Theory*, 2023a.

M. Jourdan, R. Degenne, and E. Kaufmann. An ε-best-arm identification algorithm for fixed-confidence and beyond. *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023b.

K.-S. Jun and R. Nowak. Anytime exploration for multi-armed bandits using confidence information. In *International Conference on Machine Learning* (*ICML*), pages 974–982. PMLR, 2016.

K.-S. Jun, L. Jain, H. Nassif, and B. Mason. Improved Confidence Bounds for the Linear Logistic Model and Applications to Bandits. In *International Conference on Machine Learning*, 2021.

S. Juneja and S. Krishnasamy. Sample complexity of partition identification using multi-armed bandits. In *Conference on Learning Theory* (*COLT*), 2019.

S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. PAC subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning* (*ICML*), 2012.

H. Kano, J. Honda, K. Sakamaki, K. Matsuura, A. Nakamura, and M. Sugiyama. Good arm identification via bandit feedback. *Machine Learning*, 108(5):721–745, 2019.

Z. Karnin, T. Koren, and O. Somekh. Almost optimal Exploration in multi-armed bandits. In *International Conference on Machine Learning* (*ICML*), 2013.

J. Katz-Samuels and K. Jamieson. The true sample complexity of identifying good arms. In *International Conference on Artificial Intelligence and Statistics*, 2020.

J. Katz-Samuels and C. Scott. Top Feasible Arm Identification. In *International Conference on Artificial Intelligence and Statistics* (*AISTATS*), 2019.

J. Katz-Samuels, L. Jain, K. G. Jamieson, et al. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. *Advances in Neural Information Processing Systems*, 2020.

E. Kaufmann. Contributions to the optimal solution of several bandit problems. *Université de Lille*, 2020.

E. Kaufmann and S. Kalyanakrishnan. Information complexity in bandit subset selection. In *Proceeding of the 26th Conference On Learning Theory.*, 2013.

E. Kaufmann and W. M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021.

E. Kaufmann, O. Cappé, and A. Garivier. On the Complexity of A/B Testing. In *Proceedings of the 27th Conference On Learning Theory*, 2014.

E. Kaufmann, O. Cappé, and A. Garivier. On the Complexity of Best Arm Identification in Multi-Armed Bandit Models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

E. Kaufmann, W. Koolen, and A. Garivier. Sequential test for the lowest mean: From Thompson to Murphy Sampling. In *Advances in Neural Information Processing Systems* (*NeurIPS*), 2018.

S. Kim and B. Nelson. A fully sequential procedure for indifference-zone selection in simulation. *ACM Transactions on Modeling and Computer Simulation*, 11:251–273, 2001.

T. Kocák and A. Garivier. Epsilon best arm identification in spectral bandits. In *International Joint Conference on Artificial Intelligence* (*IJCAI*), pages 2636–2642, 2021.

J. Komiyama, T. Tsuchiya, and J. Honda. Minimax optimal algorithms for fixed-budget best arm identification. In *Advances in Neural Information Processing Systems*, 2022.

C. Kone, E. Kaufmann, and L. Richert. Adaptive algorithms for relaxed pareto set identification. *Advances in Neural Information Processing Systems* (*NeurIPS*), 2023.

C. Kone, E. Kaufmann, and L. Richert. Bandit pareto set identification: the fixed budget setting. *International Conference on Artificial Intelligence and Statistics* (*AISTATS*), 2024.

T. Lai. Nearly Optimal Sequential Tests for Composite Hypotheses. *Annals of Statistics*, 16(2):856–886, 1988.

T. Lattimore and C. Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2019.

J. Lee, J. Honda, and M. Sugiyama. Thompson exploration with best challenger rule in best arm identification. *arXiv preprint arXiv:2310.00539*, 2023.

A. Locatelli, M. Gutzeit, and A. Carpentier. An optimal algorithm for the thresholding bandit problem. In *International Conference on Machine Learning* (*ICML*), 2016.

P. Lu, C. Tao, and X. Zhang. Variance-dependent best arm identification. In *Uncertainty in Artificial Intelligence*, 2021.

S. Magureanu, R. Combes, and A. Proutière. Lipschitz Bandits: Regret lower bounds and optimal algorithms. In *Proceedings on the 27th Conference On Learning Theory*, 2014.

S. Mannor and J. Tsitsiklis. The Sample Complexity of Exploration in the Multi-Armed Bandit Problem. *Journal of Machine Learning Research*, pages 623–648, 2004.

A. A. Marjani, T. Kocák, and A. Garivier. On the complexity of all ε-best arms identification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 317–332. Springer, 2022.

B. Mason, L. Jain, A. Tripathy, and R. Nowak. Finding all ε-good arms in stochastic bandits. *Advances in Neural Information Processing Systems*, 2020.

A. Maurer and M. Pontil. Empirical bernstein bounds and sample-variance penalization. In *Conference on Learning Theory*, 2009.

P. Ménard. Gradient ascent for active exploration in bandit problems. *arXiv 1905.08165*, 2019.

B. O'Donoghue and T. Lattimore. Variational bayesian optimistic sampling. *Advances in Neural Information Processing Systems*, 34:12507–12519, 2021.

R. Ouhamma, O.-A. Maillard, and V. Perchet. Online sign identification: Minimization of the number of errors in thresholding bandits. *Advances in Neural Information Processing Systems*, 34:18577–18589, 2021.

F. Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

C. Qin and W. You. Dual-directed algorithm design for efficient pure exploration. *arXiv preprint arXiv:2310.19319*, 2023.

C. Qin, D. Klabjan, and D. Russo. Improving the expected improvement algorithm. In *Advances in Neural Information Processing Systems 30* (*NIPS*), 2017.

C. Réda, E. Kaufmann, and A. Delahaye-Duriez. Top-m identification for linear bandits. In *International Conference on Artificial Intelligence and Statistics* (*AISTATS*), 2021.

C. Réda, A. Tirinzoni, and R. Degenne. Dealing with misspecification in fixed-confidence linear top-m identification. *Advances in Neural Information Processing Systems* (*NeurIPS*), 2021.

C. Riou and J. Honda. Bandit algorithms based on thompson sampling for bounded reward distributions. In *Algorithmic Learning Theory* (*ALT*), 2020.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

H. Robbins and D. Siegmund. The expected sample size of some tests of power one. *The Annals of Statistics*, 2(3):415–436, 1974.

C. Rouyer and Y. Seldin. Tsallis-inf for decoupled exploration and exploitation in multi-armed bandits. In *Conference on Learning Theory*, pages 3227–3249. PMLR, 2020.

Y. Russac, C. Katsimerou, D. Bohle, O. Cappé, A. Garivier, and W. M. Koolen. A/b/n testing with control in the presence of subpopulations. In *Advances in Neural Information Processing Systems* (*NeurIPS*), 2021.

D. Russo. Simple Bayesian algorithms for best arm identification. In *Proceedings of the 29th Conference on Learning Theory* (*COLT*), 2016.

X. Shang, R. de Heide, E. Kaufmann, P. Ménard, and M. Valko. Fixed-confidence guarantees for bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics* (*AISTATS*), 2020.

D. Shin, M. Broadie, and A. Zeevi. Tractable sampling strategies for ordinal optimization. *Operations Research*, 66(6):1693–1712, 2018.

M. Simchowitz, K. Jamieson, and B. Recht. The simulator: Understanding adaptive sampling in the moderate-confidence regime. In *Conference on Learning Theory*, 2017.

M. Soare, A. Lazaric, and R. Munos. Best Arm Identification in Linear Bandit. In *Advances in Neural Information Processing Systems*, 2014.

K. Tabata, A. Nakamura, J. Honda, and T. Komatsuzaki. A bad arm existence checking problem: How to utilize asymmetric problem structure? *Machine learning*, 109(2):327–372, 2020.

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

A. Tirinzoni and R. Degenne. On elimination strategies for bandit fixed-confidence identification. *Advances in Neural Information Processing Systems*, 2022.

A. Tirinzoni, M. Pirotta, M. Restelli, and A. Lazaric. An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. *Advances in Neural Information Processing Systems* (*NeurIPS*), 2020.

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

A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106:25–57, 2006.

A. Wald. Sequential Tests of Statistical Hypotheses. *Annals of Mathematical Statistics*, 16(2):117–186, 1945.

P.-A. Wang, R.-C. Tzeng, and A. Proutiere. Fast pure exploration via frank-wolfe. *Advances in Neural Information Processing Systems*, 2021.

P.-A. Wang, R.-C. Tzeng, and A. Proutiere. Best arm identification with fixed budget: A large deviation perspective. *Advances in Neural Information Processing Systems*, 36, 2024.

Y. Wang, Y. Han, J. Jiao, and D. Tse. Beyond the best: Estimating distribution functionals in infinite-armed bandits. *Advances in Neural Information Processing Systems*, 2022.

L. Xu, J. Honda, and M. Sugiyama. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics* (*AISTATS*), 2018.

W. You, C. Qin, Z. Wang, and S. Yang. Information-directed selection for top-two algorithms. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2850–2851. PMLR, 2023.

Y. Zhao, C. J. Stephens, C. Szepesvári, and K.-S. Jun. Revisiting simple regret: Fast rates for returning a good arm. In *40th International Conference on Machine Learning* (*ICML*), 2023.