Contributions to the Optimal Solution of Several Bandit Problems

Emilie Kaufmann

Université de Lille

HABILILITATION À DIRIGER DES RECHERCHES

Contributions to the Optimal Solution of Several Bandit Problems

Contributions à la résolution optimale de différents problèmes de bandit

présentée par Emilie Kaufmann

Ecole Doctorale des Sciences Pour l'Ingénieur
Discipline: Informatique

cnrs

CRIStAL
Centre de Recherche en Informatique, Signal et Automatique de Lille

Inria

soutenue le **13 novembre 2020** devant le jury composé de:

Francis Bach	Directeur de recherche	Inria & ENS Paris	Garant
Alexandra Carpentier	Professeure	Otto-von-Guericke-Universität Magdeburg	Examinatrice
Eric Moulines	Professeur	Ecole Polytechnique	Rapporteur
Philippe Preux	Professeur	Université de Lille	Président
Alexandre Proutière	Professeur	KTH Royal Institute of Technology	Rapporteur
Gilles Stoltz	Directeur de recherche	CNRS & Université Paris-Sud	Rapporteur
Csaba Szepesvari	Professeur	University of Alberta	Examinateur

Abstract - *Résumé*

This document presents in a unified way different results about the optimal solution of several multi-armed bandit problems. We present and analyze algorithms for sequential decision making that adaptively sample several probability distributions with unknown characteristics, in order to achieve different types of objectives. Our contributions cover two types of problems. On the one hand, we study rewards maximization in some variants of the classical bandit model and on the other hand we focus and so-called active identification problems, in which there is no incentive to maximize reward, but one should optimize exploration in order to answer some (possibly complex) question about the underlying distributions. We highlight several common tools for solving these problems. First, lower bounds, that not only permit to assess the optimality of an algorithm, but also guide the design of asymptotically optimal algorithms. We indeed provide several examples of lower-bound-inspired algorithms. Then, we emphasize the importance of time-uniform self-normalized concentration inequalities to analyze algorithms. Finally, on the algorithmic side, we present several variants of an important Bayesian principle called Thompson Sampling, which leads to easy-to-implement asymptotically optimal algorithms in some particular cases.

Ce document présente d'une manière unifiée plusieurs résultats liés à la résolution optimale de problèmes dits de bandit à plusieurs bras. Nous présentons et analysons des algorithmes pour la prise de décision séquentielle qui échantillonnent de manière adaptative des distributions de probabilités ayant des caractéristiques inconnues, dans le but de remplir différents types d'objectifs. Nous présentons des contributions pour la résolution de deux types de problèmes. D'une part nous nous intéressons à la maximization de récompenses dans des variantes du modèle de bandit classique, et d'autre part nous étudions différents problèmes d'identification active, pour lesquels l'objectif est d'optimiser l'exploration de l'environement de sorte à pouvoir répondre une question (possiblement complexe) sur les distributions sous-jacentes, mais sans la contrainte de maximiser des récompenses. Nous mettons en avant plusieurs outils communs pour traiter ces deux types de problèmes. Tout d'abord l'utilisation de bornes inférieures, qui permettent non seulement de valider l'optimalité d'un algorithme, mais qui servent aussi à guider la conception d'algorithmes asymptotiquement optimaux. Nous présentons en effet plusieurs exemples d'algorithmes inspirés par des bornes inférieures. Ensuite, nous insistons sur l'importance des inégalités de concentration auto-normalisées et uniformes en temps pour l'analyse d'algorithmes de bandits. Enfin, nous présentons plusieurs variantes d'un important principe bayésien appelé l'échantillonnage de Thompson, qui conduit à des algorithmes asymptotiquement optimaux et faciles d'implémentation dans certains cas particuliers.

Remerciements

Je remercie chaleureusement Francis Bach d'avoir accepté de se porter garant de cette habilitation à diriger des recherches. Je suis également très honorée qu'Eric Moulines, Alexandre Proutière et Gilles Stoltz aient accepté de rapporter ce manuscrit. Un grand merci pour l'intérêt que vous avez porté à ce travail, et pour vos précieux commentaires et remarques. Merci enfin à Alexandra Carpentier et Csaba Szepesvari pour leur participation au jury, et à Philippe Preux qui en a assuré la présidence.

Ces cinq années au sein de l'équipe SequeL m'ont beaucoup apporté scientifiquement et humainement. J'y ai eu l'occasion d'y côtoyer des personnes formidables, que je remercie dans ces lignes. Je remercie en particulier Philippe Preux et Odalric-Ambrym Maillard d'avoir été à mes côtés tout au long de l'aventure (qui a vu de nombreux collègues de l'équipe se faire kidnapper par les GAFAMs), et de m'avoir fourni un environnement de recherche extrêmement favorable, où je n'ai pas eu beaucoup à me préoccuper de chercher des financements. Merci aussi aux anciens membres permanents de l'équipe pour les bons moments passés ensemble, et aux nouveaux qui rejoignent l'aventure Scool, Jill-Jênn Vie (oui, elle aurait dû commencer il y a un an !), Rémy Degenne et Debabrota Basu. L'ambiance à SequeL n'aurait pas été la même sans les nombreux doctorants et post-doctorants avec qui il est toujours très agréable d'échanger, et qui, en plus de faire du bon café et de me mettre une tolée au badminton, sont toujours prêts à contribuer à, voire à impulser l'organisation d'événements scientifiques, comme EWRL en 2018 et RLSS en 2019. Merci à eux ! Un merci spécial aux relecteurs de ce document.

Parmi les meilleurs moments de ces dernières années se trouvent les nombreuses "white-board sessions" avec les étudiants avec lesquels j'ai eu la chance de travailler. Un merci spécial aux doctorant.e.s dont j'ai co-dirigé ou co-dirige en ce moment la thèse : Lilian Besson, Xuedong Shang, Omar Darwiche Domingues, Clémence Réda et Dorian Baudry. Leurs qualités les promettent à un brillant avenir ! Chaque thèse est différente, et j'ai beaucoup appris sur le "métier" de directrice de thèse au contact des collègues avec qui j'ai encadré ces thèses : Christophe Moy, qui m'a fait confiance en premier, Andrée Delahaye-Duriez, qui fait l'effort de se mettre aux bandits tandis que j'apprends la génétique, Michal Valko, que je remercie pour ses nombreux conseils, et Odalric Maillard, qui a toujours tant d'idées de projets. Et parce que je ne les ai jamais assez remerciés pour tout ce qu'ils ont fait pour moi, je salue également les meilleurs directeurs de thèse au monde, Aurélien Garivier et Olivier Cappé, à qui je dois beaucoup et qui continuent de m'inspirer maintenant que je suis à leur place.

Je remercie bien évidemment aussi tous mes co-auteurs déjà docteurs, avec une mention spéciale pour Pierre Ménard dont les talents mathématiques m'éblouissent chaque jour, Aurélien Garivier et Wouter Koolen pour une longue et fructueuse collaboration sur les problèmes d'exploration pure qui ont une place importante dans ce document et Anders Jonsson qui en venant passer un an à Lille m'a poussée à me mettre plus sérieusement à l'apprentissage par renforcement pour le projet DELTA. Au-delà des collaborateurs directs, c'est toujours un plaisir de rencontrer, lors de conférences, des membres de la communauté "bandits" francophone (qu'ils soient BADASS, BOLD ou à la recherche d'un nouvel acronyme sexy pour un projet de recherche) ou internationale. Il ne reste qu'à espérer qu'on puisse se revoir bientôt autrement qu'à travers un écran...

Enfin, je remercie ma famille et mes amis pour tous les moments de bonheur que nous partageons, et pour tant d'autres à venir.

Contents

Introduction **9**
1 Multi-Armed Bandit Problems 9
2 Theoretical Framework and Important Tools 12
3 List of Associated Publications 17

I Maximizing Rewards, with a Twist **21**

1 Optimal Solution for Variants of the Classical Multi-Armed Bandit **23**
1.1 Thompson Sampling in Transductive Settings 23
1.1.1 Thompson Sampling for Corrupt Bandits 23
1.1.2 Thompson Sampling for Dose-Finding 28
1.2 Structured Bandits 31
1.3 Thompson Sampling for Rank-One Bandits 37

2 Multi-Player Bandits **43**
2.1 Several Decentralized Bandit Problems 43
2.2 Algorithms for the Homogeneous Multi-Player MAB 46
2.2.1 The Selfish and MCTopM algorithms 46
2.2.2 Elements of analysis of MCTopM 48
2.2.3 Empirical Evaluation 51
2.3 Towards Optimal Multi-Player Algorithms 52
2.4 An Algorithm Exploiting Collisions for the Heterogeneous Case 55

II Active Identification Problems **63**

3 A Universal Stopping Rule for Active Identification **65**
3.1 Active Identification in a Bandit Model 65
3.1.1 Examples 66
3.1.2 Several mathematical frameworks 68
3.2 The Parallel GLRT Stopping Rule 69
3.2.1 Definition of the Parallel GLRT 69
3.2.2 Simple Examples 70

3.3 Correctness of the Parallel GLRT stopping rule . . . 71
3.3.1 A New Deviation Inequality . . . 71
3.3.2 Proof of Theorem 3.1: A Martingale Story . . . 73
3.4 From Sequential to Active Testing . . . 79

4 Towards Optimal and Efficient Best Arm Identification 81
4.1 Lower Bounds for Best Arm Identification and Beyond . . . 81
4.2 The Track-and-Stop Algorithm . . . 83
4.2.1 Computing the Optimal Allocation . . . 83
4.2.2 Track-and-Stop and Its Analysis . . . 86
4.3 Beyond Track-and-Stop . . . 89
4.3.1 Online Optimization and Optimism . . . 90
4.3.2 Bayesian Approaches to the Rescue . . . 92

5 Applications to Monte-Carlo Tree Search 97
5.1 A Simple Model for Planning in Games . . . 97
5.2 Monte-Carlo Tree Search by Best Arm Identification . . . 99
5.2.1 UGapE-MCTS . . . 99
5.2.2 Towards Optimal Strategies . . . 104
5.3 From Thompson Sampling to Murphy Sampling . . . 105
5.3.1 Sampling Rule: Murphy Sampling . . . 107
5.3.2 Stopping Rules . . . 110

Perspective 115

Index of Notation 121

Bibliography 125

Introduction

This manuscript describes some research directions I followed since my PhD defense in October 2014. After one very interesting year of post-doc at Inria Paris, in which I worked on community detection with Marc Lelarge and Thomas Bonald (Kaufmann et al., 2017), I joined the CRIStAL computer science lab at the University of Lille as a CNRS junior researcher, where I am also a member of the Inria team Scool (the brand new name of SequeL). Being in contact with experts of sequential decision making, I naturally went back to my main research interest: sequential learning and in particular multi-armed bandit problems. This document gives an overview of the contributions I made to this field after my PhD, in collaboration with several students and colleagues.

1 Multi-Armed Bandit Problems

The multi-armed bandit model is often associated to the simplest reinforcement learning problem, in which an agent repeatedly selects an action among a fixed set of actions, with the goal to maximize the total reward collected when performing these actions. I want to argue here that it is much more than that, and that there is no such thing as "the" multi-armed bandit problem.

From clinical trials to online content recommendation The stochastic Multi-Armed Bandit (MAB) model was historically introduced as a simple model for a sequential clinical trial (Thompson, 1933; Robbins, 1952). Imagine for example that a doctor is investigating the efficacy of some drugs for a new disease on a population of similar patients. Each drug a is associated with an unknown probability of efficacy p_a, that is unknown to the doctor prior to the trial.

p_1 p_2 p_3 p_4 p_5

For the t-th patient involved in the trial, the doctor has to select a treatment $A_t \in \{1, \dots, A\}$ in a pool of A available treatments ($A = 5$ in the above illustration). After giving the treatment, the doctor observes a response X_t that is assumed to be drawn from a Bernoulli distribution with parameter p_{A_t} such that $X_t = 1$ if the treatment is successful and $X_t = 0$ otherwise. Assuming the doctor wants to cure as many patients as possible, this response X_t can be interpreted as a reward signal. In trials targeted towards maximizing the number of patients cured, the objective can also be phrased as maximizing the total reward gathered during the trial.

This sequential interaction between the doctor and the treatments fits the more general framework of a stochastic multi-armed bandit model, in which an agent (the doctor) is facing a collection of arms (the treatments) that are unknown probability distributions. The term "arm" is used in reference to the arm of a slot-machine (also called one-armed bandit) in a casino, as drawing the arm of a slot-machine delivers some random reward. In each time step $t = 1, \dots, T$ of the interaction, the agent selects an arm A_t and observes a sample (often called reward) X_t from the distribution associated to the chosen arm A_t. The interaction is sequential in that the arm selected at time t can be *adaptively* chosen based on the observation made in previous rounds, $A_1, X_1, \dots, A_{t-1}, X_{t-1}$. The agent can *learn* from previous observations. Back to the example of clinical trials, the doctor may want to progressively focus on the treatments that look more promising based on the outcomes observed so far.

However, the drug development pipeline is complex and despite the very natural model proposed for phase III clinical trials (in which different treatments are compared, while earlier phases assess the safety and efficacy of a given treatment), it appears that bandit algorithms have been seldom used for clinical trials (see e.g. Réda et al. (2020)). More recently, different types of applications to online content optimization have justified a regain of interest for bandit models. A website that displays advertisement aims at picking, for each visitor, an add on which the visitor has a high probability to click, as clicks generate revenue (reward). A good recommender systems aims at proposing, for each user, an item that the user will like (a good rating can be seen as a reward).

This type of applications have motivated the study of different variants of multi-armed bandit models, for example contextual bandits, in which the average reward also depends on some characteristics of the user or items (see, e.g. Abbasi-Yadkori et al. (2011); Chapelle and Li (2011)). But many other variants of bandit models have been proposed in the literature such as combinatorial bandits (Chen et al., 2013) in which subsets of arms can be selected, duelling bandits (Komiyama et al., 2016) in which one can sequentially perform pairwise comparisons between arms, structured bandits (Combes et al., 2017) in which some prior knowledge about the arm means can speed up learning, or partial monitoring (Lattimore and Szepesvári, 2019) in which the reward is not directly observed.

This list is by far not exhaustive and shows that the study of multi-armed bandits is now a research field in its own, that cannot be reduced to a simple case of reinforcement learning. The interested reader can refer to the introductory and quite exhaustive survey of Lattimore and Szepesvari (2019) or to that of Slivkins (2019) to discover more bandit problems.

Should we maximize reward? A lot of these bandit problems can indeed be related to reinforcement learning as the learning process is targeted toward maximizing some notion of cumulative reward. Maximizing rewards requires to achieve a balance between exploration (learning the unknown distribution of all arms based on samples) and exploitation (trying to focus on the arms that may lead to more reward according to current knowledge). However, not all bandit problems have this incentive on exploitation, and several *pure exploration* problems have been studied in the literature. In pure exploration, one should sample the arms so as to gain relevant information on the model quickly, regardless of rewards.

In order to give a concrete example of a pure exploration problem, we can go back to the initial example of clinical trials. As explained above, in this context maximizing the sum of rewards gathered during the trial amounts to maximize the number of patients healed. However, in most cases the main purpose of clinical trials is not therapeutic. In a phase III trial, the goal is rather to identify one treatment that will later be produced and given to a much larger population than that involved in the trial. Hence one needs to be very sure that all these resources are invested in the most successful treatment, even at the cost of curing less patients during the trial. In the multi-armed bandit language this means that instead of maximizing the sum of rewards, the goal is to output a guess $\hat{a}$ for the arm $a_\star$ with largest mean reward. This guess should be as accurate as possible, and based on as few observation as possible.

This particular pure exploration task is often referred to as the best arm identification problem, and can have different mathematical formulations.

Ideally, one may want to identify the best treatment as quickly as possible while maximizing the number of patients cured. However, the bandit literature tells us that it is not possible to find a sampling strategy (A_t) that solves both objectives optimally. This has been known since the work of Bubeck et al. (2011). In a joint work with Aurélien Garivier (Garivier and Kaufmann, 2016), in which we identify the minimal number of samples needed to produce a guess that satisfies $\mathbb{P}(\hat{a} = a_\star) \geq 1 - \delta$, we also contributed to a better understanding of the difference between algorithms for (fixed-confidence) best arm identification and algorithms for maximizing rewards. This fundamental difference is further discussed in the paper Kaufmann and Garivier (2017).

Beyond the best arm identification problem, several other pure exploration problems have been studied in the literature. These alternative objectives can be informally described as sampling the arms adaptively in order to quickly learn *something* about the underlying unknown means. For example in the thresholding bandit problem (Locatelli et al., 2016) the goal is to identify all arms whose means lie below a certain threshold. This fits the more generic objective of identifying to which fold of a partitioning the vector of means of the arms belongs. This problem has been studied for example by Juneja and Krishnasamy (2019). In this document, we will present several examples of such general pure exploration problems, which we shall refer to as *active identification* problems.

Content of this document The research works presented in this document contribute to two research directions, that correspond to the two parts of this manuscript.

First, I have been working on the design and analysis of new algorithms for different reward maximization tasks, beyond the classical multi-armed bandit model. Although I cannot claim to be a "practitioner", the different works presented in part I have been motivated by different applications, or at least by colleagues working on those applications: privacy preserving in recommender systems, early stage clinical trials and adaptive channel selection for cognitive radio. In **Chapter 1**, we will present some variants of a Bayesian algorithm called Thompson Sampling for two different contexts: first, in a setting in which the reward is not observed or in which there is no clear notion of reward and then for a particular example of a structured bandit. In **Chapter 2**, we will present our contributions to the study of multi-player bandit problems, that are mostly the outcome of my first PhD co-supervision. Lilian Besson defended his PhD on multi-player bandit algorithms applied to telecommunications in November 2019, that he did in CentraleSupélec Rennes, under the supervision of Christophe Moy and myself.

The second line of research presented in this document is focused on active identification problems. My main contribution in this field is the outcome of a collaboration with Aurélien Garivier, in which we proposed a new lower bound on the sample complexity of fixed-confidence best arm identification as well as the first algorithm whose sample complexity asymptotically matches the lower bound (Garivier and Kaufmann, 2016). The Track-and-Stop strategy proposed in this work can be easily extended to other types of active identification problems, as highlighted in part II of this document. In **Chapter 3**, we will properly introduce active identification problems and present a generic *stopping rule*, based on parallel Generalized Likelihood Ratio Tests. Notably, we will highlight a new concentration inequality used to prove the correctness of this stopping rule, which was obtained with Wouter Koolen using some nice martingale techniques. In **Chapter 4**, through the best arm identification example, we will focus on *sampling rules* for active identification, and on the sample complexity of their combination with the stopping rule previously studied. In **Chapter 5**, we will investigate the (optimal) solution of two particular active identification problems that we studied with Wouter Koolen and Aurélien Garivier. Both are motivated by providing sample complexity guarantees for Monte-Carlo Tree Search algorithms.

Both parts share a few common ingredients that we emphasize in the next section, in which we also give a formal introduction to the standard multi-armed bandit model.

2 Theoretical Framework and Important Tools

We denote by $\nu_1, \nu_2, \dots, \nu_A$ the distributions associated to each arm a in $[A]$, where for each integer n, $[n]$ is a shorthand for the set $\{1, \dots, n\}$. We denote by $\mu_a = \mathbb{E}_{X\sim\nu_a}[X]$ the mean of arm a. In each round $t = 1, \dots, T$, an agent (or learner), who is unaware of the means of the arms (and possibly of the horizon T), selects an arm $A_t \in [A]$ and subsequently observes a sample X_t from the distribution ν_{A_t}, assumed to be independent from previous samples. For each $a \in [A]$, defining $(X_{a,t})_{t\in\mathbb{N}^*}$ an i.i.d. sequence distributed under ν_a, we have $X_t = X_{A_t,t}$. Equivalently, one can define $(Y_{a,s})_{s\in\mathbb{N}^*}$ to be the i.i.d. sequence of successive observations from arm a, and let $X_t = Y_{a,N_a(t)}$ where $N_a(t)$ is the number of selections of arm a up to round t.

A *sampling strategy* or *bandit algorithm* is a sequence $\mathcal{A} = (A_t)_{t\in\mathbb{N}}$ for which A_t may depend on previous observation and some exogenous randomness. Formally, A_t is $\mathcal{F}_{t-1}$ measurable where $\mathcal{F}_t = \sigma(U_0, A_1, X_1, U_1, \dots, A_t, X_t, U_t)$ is the σ-algebra generated by the observations available up to round t, where $U_t \sim \mathcal{U}([0,1])$ materializes the possible independent randomness used by the algorithm.

Performance measures In the reward maximization objective, the samples X_t are viewed as rewards and the goal of the agent is to design a strategy to maximize the expected sum of rewards, $\mathbb{E}_{\boldsymbol{\nu}}\left[\sum_{t=1}^T X_t\right]$ where $\mathbb{P}_{\boldsymbol{\nu}}$ and $\mathbb{E}_{\boldsymbol{\nu}}$ denote the probability and expectation under the bandit model with distributions $\boldsymbol{\nu} = (\nu_1, \dots, \nu_A)$. Letting $a_\star \in \operatorname{argmax}_{a\in[A]} \mu_a$ be the arm with largest mean, an oracle optimal strategy for maximizing reward consists in always selecting arm $a_\star$, whose mean is denoted by $\mu_\star$. The performance of an algorithm is often measured by the gap between the cumulative rewards of this oracle strategy and that of the algorithm, called *regret*. The (expected) regret of an algorithm $\mathcal{A}$ in T rounds is defined by

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

A simple conditioning argument permits to express the regret in terms of number of times each arm has been selected. Letting $N_a(t) = \sum_{s=1}^t \mathbb{1}(A_s = a)$ be the number of selections of arm a, it holds that

$$\mathcal{R}_{\boldsymbol{\nu}}(\mathcal{A}, T) = \sum_{a=1}^A (\mu_\star - \mu_a)\mathbb{E}_{\boldsymbol{\nu}}[N_a(T)].$$

Under different assumptions on the distributions $\nu_1, \dots, \nu_A$, one can propose algorithms with sub-linear regret, i.e. such that $\mathcal{R}_{\boldsymbol{\nu}}(\mathcal{A}, T) = o(T)$. In particular under such strategies, the ratio $N_a(T)/T$ goes to zero for sub-optimal arms a, i.e. arms such that $\mu_a < \mu_\star$. This is very different from the behavior of good sampling strategies for pure exploration problems, that have no incentive to maximize rewards and for which $N_a(T)/T$ may converge to a constant for all arms, as will be seen in part II of this document.

Strategies for pure exploration not only consist of a sampling strategy $(A_t)_{t\in\mathbb{N}}$, but also of a *stopping rule* τ, which is a stopping time with respect to the filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$ that indicates when enough exploration has been performed, and a *recommendation rule* $\hat{a}_\tau$, which is $\mathcal{F}_\tau$-measurable and provides an answer to the pure exploration problem. For the best arm identification, $\hat{a}_\tau$ is a guess for the optimal arm $a_\star$, but we will see more general questions to answer in part II of this document. In this work, we will mainly consider the fixed-confidence setting, in which the goal is to guarantee that $\mathbb{P}_{\boldsymbol{\nu}}(\hat{a}_\tau = a_\star) \geq 1 - \delta$ (for best arm identification) while minimizing the number of samples τ needed to make this recommendation, called the *sample complexity*. Other possible objectives, such as the fixed-budget setting in which τ is set to a known budget n, will be discussed in Chapter 3.

Assumptions on the arm distributions Different bandit algorithms have been designed under different assumptions on the arm distributions. The seminal work of Lai and Robbins (1985) considered

parametric distributions that are continuously parameterized by their means. Under this assumption, the authors derived an asymptotic lower bound on the regret of any uniformly efficient algorithm [1] which served as a guideline for designing so-called *asymptotically optimal* algorithm, whose regret is matching the lower bound. An asymptotically optimal algorithm called kl-UCB has been designed assuming further that the arms distributions belong to a one-parameter canonical exponential family. This algorithm can be traced back to the work of Lai (1987), and a finite-time upper bound on its regret was given by Cappé et al. (2013).

A one-parameter canonical exponential family is a set $\mathcal{P}$ of probability distributions, indexed by a real parameter θ called the natural parameter, that is defined by

$$\mathcal{P} = \{\nu_\theta, \theta \in \Theta : \nu_\theta \text{ has a density } f_\theta(x) = \exp(\theta x - b(\theta)) \text{ w.r.t. } \xi\},$$

where $\Theta = (\theta^-, \theta^+) \subseteq \mathbb{R}$ is an open interval, b a twice-differentiable and convex function (called the log-partition function) and ξ a reference measure. Examples of such distributions include Bernoulli distributions, Gaussian distributions with known variance, Poisson distributions, or Gamma distributions with known shape parameter. If $X \sim \nu_\theta$, it can be shown that $\mathbb{E}[X] = \dot{b}(\theta)$ and $\mathrm{Var}[X] = \ddot{b}(\theta) > 0$, where $\dot{b}$ (resp. $\ddot{b}$) is the derivative (resp. second derivative) of b with respect to the natural parameter θ. Thus there is a one-to-one mapping between the natural parameter θ and the mean $\mu = \dot{b}(\theta)$, and distributions in an exponential family can alternatively be parameterized by their mean. Letting $\mathcal{I} = \dot{b}(\Theta)$, for $\mu \in \mathcal{I}$ we denote by ν^μ the unique distribution in $\mathcal{P}$ that has mean $\mu : \nu^\mu = \nu_{\dot{b}^{-1}(\mu)}$. We introduce the following notation for the Kullback-Leibler divergence between two distributions in the same exponential family, for which a closed-form featuring the log-partition function can be given:

$$\begin{aligned} d(\mu, \mu') &= \mathrm{KL}\left(\nu^\mu, \nu^{\mu'}\right) = \mathbb{E}_{X\sim\nu^\mu}\left[\log \frac{f_{\dot{b}^{-1}(\mu)}(X)}{f_{\dot{b}^{-1}(\mu')}(X)}\right] \\ &= \mu \times \left(\dot{b}^{-1}(\mu) - \dot{b}^{-1}(\mu')\right) - b\left(\dot{b}^{-1}(\mu)\right) + b\left(\dot{b}^{-1}(\mu')\right) . \end{aligned} \tag{1}$$

For the particular case in which the exponential family is the set of Bernoulli distributions, we will use the special notation $\mathrm{kl}(\mu, \mu')$ for the binary relative entropy:

$$\mathrm{kl}(\mu, \mu') = \mathrm{KL}\left(\mathcal{B}(\mu), \mathcal{B}(\mu')\right) = \mu \log\left(\frac{\mu}{\mu'}\right) + (1-\mu)\log\left(\frac{1-\mu}{1-\mu'}\right) .$$

We will refer to an *exponential family bandit model* as a bandit model for which all distributions belong to an exponential family $\mathcal{P}$, of the form $\boldsymbol{\nu} = (\nu^{\mu_1}, \nu^{\mu_2}, \ldots, \nu^{\mu_A})$. Such a bandit model can be parameterized by the vector of means $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_A) \in \mathcal{I}^A$ and we will use the notation $\mathbb{P}_{\boldsymbol{\mu}}$ and $\mathbb{E}_{\boldsymbol{\mu}}$ for the probability and expectation under such a bandit model. In this document we will mostly present lower bounds and algorithms for exponential family bandits, that are parameterized by their mean vector $\boldsymbol{\mu} \in \mathcal{I}^A$. However, the lower bounds will typically apply to the more general class of distributions parameterized by their means, and many algorithms can be easily extended to some non-parametric classes of distributions such as bounded distribution or sub-Gaussian distributions.

The assumption of bounded (or sub-Gaussian) distributions is indeed very common in the bandit literature, for example Auer et al. (2002a) provide a simple finite-time upper bound on the regret of the UCB1 algorithm under these assumptions. Asymptotic optimality for bounded distributions is a more complex notion, supported by a lower bound given by Burnetas and Katehakis (1996) (that also applies to more general distributions). Algorithms matching this lower bound for bounded distributions include a non-parametric version of KL-UCB (Cappé et al., 2013) or Thompson Sampling (Riou and Honda, 2020). But bandit algorithms have been analyzed under other assumptions, such as Gaussian with unknown mean and variance (Cowan et al., 2017) or heavy-tailed distributions (Bubeck et al., 2013).

1. A uniformly efficient algorithm has a regret in $o(T^\alpha)$ for all $\alpha \in (0, 1]$ and for every bandit model in a given class.

Tools for Lower Bounds A common feature of most of the work presented in this document is that we strive to design *optimal* solutions for the considered bandit problem. Regret or sample complexity lower bound are essential to assess this optimality and both can be obtained by lower bounding the number of selections of some arms. For the sake of clarity we present the lower bound methodology for exponential family bandit models, parameterized by their vector of means $\boldsymbol{\mu}$, for which it will be used in the sequel.

To the best of my knowledge, all the lower bounds obtained in the bandit literature follow from a change of distribution argument. The idea is to find an alternative bandit model $\boldsymbol{\lambda}$ close enough to $\boldsymbol{\mu}$ but under which the algorithm is supposed to have a totally different behavior (for example due to a different optimal arm), which will give constraints on the number of selections of certain arms in the initial model. The most classical expression of a change of distribution between two bandit models $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ is, for an event $\mathcal{E} \in \mathcal{F}_t$,

$$\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}) = \mathbb{E}_{\boldsymbol{\lambda}}\left[\mathbb{1}(\mathcal{E}) \exp\left(L_t(\boldsymbol{\mu}, \boldsymbol{\lambda})\right)\right]$$

where $L_t(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \log \frac{\ell(X_1,\dots,X_t;\boldsymbol{\mu})}{\ell(X_1,\dots,X_t;\boldsymbol{\lambda})} = \sum_{s=1}^{t} \log \frac{f_{\dot{b}^{-1}(\mu_{A_s})}(X_s)}{f_{\dot{b}^{-1}(\lambda_{A_s})}(X_s)}$ is the log-likelihood ratio of the observation. This elementary change of distribution has been a key ingredient for the derivation of lower bounds in the bandit literature. The most notable example is the famous regret lower bound of Lai and Robbins (1985), and further examples notably include Audibert et al. (2010). In particular, it follows from the Lai and Robbins' lower bound that a uniformly efficient algorithm $\mathcal{A}$ for a given exponential family bandit models satisfies, for all sub-optimal arm a,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]}{\log(T)} \geq \frac{1}{d(\mu_a, \mu_\star)} \,,$$

which yields a lower bound on its regret as $\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) = \sum_{a:\mu_a<\mu_\star} (\mu_\star - \mu_a)\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]$.

More recently, several works came up with a simpler expression for the change of distribution, that directly relates the expected log-likelihood (that can be easily related to the number of selections of each arm) to the probabilities $\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E})$ and $\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E})$, and leads to shorter lower bound proofs. For example Combes and Proutière (2014b) and Kaufmann et al. (2016) independently proposed such a change-of-distribution lemma. A more elegant information-theoretic proof and expression was given by Garivier et al. (2019b) and is presented below. Denoting by

$$I_t = (U_0, A_1, X_1, U_1, \dots, A_t, X_t, U_t)$$

the information available in round t (so that A_t is a deterministic function of I_{t-1}), the data-processing inequality permits to lower bound the Kullback-Leibler divergence of the distribution of I_t under two different bandit models parameterized by $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$.

Lemma 0.1. *Let $\boldsymbol{\mu} = (\mu_1, \dots, \mu_A)$ and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_A)$ be two bandit models. Let τ be a stopping time w.r.t. $(\mathcal{F}_t)_{t\in\mathbb{N}}$ where $\mathcal{F}_t = \sigma(I_t)$, that is almost surely finite. For every event $\mathcal{E} \in \mathcal{F}_\tau$,*

$$\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{I_\tau}, \mathbb{P}_{\boldsymbol{\lambda}}^{I_\tau}\right) \geq \mathrm{kl}\left(\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}), \mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E})\right).$$

Proof. The data-processing inequality states that, if P and Q are some probability distributions on the same measurable space $\mathcal{X}$ and if $f : \mathcal{X} \to \mathcal{Y}$ is a measurable function, then $\mathrm{KL}(P, Q) \geq \mathrm{KL}(P^f, Q^f)$, where P^f (resp. Q^f) denotes the push-forward measure of P (resp. Q) by f. Using this, we get

$$\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{I_\tau}, \mathbb{P}_{\boldsymbol{\lambda}}^{I_\tau}\right) \geq \mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{\mathbb{1}_\mathcal{E}}, \mathbb{P}_{\boldsymbol{\lambda}}^{\mathbb{1}_\mathcal{E}}\right) = \mathrm{kl}\big(\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}), \mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E})\big)\,.$$

□

By definition of the Kullback-Leibler divergence, observe that $\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{I_\tau}, \mathbb{P}_{\boldsymbol{\lambda}}^{I_\tau}\right) = \mathbb{E}_{\boldsymbol{\mu}}\left[L_\tau(\boldsymbol{\mu}, \boldsymbol{\lambda})\right]$ and by Wald's inequality, if τ is almost surely finite, we have $\mathbb{E}_{\boldsymbol{\mu}}\left[L_\tau(\boldsymbol{\mu}, \boldsymbol{\lambda})\right] = \sum_{a=1}^{A} \mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau)] d(\mu_a, \lambda_a)$. By choosing the event $\mathcal{E}$ appropriately, one can easily control the right hand side of the inequality in Lemma 0.1. The following corollary gives an example of such a choice and will be useful to prove lower bounds for regret minimization in different variants of the classical bandit model.

Lemma 0.2. *Let $\mathcal{A}$ be an algorithm that is uniformly efficient on a class of bandit models $\mathcal{M}$, i.e. for all $\boldsymbol{\mu} \in \mathcal{M}$, for any sub-optimal arm a in $\boldsymbol{\mu}$, $\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] = o(T^\alpha)$ for all $\alpha \in (0,1)$. Then for all bandit models $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ that do not have an optimal arm in common,*

$$\lim_{T\to\infty} \frac{\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{I_T}, \mathbb{P}_{\boldsymbol{\lambda}}^{I_T}\right)}{\log(T)} \geq 1.$$

Proof. Let $\mathcal{A}_{\boldsymbol{\mu}}$ and $\mathcal{A}_{\boldsymbol{\lambda}}$ be the set of optimal arms in the bandit models parameterized by $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$. We assume that $\mathcal{A}_{\boldsymbol{\mu}} \cap \mathcal{A}_{\boldsymbol{\lambda}} = \varnothing$. For $T \in \mathbb{N}^*$, we let $\mathcal{E}_T$ be the event

$$\mathcal{E}_T = \left\{ \sum_{a\in\mathcal{A}_{\boldsymbol{\mu}}} N_a(T) \leq T/2 \right\}.$$

The event $\mathcal{E}_T$ belongs to $\mathcal{F}_T$ and has intuitively a small probability under $\boldsymbol{\mu}$ in which the optimal arms in $\mathcal{A}_{\boldsymbol{\mu}}$ should be selected a lot, and a large probability under $\boldsymbol{\lambda}$ in which $\mathcal{A}_{\boldsymbol{\mu}}$ only contains sub-optimal arms. Using Markov's inequality, we can make this more formal:

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\mu}}\left(\mathcal{E}_T\right) &= \mathbb{P}_{\boldsymbol{\mu}}\left(\sum_{a\notin\mathcal{A}_{\boldsymbol{\mu}}} N_a(T) > T/2\right) \leq \frac{2\sum_{a\notin\mathcal{A}_{\boldsymbol{\mu}}} \mathbb{E}_{\boldsymbol{\mu}}\left[N_a(T)\right]}{T}, \\
\mathbb{P}_{\boldsymbol{\lambda}}\left(\overline{\mathcal{E}_T}\right) &= \mathbb{P}_{\boldsymbol{\lambda}}\left(\sum_{a\in\mathcal{A}_{\boldsymbol{\mu}}} N_a(T) > T/2\right) \leq \frac{2\sum_{a\in\mathcal{A}_{\boldsymbol{\mu}}} \mathbb{E}_{\boldsymbol{\lambda}}\left[N_a(T)\right]}{T}.
\end{aligned}$$

Letting $u_T = 2\sum_{a\notin\mathcal{A}_{\boldsymbol{\mu}}} \mathbb{E}_{\boldsymbol{\mu}}\left[N_a(T)\right]$ and $v_T = 2\sum_{a\in\mathcal{A}_{\boldsymbol{\mu}}} \mathbb{E}_{\boldsymbol{\lambda}}\left[N_a(T)\right]$ by the assumption made on the algorithm, we know that $u_T, v_T = o(T^\alpha)$ for all $\alpha \in (0,1]$.

Using a lower bound on the binary relative entropy given in Garivier et al. (2019b), namely $\mathrm{kl}(p,q) \geq (1-p)\log(1/(1-q)) - \log(2)$, one can write

$$\mathrm{kl}\left(\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_T), \mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_T)\right) \geq \left(1 - \frac{u_T}{T}\right)\log\left(\frac{T}{v_T}\right) - \log(2).$$

Using the above properties of u_T and v_T yields $\lim_{T\to\infty}\left(1 - \frac{u_T}{T}\right) = 1$ and $\log(T/v_T) \sim \log(T)$ which concludes the proof together with the inequality in Lemma 0.1.

□

The change of distribution lemmas introduced in this paragraph will be useful in several places in this document. First, Lemma 0.2 will be used in part I to prove variants of the Lai and Robbins' lower bound for more complex rewards maximization problems. Then, Lemma 0.1 will be useful in part II to prove sample complexity lower bounds for active identification problems. We will also see that lower bounds are not only useful for checking the (asymptotic) optimality of an algorithm, but that they can also guide the design of algorithms. We will indeed see several examples of *lower bound inspired* algorithms.

Tools for algorithms: concentration and posterior distributions Different exploration mechanisms exist for solving the exploration/exploitation trade-off inherent to rewards maximization, or for pure exploration. A common feature of these mechanisms is that they do not only rely on point estimates of the unknown means. For example, the greedy strategy that always selects the arm with largest empirical mean is known to have a linear regret.

A first approach consists in leveraging *confidence intervals* on the unknown means of the arm. For reward maximization, the celebrated optimism in face of uncertainty principle recommends to pick the arm that can lead to the largest possible pay-off which corresponds to the arm with highest Upper Confidence Bound (UCB) (Agrawal, 1995; Auer et al., 2002a). For best arm identification, algorithms leveraging both upper and lower confidence bounds have been proposed, like LUCB (Kalyanakrishnan et al., 2012).

To build good confidence intervals, tight concentration inequalities are needed, which depend on the assumptions made on the arm distributions. For exponential bandit models, the Chernoff inequality (which follows from applying the Crámer-Chernoff method) takes the following form, which features the KL-divergence function $d(\cdot,\cdot)$. For every arm $a \in [A]$,

$$\text{for } x > \mu_a,\ \mathbb{P}\left(\hat{\mu}_{a,s} > x\right) \leq e^{-sd(\mu_a,x)} \quad \text{and} \quad \text{for } x < \mu_a,\ \mathbb{P}\left(\hat{\mu}_{a,s} < x\right) \leq e^{-sd(\mu_a,x)}\ ,$$

where $\hat{\mu}_{a,s} = \frac{1}{s}\sum_{i=1}^{s} Y_{a,s}$ is the empirical mean of arm a based on the first s observation from this arm. Exploiting the monotonicity properties of the divergence d, one can also write, for all $u > 0$,

$$\mathbb{P}\left(sd^{+}\left(\hat{\mu}_{a,s},\mu_a\right) > u\right) \leq e^{-u} \quad \text{and} \quad \mathbb{P}\left(sd^{-}\left(\hat{\mu}_{a,s},\mu_a\right) > u\right) \leq e^{-u}\ ,$$

where $d^{+}(x,y) = d(x,y)\mathbb{1}_{(x\leq y)}$ and $d^{-}(x,y) = d(x,y)\mathbb{1}_{(x\geq y)}$. Hence we see that in an exponential family, deviations are naturally measured with respect to the associated divergence d, which is not a standard quadratic distance (except in the case of Gaussian distribution with known variance σ^2 for which $d(x,y) = \frac{(x-y)^2}{2\sigma^2}$).

One of the tricky aspects of the analysis of a bandit algorithm is that the number of observations from each arm at a current stage of the algorithm, $N_a(t)$, is itself a random variable, so one cannot simply replace s by $N_a(t)$ in the above inequalities. To circumvent this issue, one need to resort to self-normalized inequalities, such that the one below, in which we denote by $\hat{\mu}_a(t) = \hat{\mu}_{a,N_a(t)}$ the empirical mean of the observation collected from arm a after t rounds.

Lemma 0.3 (Garivier and Cappé (2011))**.** *For every exponential family bandit model with divergence function d, for every arm $a \in [A]$, for $u > 0$,*

$$\mathbb{P}\Big(N_a(t)d^{+}\left(\hat{\mu}_a(t),\mu_a\right) > u\Big) \leq \mathbb{P}\Big(\exists s \in [t] : sd^{+}\left(\hat{\mu}_{a,s},\mu_a\right) > u\Big) \leq e\lceil u\log(t)\rceil \exp(-u)\ .$$

This inequality permits to calibrate the upper confidence bound used by the kl-UCB algorithm (Cappé et al., 2013), which is

$$\mathrm{UCB}_a(t) = \max\left\{q : N_a(t)d(\hat{\mu}_a(t),q) \leq \log(t) + c\log\log(t)\right\}$$

and satisfies $\mathbb{P}\left(\mu_a \leq \mathrm{UCB}_a(t)\right) \gtrsim 1 - \left(t\log(t)^{c-2}\right)^{-1}$. KL-based confidence intervals will be used in several places in this document. We will also present some new concentration tools that allow to measure deviations for multiple arms at the same time, and can further provide deviation that are uniform in time for every $t \in \mathbb{N}$. A part of Chapter 3 will be dedicated to the presentation of this new inequality (Theorem 3.1) and its proof.

A popular alternative to the use of confidence intervals is the use of Bayesian methods, notably *Thompson Sampling*, also known as posterior sampling. Thompson Sampling was proposed as the very

first bandit algorithm strategy for two-armed Bernoulli bandit (Thompson, 1933) and has recently gained a lot of popularity for rewards maximization in complex bandit models (Agrawal and Goyal, 2013b; Riquelme et al., 2018) or reinforcement learning (Osband et al., 2013), see Russo et al. (2018) for a tutorial. On the theoretical side, Thompson Sampling is known to be asymptotically optimal in exponential family bandit models (Korda et al., 2013).

Given a prior distribution on the means $\boldsymbol{\mu}$, the algorithm maintains a posterior distribution Π_t on $\boldsymbol{\mu}$, defined as the conditional distribution of the vector of means (seen as a random vector drawn from the prior) given the observation made in the first t rounds. Thompson Sampling is a randomized algorithm which selects each arm according to its posterior probability of being optimal. Instead of computing the posterior probability of each arm to be the best, which can be numerically costly, a simple implementation consists in drawing a possible mean vector $\theta(t)$ from Π_t and picking the arm with largest entry $\theta_a(t)$. In an exponential family bandit model, a common choice of prior distribution consists in using independent conjugate prior for each μ_a. For example, in Bernoulli bandits with a uniform prior on each mean, the posterior distribution on μ_a takes the simple form

$$\pi_a(t) = \text{Beta}\left(S_a(t) + 1, N_a(t) - S_a(t) + 1\right)$$

where $S_a(t) = \sum_{s=1}^{t} X_{a,t}\mathbb{1}\left(A_s = a\right)$ is the sum of ones observed from arm a and $N_a(t) - S_a(t)$ is the sum of zeros.

Thompson Sampling will appear in both parts of this document. In part I, we will see that some variants of this algorithm can be proposed for more sophisticated rewards maximization problems, and remain asymptotically optimal. In part II, we will present two adaptations of Thompson Sampling to two different active identification problems that also achieve some notion of optimality.

3 List of Associated Publications

This document highlights a selection of contributions that have been published in journals or mostly in conference proceedings, that are very important in the machine learning community. A list of the published work posterior to my PhD defense and not related to the PhD work is given below. I highlight in bold the student with whom I have collaborated. This includes the PhD students I co-supervise(d):

- Lilian Besson, with Christophe Moy (Université de Rennes), 2016-2019
- Xuedong Shang, with Michal Valko (Inria Lille, now Deepmind), 2017-
- Omar Darwiche Domingues, with Michal Valko (Inria Lille, now Deepmind), 2018-
- Clémence Réda, with Andrée Delahaye-Duriez (INSERM, Paris), 2019-
- Dorian Baudry, with Odalric-Ambrym Maillard (Inria Lille), 2019-

but also one master student (Cindy Trinh), two PhD students who did research visits in SequeL (Maryam Aziz and Rianne de Heide) and other collaborators.

Journal papers (5)

- *Machine learning applications in drug development.* **Clémence Réda**, Emilie Kaufmann, Andrée Delahaye-Duriez. Computational and Structural Biotechnology Journal 18: 241-252, 2020.
- *Asymptotically Optimal Algorithms for Budgeted Multiple Play Bandits.* Alexander Luedtke, Emilie Kaufmann and Antoine Chambaz. Machine Learning 108(11): 1919-1949, 2019.
- *A Spectral Algorithm with Additive Clustering for the Recovery of Overlapping Communities in Networks*. Emilie Kaufmann, Thomas Bonald and Marc Lelarge.
  Theoretical Computer Science, Vol 742: 3-26, 2018.
- *Learning the distribution with largest mean: two bandit frameworks.* Emilie Kaufmann and Aurélien Garivier. ESAIM: Proceedings and Surveys, Vol 60:114-131, 2017.

- *On Bayesian Index Policies for Sequential Resource Allocation*. Emilie Kaufmann. Annals of Statistics, Vol 46(2): 842-865, 2017.

Paper in international conferences with proceedings (17)

- *Sub-sampling for Efficient Non-Parametric Bandit Exploration*. **Dorian Baudry**, Emilie Kaufmann, Odalric-Ambrym Maillard. Advances in Neural Processing Systems (NeurIPS), 2020.
- *Planning in Markov Decision Processes with Gap-Dependent Sample Complexity*. Anders Jonsson, Emilie Kaufmann, Pierre Ménard, **Omar Darwiche Domingues**, **Edouard Leurent** and Michal Valko. Advances in Neural Processing Systems (NeurIPS), 2020.
- *Fixed Confidence Guarantees for Bayesian Best Arm Identification*. **Xuedong Shang**, **Rianne de Heide**, Emilie Kaufmann, Pierre Ménard and Michal Valko. International Conference on Artificial Intelligence and Statistics (AISTATS), 2020.
- *A Practical Algorithm for Multiplayer Bandits when Arm Means Vary Among Players*. **Etienne Boursier**, Emilie Kaufmann, Abbas Mehrabian and Vianney Perchet. International Conference on Artificial Intelligence and Statistics (AISTATS), 2020.
- *Solving Bernoulli Rank-One Bandits with Unimodal Thompson Sampling.* **Cindy Trinh**, Emilie Kaufmann, Claire Vernade and Richard Combes. International Conference on Algorithmic Learning Theory (ALT), 2020.
- *General Parallel Optimization without Metric*. Xuedong Shang, Emilie Kaufmann and Michal Valko. International Conference on Algorithmic Learning Theory (ALT), 2019.
- *Sequential Test for the Lowest Mean: From Thompson to Murphy Sampling*. Emilie Kaufmann, Wouter Koolen and Aurélien Garivier. Advances in Neural Information Processing Systems (NeurIPS), 2018.
- *Multi-Player Bandits Revisited*. **Lilian Besson** and Emilie Kaufmann. International Conference on Algorithmic Learning Theory (ALT), 2018.
- *Corrupt Bandits for Preserving Local Privacy*. **Pratik Gajane**, Tanguy Urvoy and Emilie Kaufmann. International Conference on Algorithmic Learning Theory (ALT), 2018.
- *Pure Exploration in Infinite Bandit Models with Fixed Confidence*. **Maryam Aziz**, **Jesse Anderton**, Emilie Kaufmann and Javed Aslam. International Conference on Algorithmic Learning Theory (ALT), 2018.
- *Aggregation of Multi-Armed Bandits learning algorithms for Opportunistic Spectrum Access*. **Lilian Besson**, Emilie Kaufmann, Christophe Moy. IEEE Wireless Communications and Networking Conference (WCNC), 2018.
- *Monte-Carlo Tree Search by Best Arm Identification*. Emilie Kaufmann and Wouter M. Koolen. Advances in Neural Processing Systems (NeurIPS), 2017.
- *Multi-Armed Bandit Learning in IoT Networks: Learning helps even in non-stationary settings*. **Rémi Bonnefoi**, **Lilian Besson**, Christophe Moy, Emilie Kaufmann and Jacques Palicot. International Conference on Cognitive Radio Oriented Wireless Networks (CROWNCOM), 2017.
- *On Explore-Then-Commit Strategies*. Aurélien Garivier, Emilie Kaufmann and Tor Lattimore. Advances in Neural Processing Systems (NeurIPS), 2016.
- *A Spectral Algorithm with Additive Clustering for the Recovery of Overlapping Communities in Networks*. Emilie Kaufmann, Thomas Bonald and Marc Lelarge. International Conference on Algorithmic Learning Theory (ALT), 2016.
- *Maximin Action Identification: a New Bandit Framework for Games*. Aurélien Garivier, Emilie Kaufmann and Wouter M. Koolen. 29th Conference on Learning Theory (COLT), 2016.
- *Optimal Best-Arm Identification with Fixed Confidence*. Aurélien Garivier and Emilie Kaufmann. 29th Conference on Learning Theory (COLT), 2016.

Papers in international workshops (5)

- *Adaptive Reward-Free Exploration*. Emilie Kaufmann, Pierre Ménard, **Omar Darwiche Domingues**, Anders Jonsson, **Edouard Leurent** and Michal Valko. Theoretical Foundations of Reinforcement Learning Workshop @ ICML, 2020.
- *A Kernel-Based Approach to Non-Stationary Reinforcement Learning in Metric Spaces*. **Omar Darwiche Domingues**, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann and Michal Valko. Theoretical Foundations of Reinforcement Learning Workshop @ ICML, 2020.
- *A simple dynamic bandit algorithm for hyper-parameter tuning*. **Xuedong Shang**, Emilie Kaufmann and Michal Valko. AutoML workshop @ ICML, 2019.
- *Adaptive Black-Box Optimization Got Easier: HCT Only Needs Local Smoothness*. **Xuedong Shang**, Emilie Kaufmann and Michal Valko. European Workshop on Reinforcement Learning (EWRL), 2018.
- *Corrupt Bandits*. **Pratik Gajane**, Tanguy Urvoy and Emilie Kaufmann. European Workshop on Reinforcement Learning (EWRL), 2016.

Papers in French conferences or journals (2)

- *Analyse non asymptotique d'un test séquentiel de détection de ruptures et application aux bandits non stationnaires*. **Lilian Besson** and Emilie Kaufmann. GRETSI, 2019.
- *Modèles de bandit : une histoire bayésienne et fréquentiste*. Emilie Kaufmann. MATAPLI 109:51-64, 2016.

Submitted papers (7)

- *Non-Asymptotic Sequential Tests for Overlapping Hypotheses and application to near optimal arm identification in bandit models*. Aurélien Garivier and Emilie Kaufmann. Accepted for publication in Sequential Analysis, to appear in 2021. *arXiv:1905.03495*
- *On Multi-Armed Bandit Designs for Dose-Finding Trials*. Maryam Aziz, Emilie Kaufmann and Marie-Karelle Rivière. Accepted for publication in the Journal of Machine Learning Research, to appear in 2021. *arXiv:1903.07082*
- *Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals*. Emilie Kaufmann and Wouter M. Koolen. Under revision for the Journal of Machine Learning Research, 2018. *arXiv:1811.11419*
- *A Kernel-Based Approach to Non-Stationary Reinforcement Learning in Metric Spaces*. **Omar Darwiche Domingues**, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann and Michal Valko. Submitted to AISTATS 2021. *arXiv:2007.05078*
- *Regret Bounds for Kernel-Based Reinforcement Learning*. **Omar Darwiche Domingues**, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann and Michal Valko. Submitted to ALT 2021. *arXiv:2004.05599*
- *Episodic Reinforcement Learning in Finite MDPs: Minimax Lower Bounds Revisited* . **Omar Darwiche Domingues**, Pierre Ménard, Emilie Kaufmann and Michal Valko. Submitted to ALT 2021. *arXiv:2010.03531*
- *Adaptive Reward-Free Exploration*. Emilie Kaufmann, Pierre Ménard, **Omar Darwiche Domingues**, Anders Jonsson, **Edouard Leurent** and Michal Valko. Submitted to ALT 2021. *arXiv:2006.06294*

Part I

Maximizing Rewards, with a Twist

Chapter 1

Optimal Solution for Variants of the Classical Multi-Armed Bandit

In this chapter we study some extensions of the standard rewards maximization problem in a multi-armed armed bandit problem, that are motivated by different applications.

On the algorithmic side, we mostly explore how to design variants of Thompson Sampling in two different contexts. We first consider *transductive settings*, in which the feedback observed during the interaction is not the reward that the agent seeks to maximize. Then we present a first step in designing a variant of Thompson Sampling for *structured bandits* with the example of Unimodal Thompson Sampling motivated by rank-one bandits. For these different settings, our goal remains the design of *asymptotically optimal* algorithms and we therefore explain how the lower bound methodology presented in the Introduction can lead to more sophisticated lower bounds with simple proofs.

1.1 Thompson Sampling in Transductive Settings

In the classical bandit setting, a learner tries to maximize her total reward and the feedback she obtained during learning is the observation of the rewards themselves. In this section, we present two variants of this problem that have in common that the feedback observed during learning is only loosely related to the rewards that the learner aims at maximizing.

1.1.1 Thompson Sampling for Corrupt Bandits

Motivated by (local) privacy preserving, we studied with Patrick Gajane, who was then doing a CIFRE PhD at Orange Labs, and his advisor Tanguy Urvoy, the problem of maximizing rewards based on a feedback which is a stochastic transformation of the rewards, when the link between the reward and feedback distributions is known.

More precisely, the setting, that we call "corrupt bandits" is the following. A corrupt bandit model is described by A reward distributions $\{\nu_a\}_{a=1\dots A}$, A feedback distributions $\{\varsigma_a\}_{a=1,\dots,A}$, and a list of *mean-corruption functions* $\{g_a\}_{a=1,\dots,A}$. The reward and feedback distributions are unknown to the learner, while the mean-corruption functions are assumed to be known. At round t, if the learner selects arm $A_t = a$, she receives a reward R_t drawn from the distribution ν_a with mean μ_a and observes a feedback F_t drawn from the distribution ς_a with mean λ_a. We assume that, for each arm, there exists a loose link between the reward and the feedback through the corruption function g_a which maps the mean of the reward distribution to the mean of the feedback distribution : for all $a \in [A]$, $g_a(\mu_a) = \lambda_a$.

The motivation of this setting stems from the use of a privacy-preserving mechanism in a recom-

mender system. To avoid disclosing the tastes of each user (materialized by the click, or rating that plays the role of a reward) to a local observer having access to its database, the system stores on purpose "corrupted" versions of these rewards. Based on this corrupted feedback, the goal is still to propose items which generate large expected rewards. If the rewards are binary, randomized response can be used (Warner, 1965): for each arm a the noisy feedback F is generated from the reward R is such a way that $\mathbb{P}(F = x|R = y) = \mathcal{M}_a(y, x)$, where $\mathcal{M}_a$ is some corruption matrix

$$\mathcal{M}_a = \begin{matrix} \\ 0 \\ 1 \end{matrix} \begin{matrix} \begin{matrix} 0 & \quad 1 \end{matrix} \\ \begin{bmatrix} p_{00}(a) & 1-p_{11}(a) \\ 1-p_{00}(a) & p_{11}(a) \end{bmatrix} \end{matrix} \tag{1.1}$$

For an arm a with expected reward μ_a the expected feedback under this corrupting scheme is $g_a(\mu_a) = 1-p_{00}(a)+[p_{00}(a)+p_{11}(a)-1]\cdot\mu_a$. In the paper Gajane et al. (2018), we proposed two asymptotically optimal algorithms for the corrupt bandit problems, for general corruption functions that are assumed to be continuous and monotone, when the reward and feedback are binary. [1]

A binary corrupt bandit problem with given corruption functions $\{g_a\}_{a=1\in[A]}$ can be parameterized by its vector of means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_A)$: when the learner selects arm a, a Bernoulli reward with mean μ_a and a Bernoulli feedback with mean $\lambda_a^{\boldsymbol{\mu}} = g_a(\mu_a)$ are generated.

Regret and best achievable performance Despite observing only the feedback, the learner still aims at maximizing the sum of her (unobserved) rewards. Hence the regret is measured with respect to the strategy always playing the arm with largest mean as in the classical setting:

$$\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) = \mu_\star T - \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} R_t\right] = \sum_{a\in[A]} \Delta_a^{\boldsymbol{\mu}} \mathbb{E}_{\boldsymbol{\mu}}\left[N_a(T)\right],$$

where $\mu_\star = \max_a \mu_a$ and $\Delta_a^{\boldsymbol{\mu}} = \mu_\star - \mu_a$.

One can extend the notion of uniformly efficient algorithms of Lai and Robbins (1985) to the binary corrupt bandit problem with a given family of corruption functions: an algorithm $\mathcal{A}$ is uniformly efficient if, for every $\boldsymbol{\mu} \in [0,1]^A$, for every $\alpha \in (0,1)$, its regret satisfies $\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) = o(T^\alpha)$. One can use the powerful Lemma 0.2 presented in the Introduction to derive an asymptotic lower bound on the number of selections of each sub-optimal arm. Consider a corrupt bandit instance $\boldsymbol{\mu}$ for which we assume to fix the ideas that arm 1 is the best arm. Let $\boldsymbol{\mu}'$ be any corrupt bandit instance in which arm 1 is sub-optimal. Then Lemma 0.2 yields

$$\lim_{T\to\infty} \frac{\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{I_T}, \mathbb{P}_{\boldsymbol{\mu}'}^{I_T}\right)}{\log(T)} \geq 1.$$

where we recall that I_t is the information available for the learner after t observations. In the corrupted setting, the information contains the feedback, and not the reward:

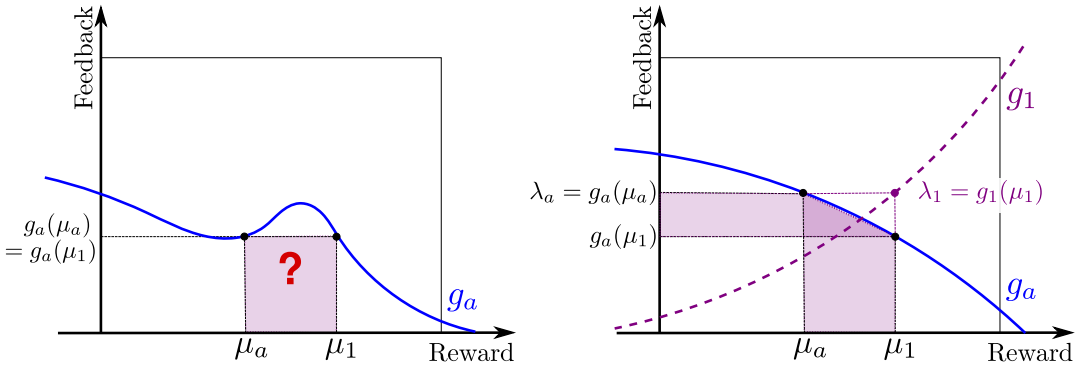
$$I_t = (U_0, A_1, F_1, U_1, \ldots, A_t, F_t, U_t).$$

Hence, one can show that $\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{I_T}, \mathbb{P}_{\boldsymbol{\mu}'}^{I_T}\right) = \sum_{a=1}^{A} \mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]\mathrm{kl}\left(\lambda_a^{\boldsymbol{\mu}}, \lambda_a^{\boldsymbol{\mu}'}\right)$. The final lower bound stated below follows from selecting the alternative instance $\boldsymbol{\mu}'$ defined by

$$\mu'_b = \begin{cases} \mu_1 + \varepsilon, & \text{if } b = a \\ \mu_b & \text{otherwise} \end{cases}$$

for each sub-optimal arm a (for which $\lambda_a^{\boldsymbol{\mu}'} = g_a(\mu_1 + \varepsilon)$) and by letting ε go to zero.

1. The proposed strategies can be adapted to sub-Gaussian distributions or to rewards that belong to another one-dimensional exponential family, but we present the binary case for the sake of simplicity.



(a) Uninformative g_a function (b) Informative g_a function

Figure 1.1 – In Figure 1.1(a), g_a such that $\lambda_a = g_a(\mu_1)$ thereby making it impossible to discern arm a from the optimal arm 1 given the mean feedback. In Figure 1.1(b), a steep monotonic g_a transforms the reward gap $\Delta_a^{\boldsymbol{\mu}} = \mu_1 - \mu_a$ into a clear gap between λ_a and $g_a(\mu_1)$.

Theorem 1.1. *Given continuous corruption functions $\{g_a\}_{a\in A}$, an algorithm $\mathcal{A}$ that is uniformly efficient for the Bernoulli corrupt bandit problem satisfies, for any sub-optimal arm a,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]}{\log(T)} \geq \frac{1}{\mathrm{kl}\left(\lambda_a^{\boldsymbol{\mu}}, g_a(\mu_\star)\right)},$$

where $\mathrm{kl}(x,y) = x\log\left(\frac{x}{y}\right) + (1-x)\log\left(\frac{1-x}{1-y}\right)$ *is the binary relative entropy.*

The lower bound reveals that the divergence between the mean feedback for arm a and the image of the optimal reward $\mu_\star$ with g_a plays a crucial role in distinguishing arm a from the optimal arm. The shape of the g_a function in the neighborhood of both a and $a_\star$ has a great impact on the information that the learner can extract from the received feedback. Particularly, if the g_a function is non-monotonic and $g_a(\mu_\star) = g_a(\mu_a)$, it may be impossible to distinguish between arm a and the optimal arm, arm 1. See Figure 1.1(a) for an illustration. To avoid this problem, we will propose algorithms under the extra assumption that each corruption function g_a is strictly monotonic. The inverse function of g_a is therefore well defined and we denote it by g_a^{-1}. Such an informative corruption function is shown in Figure 1.1(b). To clarify that the gap between λ_a and $\lambda_\star = g_{a_\star}(\mu_\star)$ is not relevant here, we also add in Figure 1.1(b), a corruption function $g_{a_\star}$ which differs from g_a and causes fortuitously the two arms a and $a_\star$ to have the same mean feedback with different interpretations in terms of mean rewards.

The TSCF algorithm We now explain how to adapt an algorithm which is asymptotically optimal in the classical setting, Thompson Sampling, to the corrupt bandit problem. The kl-UCB algorithm (Cappé et al., 2013) can be similarly adapted as explained in the paper (Gajane et al., 2018), but we focus on Thompson Sampling in this document, stated as Algorithm 1 below.

Given a uniform prior distribution over the mean *feedback* of each arm, $\lambda_a^{\boldsymbol{\mu}}$, the algorithm updates a posterior distribution after the observation of a Bernoulli feedback with mean $\lambda_a^{\boldsymbol{\mu}}$ each time arm a is sampled. The posterior distribution of $\lambda_a^{\boldsymbol{\mu}}$ after t rounds of the algorithm is

$$\pi_a(t) = \mathrm{Beta}\left(S_a(t)+1, N_a(t) - S_a(t) + 1\right),$$

where $S_a(t) = \sum_{s=1}^t F_s \mathbb{1}_{(A_s=a)}$ is the sum of feedback observed on arm a and $N_a(t) = \sum_{s=1}^t \mathbb{1}_{(A_s=a)}$ is the number of selections of arm a. Thompson Sampling draws at time $t+1$ one sample $\theta_a(t)$ from each

Algorithm 1 Thompson Sampling for MAB with corrupted feedback (TS-CF)

Input: Horizon T, corruption functions $g_1, \dots, g_K$
Initialization: for each $a \in [A]$, $N_a = 0$, $S_a = 0$
for $t = 1 \dots T$ **do**
  For each arm $a \in [A]$, sample $\theta_a \sim \text{Beta}(S_a + 1, N_a - S_a + 1)$
  Select arm $A_t = \arg\max_a g_a^{-1}(\theta_a)$ and observe feedback F_t
  Update number of visits and sum of feedback:
  $N_{A_t} \leftarrow N_{A_t} + 1$, $S_{A_t} \leftarrow S_{A_t} + F_t$
end for

posterior distribution $\pi_a(t)$. As the mapping between mean feedback and mean reward is known, the algorithm then selects A_{t+1} as the arm for which $g_a^{-1}(\theta_a(t))$ is largest. It can easily be checked that the the probability to select arm a at round t is exactly the posterior probability that arm a has largest mean reward, which is the general philosophy of Thompson Sampling.

We prove the following regret bound for this TS-CF algorithm. The proof follows the analysis of Agrawal and Goyal (2013a) to upper bound $\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]$ for each a such that $\mu_a < \mu_\star$. It requires to distinguish 4 cases depending on whether g_1 and g_a are increasing or decreasing. Some of them require adaptations of the arguments of Agrawal and Goyal (2013a), that are highlighted in the sketch below.

Theorem 1.2. *Let $\boldsymbol{\mu}$ be a Bernoulli corrupt bandit model with continuous and monotone corruption functions. For every $\varepsilon > 0$, there exists a constant $C_\varepsilon = C(\varepsilon, \boldsymbol{\mu}, \{g_a\}_{a\in[A]})$ such that*

$$\mathcal{R}_{\boldsymbol{\mu}}(\text{TS-CF}, T) \le (1+\varepsilon) \sum_{a \in [A]\setminus\{a_\star\}} \frac{\Delta_a^{\boldsymbol{\mu}} \log(T)}{\text{kl}(\lambda_a^{\boldsymbol{\mu}}, g_a(\mu_\star))} + C_\varepsilon.$$

Sketch of proof. To ease the notation, assume arm 1 is optimal and fix some sub-optimal arm a. We define two high-probability events on which the empirical mean $\hat{\lambda}_a(t)$ of the collected feedback and the posterior sample $\theta_a(t)$ do not deviate too much from $\lambda_a^{\boldsymbol{\mu}}$. Introducing two thresholds x and y (to be chosen at the end of the proof) such that $\lambda_a^{\boldsymbol{\mu}} < x < y < g_a(\mu_1)$ (resp. $\lambda_a^{\boldsymbol{\mu}} > x > y > g_a(\mu_1)$) when g_a is increasing (resp. decreasing), we let

$$E_a^\lambda(t) = \left\{g_a^{-1}(\hat{\lambda}_a(t)) \le g_a^{-1}(x)\right\} \quad \text{and} \quad E_a^\theta(t) = \{g_a^{-1}(\theta_a(t)) \le g_a^{-1}(y)\}$$

and upper bound the number of sub-optimal selections as follows:

$$\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] = \sum_{t=0}^{T-1} \mathbb{P}_{\boldsymbol{\mu}}(A_{t+1} = a, E_a^\lambda(t), E_a^\theta(t)) + \sum_{t=0}^{T-1} \mathbb{P}_{\boldsymbol{\mu}}(A_{t+1} = a, E_a^\lambda(t), \overline{E_a^\theta(t)}) + \sum_{t=0}^{T-1} \mathbb{P}_{\boldsymbol{\mu}}(A_{t+1} = a, \overline{E_a^\lambda(t)}). \quad (1.2)$$

Controlling the **third term** in (1.2) amounts to upper bounding $\sum_{s=1}^T \mathbb{P}\left(\hat{\lambda}_{a,s} > x\right)$ for $x > \lambda_a^{\boldsymbol{\mu}}$ (increasing g_a) or $\sum_{s=1}^T \mathbb{P}\left(\hat{\lambda}_{a,s} < x\right)$ for $x < \lambda_a^{\boldsymbol{\mu}}$ (decreasing g_a) where $\hat{\lambda}_{a,s}$ is the empirical mean of the first s observed feedbacks from arm a. In both cases, using the Chernoff inequality permits to prove that

$$\sum_{t=0}^{T-1} \mathbb{P}_{\boldsymbol{\mu}}(A_{t+1} = a, \overline{E_a^\lambda(t)}) \le 1 + \frac{1}{\text{kl}\left(x, \lambda_a^{\boldsymbol{\mu}}\right)}.$$

Controlling the **second term** in (1.2) amounts to upper bounding

$$\sum_{t=0}^{T-1} \mathbb{P}_{\boldsymbol{\mu}}\left(A_{t+1} = a, \hat{\lambda}_a(t) \le x, \theta_a(t) > y\right) \quad \text{or} \quad \sum_{t=0}^{T-1} \mathbb{P}_{\boldsymbol{\mu}}\left(A_{t+1} = a, \hat{\lambda}_a(t) \ge x, \theta_a(t) < y\right)$$

when g_a is increasing or decreasing. While the increasing case is directly handled by Lemma 3 in Agrawal and Goyal (2013a), for the decreasing case, we need to prove a counterpart of this result. Leveraging the Beta-Binomial trick together with the Chernoff inequality, one can prove in both cases that

$$\sum_{t=0}^{T-1} \mathbb{P}_{\boldsymbol{\mu}}(A_{t+1} = a, E_a^{\lambda}(t), \overline{E_a^{\theta}(t)}) \leq \frac{\log(T)}{\mathrm{kl}(x', y)} + 1$$

for any x' in (x, y) and T large enough (see Lemma 4 in Gajane et al. (2018)).

Lemma 1.3 (Beta-binomial trick)**.** *Let $F^{Beta}_{\alpha,\beta}$ denote the cdf of a Beta distribution with parameter α and β and $F^{Bin}_{n,p}$ denote the cdf of a Binomial distribution with parameters n and p. If α and β are integers,*

$$F^{Beta}_{\alpha,\beta}(w) = 1 - F^{Bin}_{\alpha+\beta-1,w}(\alpha - 1)\ .$$

Proving that the **first term** in (1.2) is a constant is the most challenging part of the analysis of Thompson Sampling, even for the classical algorithm. For the corrupt bandit problem, one first generalizes the trick which relates the probability of selecting arm a to that of selecting arm 1:

Lemma 1.4. *Letting $p_{a,t} := \mathbb{P}\left(g_1^{-1}(\theta_1(t)) > g_a^{-1}(y) \middle| \mathcal{F}_t\right)$, it holds that*

$$\mathbb{P}_{\boldsymbol{\mu}}\left(A_{t+1} = a, E_a^{\theta}(t), E_a^{\lambda}(t) \middle| \mathcal{F}_t\right) \leq \frac{(1 - p_{a,t})}{p_{a,t}} \mathbb{P}_{\boldsymbol{\mu}}\left(A_{t+1} = 1, E_a^{\theta}(t), E_a^{\lambda}(t) \middle| \mathcal{F}_t\right)\ .$$

Lemma 1.4 permits to upper bound

$$\sum_{t=0}^{T-1} \mathbb{P}_{\boldsymbol{\mu}}(A_{t+1} = a, E_a^{\lambda}(t), E_a^{\theta}(t)) \leq \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\mu}}\left[\frac{1 - p_{a,t}}{p_{a,t}} \mathbb{1}_{(A_{t+1}=1)}\right] \leq \sum_{s=0}^{T-1} \left(\mathbb{E}_{\boldsymbol{\mu}}\left[\frac{1}{p_{a,\tau_s+1}}\right] - 1\right),$$

where τ_s is the instant of the s-th selection of arm 1. To prove that the first term in (1.2) is a constant, we prove that

$$\mathbb{E}_{\boldsymbol{\mu}}\left[\frac{1}{p_{a,\tau_s+1}}\right] \leq 1 + f(s) \tag{1.3}$$

for some function f that satisfies $\sum_s f(s) < \infty$.

Now one can observe that

- when g_1 is increasing, $p_{a,t} = \mathbb{P}_{\boldsymbol{\mu}}(\theta_1(t) > \tilde{y} | \mathcal{F}_t)$ for some $\tilde{y} < \lambda_1^{\boldsymbol{\mu}}$
- when g_1 is decreasing, $p_{a,t} = \mathbb{P}_{\boldsymbol{\mu}}(\theta_1(t) < \tilde{y} | \mathcal{F}_t)$ for some $\tilde{y} > \lambda_1^{\boldsymbol{\mu}}$

When g_1 is increasing, (1.3) directly follows from the technical Lemma 4 in Agrawal and Goyal (2013a). When g_1 is decreasing, exploiting some symmetries allows to leverage the same arguments to prove (1.3). In the decreasing case, thanks to the Beta-Binomial trick,

$$p_{a,\tau_s+1} = F^{\mathrm{Beta}}_{(S_1(\tau_s)+1, s-S_1(\tau_s)+1)}(\tilde{y}) = 1 - F^{\mathrm{Bin}}_{(s+1,\tilde{y})}(S_1(\tau_s)) \quad \text{and} \quad S_1(\tau_s) \sim \mathrm{Bin}\left(s, \lambda_1^{\boldsymbol{\mu}}\right).$$

Using that $f^{\mathrm{Bin}}_{n,p}(j) = f^{\mathrm{Bin}}_{n,1-p}(n-j)$ and $F^{\mathrm{Bin}}_{n,p}(j) = 1 - F^{\mathrm{Bin}}_{n,1-p}(n-j-1)$ where $F^{\mathrm{Bin}}_{n,p}$ and $f^{\mathrm{Bin}}_{n,p}$ respectively denote the cdf and pdf of a Binomial distribution with parameters n and p, one can write

$$\mathbb{E}_{\boldsymbol{\mu}}\left[\frac{1}{p_{a,\tau_s+1}}\right] = \sum_{j=0}^{s} \frac{f^{\mathrm{Bin}}_{(s,\lambda_1^{\boldsymbol{\mu}})}(j)}{1 - F^{\mathrm{Bin}}_{(s+1,\tilde{y})}(j)} = \sum_{j=0}^{s} \frac{f^{\mathrm{Bin}}_{(s,1-\lambda_1^{\boldsymbol{\mu}})}(s-j)}{F^{\mathrm{Bin}}_{(s+1,1-\tilde{y})}(s-j)} = \sum_{j=0}^{s} \frac{f^{\mathrm{Bin}}_{(s,1-\lambda_1^{\boldsymbol{\mu}})}(j)}{F^{\mathrm{Bin}}_{(s+1,1-\tilde{y})}(j)}\ .$$

In the proof of their Lemma 4, Agrawal and Goyal (2013a) provide an upper bound on the quantity $\sum_{j=0}^{s} f^{\mathrm{Bin}}_{(s,c)}(j)/F^{\mathrm{Bin}}_{(s+1,d)}(j)$ whenever c is larger that d. This bound can be used here as $1 - \lambda_1^{\boldsymbol{\mu}} > 1 - \tilde{y}$ and it permits to prove that (1.3) holds.

Putting things together, if for example g_a is increasing we proved that for every $\lambda_a^{\boldsymbol{\mu}} < x' < y < g_a(\mu_1)$ there exists a constant $C_a(\boldsymbol{\mu}, x', y)$ such that

$$\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq \frac{1}{\mathrm{kl}(x', y)} \log(T) + C_a(\boldsymbol{\mu}, x', y).$$

Choosing x' and y such that $\mathrm{kl}(x', y) = \mathrm{kl}(\lambda_a^{\boldsymbol{\mu}}, g_a(\mu_1))/(1+\varepsilon)$ and summing over sub-optimal arms yields the conclusion.

□

A trade-off between privacy and regret Let us go back to our motivating example of designing a recommender system that learns to display items that the users like using only the stored noisy versions of their preferences. If the stored feedback is obtained with a randomized response mechanism (given by the matrix (1.1)) which is identical for each arm ($p_{00}(a) = p_{00}$ and $p_{11}(a) = p_{11}$), it follows from Theorem 1.2 (and Pinsker's inequality: $\mathrm{kl}(x, y) \geq 2(x-y)^2$) that the regret of TS-CF is of order

$$\underbrace{\frac{1}{|p_{00} + p_{11} - 1|^2}}_{\text{multiplicative factor}} \times \underbrace{\left(\sum_{a \neq a_\star}^{K} \frac{\log(T)}{2\Delta_a^{\boldsymbol{\mu}}} \right)}_{\substack{\text{best possible regret} \\ \text{(sub-Gaussian approx.)}}}.$$

Hence the more secure feedback comes at a multiplicative cost in the regret, this cost being small if both p_{00} and p_{11} are large. It is therefore interesting to find the smallest possible cost which guarantees the randomized response scheme to be *differentially private*.

Differential Privacy (DP) is a notion introduced by Dwork et al. (2006) to quantify how much sharing information about a database can reveal information about individual entries in this database. In our context, a corruption scheme $\tilde{g}$ (generating feedback from rewards) is ε-differentially private if for all rewards sequences $R_1, \dots, R_t$ and $R'_1, \dots, R'_t$ that differ by at most one entry $\mathbb{P}\left(\tilde{g}(R_1, \dots, R_t) \in \mathcal{S}\right) \leq e^{\varepsilon} \cdot \mathbb{P}\left(\tilde{g}(R'_1, \dots, R'_t) \in \mathcal{S}\right)$. Wang et al. (2016) show that randomized response is differentially private provided that $\max(p_{00}/(1-p_{11}), p_{11}/(1-p_{00})) \leq e^{\varepsilon}$. They also show that maximizing $p_{00} + p_{11}$ while maintaining ε-DP requires to chose $p_{00} = p_{11} = e^{\varepsilon}/(1+e^{\varepsilon})$, which yields the smallest multiplicative factor of $(1+e^{\varepsilon})^2/(1-e^{\varepsilon})^2 \simeq 2/\varepsilon^2$.

We note the the notion of ε-DP that we require here is *local*: the entries of the bandit algorithms (feedback) are required to be an ε-DP transformation of the rewards. Other notions of privacy can be defined in the context of bandit algorithms. For example, one may want the selected actions not to leak too much information about the rewards, see, e.g. Tossou and Dimitrakakis (2016).

1.1.2 Thompson Sampling for Dose-Finding

The second "transductive" setting for which I studied Thompson Sampling is motivated by a different field of applications: clinical trials. I've been very curious how to actually apply bandit algorithms for this purpose, and thanks to discussions with Marie-Karelle Rivière, a biostatistician at Sanofi, I discovered the dose-finding problem, which occurs at early stages of clinical trials. I briefly described the work we did on this problem with Maryam Aziz, who did an internship with us in 2016 (Aziz et al., 2018).

In an early stage clinical trial (phase I/II), the goal is to identify a good dosage of a given drug that should be used in further phases of the trial, in which the drug will be compared to other treatments or a placebo. In particular, doctors are interested in finding the Maximum Tolerated Dose (MTD), defined as the dose whose probability to be toxic (i.e. the patient shows severe side effects) is closest to some pre-specified threshold θ, among a set of A candidate doses (typically A varies between 5 to 10). Letting

p_a denote the probability that dose a is toxic, the MTD is formally defined as

$$a_\star \in \underset{a \in [A]}{\operatorname{argmin}} |\theta - p_a|$$

In the context of oncology, efficacy comes at a price of a certain level of toxicity and the threshold θ is typically set to $\theta = 0.3$. Figure 1.2 illustrates the MTD in a situation in which the probability of toxicity is increasing with the dose. This assumption is common for single-drug trials but is not meaningful when combinations of drugs are on trial. Forgetting about possible monotonicity constraints for a while, we now formalize sequential dose-finding as a variant of a multi-armed bandit problem.

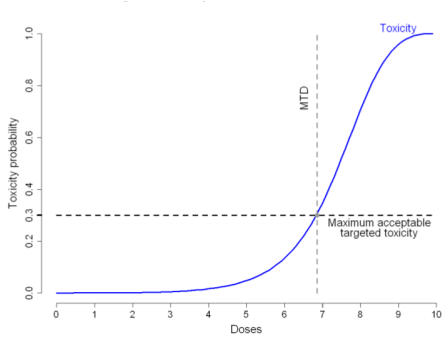


Figure 1.2 – The Maximal Tolerated Dose

In a dose-finding study, a dose $A_t \in [A]$ is selected for the t-th patient, and a binary outcome X_t indicating whether a harmful side-effect occurred is observed. We assume that X_t is drawn from a Bernoulli distribution with mean p_{A_t}, independently from previous observations. The selection rule A_t is sequential and A_t can only depend on the previously allocated doses and the observed outcomes $A_1, X_1, \ldots, A_{t-1}, X_{t-1}$ (and some possible exogenous randomness). A guess $\hat{a}_t$ for the MTD $a_\star$ may also be proposed after the first t patients. The goal of a dose-finding study can be twofold. A first possible objective is to minimize the probability of error $\mathbb{P}(\hat{a}_t \neq a_\star)$, which is a variant of a best arm identification problem, discussed further in part II of this document. Then, if the goal of the trial is therapeutic, another objective is to treat as many patients as possible with the MTD, which is the most efficient tolerable dose. In either case, the feedback observed (the toxicity of the given dose) cannot be viewed as a reward that we seek to maximize.

In the "therapeutic view" on the trial, a possible (unobserved) reward associated to the dose selected in round t could be $R_t = \mathbb{1}(A_t = a_\star)$ and the target behavior of the algorithm is then to maximize $\mathbb{E}_{\boldsymbol{p}}[N_{a_\star}(T)]$ (in a Bernoulli bandit model parameterized by $\boldsymbol{p} = (p_1, \ldots, p_A)$). This goal makes sense in oncology in which patients involved in the trial are in need for a therapeutic effect, hence should ideally be treated at the dose with largest efficacy given acceptable side effects, which is the MTD. This defines a variant of the classical MAB problem where the goal is to maximize the number of selections of some target arm $a_\star$, which is not necessarily the arm with largest mean. We explain below for the MTD identification example how the standard MAB theory can be extended to cover this case.

First, the lower bound methodology presented in the Introduction can be used for the MTD identification problem as well, introducing an appropriate definition of uniform efficiency, directly expressed in terms of the sub-optimal selections.

Theorem 1.5. *We define a uniformly efficient algorithm as an algorithm satisfying for all possible toxicity probabilities* $\boldsymbol{p} = (p_1, \ldots, p_A)$*, for all arm* $a : |\theta - p_a| \neq |\theta - p_{a_\star}|$*,* $\mathbb{E}_{\boldsymbol{p}}[N_a(T)] = o(T^\alpha)$ *for all*

$\alpha = (0, 1]$. *Any uniformly efficient algorithm satisfies, for* $\boldsymbol{p}$ *such that* $p_{a_\star} \neq \theta$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{p}}[N_a(T)]}{\log(T)} \geq \frac{1}{\mathrm{kl}(p_a, d_a^\star)}, \quad \textit{where} \quad d_a^\star := \operatorname*{argmin}_{d\in\{p_{a_\star}, 2\theta - p_{a_\star}\}} |p_a - d|.$$

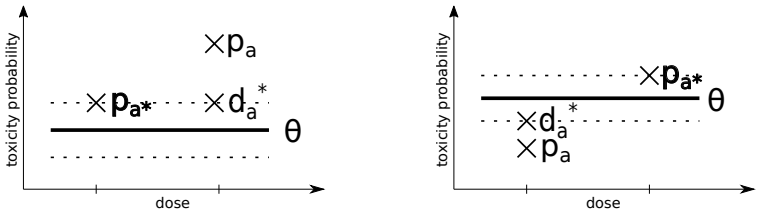


Figure 1.3 – Optimal dose d_a^* defined in Theorem 1.5.
d_a^* is either the toxicity of the MTD or that of its symmetric with respect to θ

Next, we show that Thompson Sampling with independent uniform prior distributions over each probability of toxicity, formally stated in Algorithm 2, is matching the lower bound of Theorem 1.5. Given a sample $\theta_a(t)$ from the posterior distribution on p_a, the arm selected at round $t+1$ is $A_{t+1} = \mathrm{argmin}_a |\theta - \theta_a(t)|$. This randomized algorithm satisfies the Thompson Sampling property: $\mathbb{P}(A_{t+1} = a|\mathcal{F}_t)$ is equal to the posterior probability that arm a is the MTD.

Algorithm 2 Independent Thompson Sampling for MTD Identification

Input: Horizon T, threshold θ.
Initialization: for each $a \in [A]$, $N_a = 0$, $S_a = 0$
for $t = 1 \dots T$ **do**
  For each arm $a \in [A]$, sample $\theta_a \sim \mathrm{Beta}(S_a + 1, N_a - S_a + 1)$
  Select arm $A_t = \arg\min_a |\theta_a - \theta|$ and observe toxic outcome X_t
  Update number of trials and toxic outcomes:
  $N_{A_t} \leftarrow N_{A_t} + 1$, $S_{A_t} \leftarrow S_{A_t} + X_t$
end for

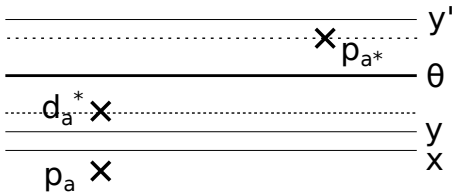
We prove the following upper bound on the number of sub-optimal selections.

Theorem 1.6. *For all* $\varepsilon > 0$, *there exists a constant* $C_{\varepsilon,\theta,\boldsymbol{p}}$ *(depending on* ε, *the threshold* θ *and the vector of toxicity probabilities* $\boldsymbol{p}$*) such that Algorithm 2 satisfies, for all* a *:* $|p_a - \theta| \neq |\theta - p_{a_\star}|$,

$$\mathbb{E}_{\boldsymbol{p}}[N_a(T)] \leq \frac{1+\varepsilon}{\mathrm{kl}(p_a, d_a^*)} \log(T) + C_{\varepsilon,\theta,\boldsymbol{p}}\,.$$

Sketch of proof. We follow the same approach as for the analysis of the TS-CF algorithm in the previous section: we adapt the analysis of Agrawal and Goyal (2013a) to upper bound the number of selections of a sub-optimal arm a. We use the same decomposition as in (1.2) but we modify the definition of the high-probability events, here $E_a^p(t)$ and $E_a^\theta(t)$, to be adapted to the new notion of optimal arm $a_\star$.

To fix the ideas, we consider the case $p_{a_\star} \geq \theta > p_a$. In that case $d_a^\star = 2\theta - p_{a_\star}$ satisfies $p_a < d_a^\star \leq \theta$:



Let $x, y \in]0,1[^2$ be such that $p_a < x < y < d_a^\star$ (chosen at the end of the proof). We let $y' = 2\theta - y > \theta$ be symmetric to y with respect to the threshold (see the illustration above). Letting $\hat{p}_a(t)$ be the empirical mean of the toxicity responses gathered from dose a up to the end of round t and $\theta_a(t)$ be the sample from the Beta posterior on p_a after t rounds, we define $E_a^p(t) = (\hat{p}_a(t) \leq x)$ and $E_a^\theta(t) = (\theta_a(t) \leq y)$.

The proof is then very similar to that of Theorem 1.2 and consists in upper bounding each of the three terms in (1.2) by using concentration inequalities and the Beta-binomial trick. As before, the most delicate part is the first term. Introducing this time $p_{a,t} := \mathbb{P}\Big(\theta_{a_\star}(t) \in [y, y'] \Big| \mathcal{F}_t\Big)$, we prove that

$$\mathbb{P}_{\boldsymbol{p}}\left(A_{t+1} = a \,\middle|\, E_a^\theta(t+1), \mathcal{F}_t\right) \leq \frac{1 - p_{a,t}}{p_{a,t}} \mathbb{P}_{\boldsymbol{p}}\left(A_{t+1} = a_\star \,\middle|\, E_k^\theta(t+1), \mathcal{F}_t\right).$$

As in the previous proof, it remains to upper bound $\sum_{s=0}^{T-1} \mathbb{E}_{\boldsymbol{p}}\left[\frac{1}{p_{a,\tau_s+1}}\right]$ by a constant, where τ_s is the instant of the s-th selection of the MTD $a_\star$. Observe however the different definition of $p_{a,t}$, which is now the posterior probability of p_a being in an interval $[y, y']$ instead of p_a being simply larger or smaller than some threshold. The proof requires therefore a few extra technicalities, that are detailed in Appendix A of Aziz et al. (2018).

□

In the paper Aziz et al. (2018) the theoretical results introduced in this section are presented as a sanity-check for the use of more sophisticated variants of Thompson Sampling for MTD identification. Indeed, both our upper and lower bounds are asymptotic in the number of patients T, whereas for real clinical trials we typically care about small sample sizes. In order to propose practically meaningful designs for MTD identification, we propose the use of Thompson Sampling with a prior enforcing a certain structure on the toxicity probabilities (for example a toxicity increasing in the dose as in Figure 1.2). We show that Thompson Sampling compares favorably with standard baselines for MTD identification, notably the Continuous Reassessment Method (O'Quigley et al., 1990).

Through two examples, we illustrated the flexibility of Thompson Sampling, which can be easily adapted when the optimal arm is not the arm with largest mean feedback. Interestingly, while for corrupt bandits we also proposed an adaptation of the optimistic kl-UCB algorithm (Gajane et al., 2018), defining a UCB-like algorithm that maximizes the number of selections of some arbitrary arm, as needed in the dose finding problem, is not as straightforward. Indeed, Thompson Sampling only requires to be able to define some notion of optimal arm, whereas the optimism principle requires the optimal arm to be the maximizer of some expected payoff.

1.2 Structured Bandits

Going back to the setting in which the learner does observe the actual signal she seeks to maximize, the vanilla bandit problem can be complexified in several other ways. Over the past years, I have been particularly interested to see an increasing number of papers dealing with *structured bandits*. In our parametric setting, the structure is some prior knowledge that the vector of means $\boldsymbol{\mu} = (\mu_1, \dots, \mu_A)$ lies in some *known* subset $\mathcal{S}$ of the set of all possible means. A good algorithm should exploit this knowledge in order to have a regret which is smaller than that of a UCB algorithm agnostic to $\mathcal{S}$.

Here are three interesting examples of structures that have been studied in the literature:

- **Lipschitz bandits** (Magureanu et al., 2014) in which the reward of arm a is a Lipschitz function of some characteristic x_a. Both the $(x_a)_{a\in[A]}$ and the Lipschitz constant L are known in

$$\mathcal{S}_{\text{Lipschitz}} = \left\{\boldsymbol{\mu} = (\mu_{x_1}, \dots, \mu_{x_A}) : \forall (a,b) \in [A]^2,\ |\mu_{x_a} - \mu_{x_b}| \leq L|x_a - x_b|\right\}.$$

- **Unimodal bandits** (Combes and Proutière, 2014a) in which it is known that the mean reward is first increasing, then decreasing:

$$\mathcal{S}_{\text{unimodal}} = \left\{ \boldsymbol{\mu} = (\mu_1, \dots, \mu_A) : \exists a \in [A] : \mu_1 \leq \cdots \leq \mu_a \text{ and } \mu_a \geq \mu_{a+1} \geq \cdots \geq \mu_A \right\}.$$

  This notion was extended to that of *graphical unimodal bandits*, detailed in Section 1.3.
- **Linear bandits** in which $\mu_a = x_a^\top \theta$ for some known context vector $x_a \in \mathbb{R}^d$ describing arm a and some unknown regression parameter $\theta \in \mathbb{R}^d$. Letting $X \in \mathbb{R}^{K \times d}$ the matrix whose row a is $x_a^\top$,

$$\mathcal{S}_{\text{linear}} = \left\{ \boldsymbol{\mu} = (X\theta)^\top \mid \theta \in \mathbb{R}^d \right\}.$$

  Linear bandits have been studied a lot in the multi-armed bandit literature (see, e.g., Abbasi-Yadkori et al. (2011) and references therein) yet Lattimore and Szepesvári (2017) are the first to present an optimal algorithm for linear bandits with a finite number of arms.

Of course, this framework allows to recover **classical bandits** in which there is no structure knowledge, by defining $\mathcal{S}_{\text{classical}} = \mathcal{I}^A$ to be the set of all possible means. Other examples of structures are given by Lattimore and Munos (2014), Kwon et al. (2017) or Jedor et al. (2019).

A regret lower bound for structured bandits A counterpart of the Lai and Robbins lower bound for structured bandits was first given by Agrawal et al. (1989) when the set $\mathcal{S}$ is finite. The case $|\mathcal{S}| = +\infty$ is covered by the more general work of Graves and Lai (1997), which also applies to controlled Markov chains. We now explain how this lower bound for structured bandits can be deduced from Lemma 0.2.

Letting $a_\star(\boldsymbol{\mu})$ denote the set of optimal arms in the bandit model parameterized by $\boldsymbol{\mu}$, we introduce the following sets of alternative (structured) bandit models:

$$\begin{aligned} \text{Alt}_{\mathcal{S}}(\boldsymbol{\mu}) &= \{\boldsymbol{\lambda} \in \mathcal{S} : a_\star(\boldsymbol{\mu}) \cap a_\star(\boldsymbol{\lambda}) = \varnothing\} \\ B_{\mathcal{S}}(\boldsymbol{\mu}) &= \{\boldsymbol{\lambda} \in \mathcal{S} : \forall a \in a_\star(\boldsymbol{\mu}), \mu_a = \lambda_a \text{ and } a_\star(\boldsymbol{\mu}) \cap a_\star(\boldsymbol{\lambda}) = \varnothing\}. \end{aligned}$$

$\text{Alt}_{\mathcal{S}}(\boldsymbol{\mu})$ is the set of bandit instances that have no common optimal arm with the bandit instance $\boldsymbol{\mu}$. $B_{\mathcal{S}}(\boldsymbol{\mu}) \subseteq \text{Alt}_{\mathcal{S}}(\boldsymbol{\mu})$ considers only alternatives in which all arms in $a_\star(\boldsymbol{\mu})$ are unchanged.

The notion of uniformly efficient algorithms is straightforwardly extended to a structure $\mathcal{S}$ as follows: algorithm $\mathcal{A}$ is uniformly efficient if $\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) = o(T^\alpha)$ for all $\boldsymbol{\mu} \in \mathcal{S}$ and all $\alpha \in (0,1)$. This implies in particular that for all $a \notin a_\star(\boldsymbol{\mu})$, $\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] = o(T^\alpha)$ for all $\alpha \in (0,1)$. Recalling that $d(\mu, \mu')$ denotes the Kullback-Leibler divergence between the distribution of mean μ and that of mean μ' in a context of arms parameterized by their means (e.g., exponential family bandit models), Lemma 0.2 yields

Lemma 1.7. *Let $\mathcal{A}$ be a uniformly efficient bandit algorithm for a structure $\mathcal{S}$. Then,*

$$\forall \boldsymbol{\lambda} \in \text{Alt}_{\mathcal{S}}(\boldsymbol{\mu}), \quad \liminf_{T \to \infty} \frac{\sum_{a=1}^A d(\mu_a, \lambda_a) \mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]}{\log(T)} \geq 1.$$

This result is similar to the first statement in Theorem 1 of Graves and Lai (1997) (Equation (2.16)), the difference being that the set $\text{Alt}_{\mathcal{S}}(\boldsymbol{\mu})$ in Lemma 1.7 is replaced by the smaller set $B_{\mathcal{S}}(\boldsymbol{\mu})$. Considering alternative instances in the set $B_{\mathcal{S}}(\boldsymbol{\mu})$ is indeed sufficient to obtain the Graves and Lai lower bound, stated below as Theorem 1.8. However for some particular structures (such as linear bandits) considering only alternatives in $B_{\mathcal{S}}(\boldsymbol{\mu})$, that cannot change the marginal distributions of the optimal arms, may be restrictive, and exploiting Lemma 1.7 might lead to tighter lower bounds.

Theorem 1.8 (Graves and Lai lower bound)**.** *Let $\mathcal{A}$ be a uniformly efficient bandit algorithm for the structure $\mathcal{S}$. Then for every instance $\boldsymbol{\mu} \in \mathcal{S}$,*

$$\liminf_{T\to\infty} \frac{\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A},T)}{\log(T)} \geq C_{\mathcal{S}}(\boldsymbol{\mu})$$

where

$$C_{\mathcal{S}}(\boldsymbol{\mu}) = \inf\left\{ \sum_{a\notin a_\star(\boldsymbol{\mu})} (\mu_\star - \mu_a)c_a \;\middle|\; \begin{array}{c} \forall a\notin a_\star(\boldsymbol{\mu}),\ c_a\geq 0, \\ \forall \boldsymbol{\lambda}\in B_{\mathcal{S}}(\boldsymbol{\mu}),\ \sum_{a\notin a_\star(\theta)} d(\mu_a,\lambda_a)c_a \geq 1 \end{array} \right\} \tag{1.4}$$

$$= \inf\left\{ \sup_{\boldsymbol{\lambda}\in B_{\mathcal{S}}(\boldsymbol{\mu})} \frac{\sum_{a\notin a_\star(\boldsymbol{\mu})}(\mu_\star-\mu_a)x_a}{\sum_{a\notin a_\star(\boldsymbol{\mu})} d(\mu_a,\lambda_a)x_a} \;\middle|\; x_a\geq 0, \sum_{a\notin a_\star(\boldsymbol{\mu})} x_a = 1 \right\}. \tag{1.5}$$

The form (1.4) for the complexity quantity $c_{\mathcal{S}}(\boldsymbol{\mu})$ is the one given by Graves and Lai (1997) while the form (1.5) is the one presented by Agrawal et al. (1989) originally for $|\mathcal{S}| < \infty$. The equivalence between the two forms will be proved below but we will start by proving Theorem 1.8 for the expression (1.4). This derivation is presented as a trivial corollary of Lemma 1.7 by Graves and Lai (1997), who do not provide a proof. However, some care is needed to provide a correct proof of this result, as was noted by a colleague who reminded me that taking the liminf is *not* a linear operation. The proof given below is the outcome of a discussion with Richard Combes, Gilles Stoltz and Claire Vernade.

Proof of Theorem 1.8 We first assume that $\ell(\boldsymbol{\mu}) := \liminf_{T\to\infty} \mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A},T)/\log(T)$ is finite (otherwise the lower bound is trivial). By definition of the liminf, there exists a sequence $(T_i)_{i\in\mathbb{N}}$ such that

$$\liminf_{T\to\infty} \frac{\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A},T)}{\log(T)} = \lim_{i\to\infty} \left[\sum_{a\notin a_\star(\boldsymbol{\mu})} (\mu_\star-\mu_a) \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T_i)]}{\log(T_i)} \right] = \ell(\boldsymbol{\mu}).$$

This convergence implies that for each $a \notin a_\star(\boldsymbol{\mu})$ the sequence $(\mathbb{E}_{\boldsymbol{\mu}}[N_a(T_i)]/\log(T_i))$ is bounded. Therefore there exists a sub-sequence $(T'_i)_{i\in\mathbb{N}}$ of $(T_i)_{i\in\mathbb{N}}$ and non-negative values $c_a \in \mathbb{R}^+$ such that

$$\forall a \notin a_\star(\boldsymbol{\mu}), \quad \lim_{i\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T'_i)]}{\log(T'_i)} = c_a.$$

Hence,

$$\liminf_{T\to\infty} \frac{\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A},T)}{\log(T)} = \lim_{i\to\infty} \left[\sum_{a\notin a_\star(\boldsymbol{\mu})} (\mu_\star-\mu_a) \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T'_i)]}{\log(T'_i)} \right] = \sum_{a\notin a_\star(\boldsymbol{\mu})} (\mu_\star-\mu_a)c_a. \tag{1.6}$$

Now, using Lemma 1.7, we know that for all $\boldsymbol{\lambda} \in \mathrm{Alt}_{\mathcal{S}}(\boldsymbol{\mu})$

$$\liminf_{T\to\infty} \left[\sum_{a=1}^{A} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]}{\log(T)} d(\mu_a,\lambda_a) \right] \geq 1\,.$$

In particular, for all $\boldsymbol{\lambda} \in B_{\mathcal{S}}(\boldsymbol{\mu})$,

$$\begin{aligned} \liminf_{i\to\infty} \left[\sum_{a\notin a_\star(\boldsymbol{\mu})} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T'_i)]}{\log(T'_i)} d(\mu_a,\lambda_a) \right] &\geq 1\,, \\ \sum_{a\notin a_\star(\boldsymbol{\mu})} c_a d(\mu_a,\lambda_a) &\geq 1\,. \end{aligned}$$

Hence the quantities c_a in (1.6) satisfy $\forall \boldsymbol{\lambda} \in B_{\mathcal{S}}(\boldsymbol{\mu}), \sum_{a \notin a_\star(\boldsymbol{\mu})} c_a d(\mu_a, \lambda_a) \geq 1$, which leads to (1.4).

We now establish the equivalence between (1.4) and (1.5), by introducing the two sets of constraints

$$\mathcal{C}_{\boldsymbol{\mu}} = \left\{ x \in \mathbb{R}^A : x_a \geq 0, \sum_{a \notin a_\star(\boldsymbol{\mu})} x_a = 1 \right\} \text{ and } \mathcal{D}_{\boldsymbol{\mu}} = \left\{ c \in \mathbb{R}^A : c_a \geq 0, \inf_{\boldsymbol{\lambda} \in B_{\mathcal{S}}(\boldsymbol{\mu})} \sum_{a \notin a_\star(\boldsymbol{\mu})} d(\mu_a, \lambda_a) c_a \geq 1 \right\}.$$

We first observe that if $x \in \mathcal{C}_{\boldsymbol{\mu}}$, any c such that $\forall a \notin a_\star(\boldsymbol{\mu})$, $c_a = \frac{x_a}{\inf_{\boldsymbol{\lambda} \in B_{\mathcal{S}}(\boldsymbol{\mu})} \sum_{a \notin a_\star(\boldsymbol{\mu})} d(\mu_a, \lambda_a) x_a}$ belongs to $\mathcal{D}_{\boldsymbol{\mu}}$, which yields

$$(1.5) = \inf_{x \in \mathcal{C}_{\boldsymbol{\mu}}} \left(\sum_{a \notin a_\star(\boldsymbol{\mu})} (\mu_\star - \mu_a) \frac{x_a}{\inf_{\boldsymbol{\lambda} \in B_{\mathcal{S}}(\boldsymbol{\mu})} \sum_{a \notin a_\star(\boldsymbol{\mu})} d(\mu_a, \lambda_a) x_a} \right) \geq \inf_{c \in \mathcal{D}_{\boldsymbol{\mu}}} \left(\sum_{a \notin a_\star(\boldsymbol{\mu})} (\mu_\star - \mu_a) c_a \right) = (1.4).$$

Then, observing that if $c \in \mathcal{D}_{\boldsymbol{\mu}}$, any x such that $\forall a \notin a_\star(\boldsymbol{\mu})$, $x_a = \frac{c_a}{\sum_{a \notin a_\star(\boldsymbol{\mu})} c_a}$ belongs to $\mathcal{C}_{\boldsymbol{\mu}}$ yields

$$\begin{aligned}
(1.4) &= \inf_{c \in \mathcal{D}_{\boldsymbol{\mu}}} \frac{\sum_{a \notin a_\star(\boldsymbol{\mu})} (\mu_\star - \mu_a) c_a}{1} \geq \inf_{c \in \mathcal{D}_{\boldsymbol{\mu}}} \frac{\sum_{a \notin a_\star(\boldsymbol{\mu})} (\mu_\star - \mu_a) c_a}{\inf_{\boldsymbol{\lambda} \in B_{\mathcal{S}}(\boldsymbol{\mu})} \sum_{a \notin a_\star(\boldsymbol{\mu})} d(\mu_a, \lambda_a) c_a} \\
&= \inf_{c \in \mathcal{D}_{\boldsymbol{\mu}}} \sup_{\boldsymbol{\lambda} \in B_{\mathcal{S}}(\boldsymbol{\mu})} \frac{\sum_{a \notin a_\star(\boldsymbol{\mu})} (\mu_\star - \mu_a) \frac{c_a}{\sum_{a \notin a_\star(\boldsymbol{\mu})} c_a}}{\sum_{a \notin a_\star(\boldsymbol{\mu})} d(\mu_a, \lambda_a) \frac{c_a}{\sum_{a \notin a_\star(\boldsymbol{\mu})} c_a}} \geq \inf_{x \in \mathcal{C}_{\boldsymbol{\mu}}} \sup_{\boldsymbol{\lambda} \in B_{\mathcal{S}}(\boldsymbol{\mu})} \frac{\sum_{a \notin a_\star(\boldsymbol{\mu})} (\mu_\star - \mu_a) x_a}{\sum_{a \notin a_\star(\boldsymbol{\mu})} d(\mu_a, \lambda_a) x_a} = (1.5).
\end{aligned}$$

□

Computing the lower bound The quantity $C_{\mathcal{S}}(\boldsymbol{\mu})$ may be very hard to compute for general structured bandits, as it requires to minimize a linear function under an *infinite* number of linear constraints:

$$\forall \boldsymbol{\lambda} \in B_{\mathcal{S}}(\boldsymbol{\mu}), \sum_{a \notin a_\star(\boldsymbol{\mu})} d(\mu_a, \lambda_a) c_a \geq 1.$$

In order to make it a finite number of constraints, one can introduce for each $a \notin a_\star(\boldsymbol{\mu})$ the closest alternative in $B_{\mathcal{S}}(\boldsymbol{\mu})$ for which a is optimal, defined as

$$\lambda^a_{\mathcal{S}}(\boldsymbol{\mu}, \boldsymbol{c}) \in \operatorname*{argmin}_{\boldsymbol{\lambda} \in B_{\mathcal{S}}(\boldsymbol{\mu}) :\, a \in a_\star(\boldsymbol{\lambda})} \left[\sum_{i \notin a_\star(\boldsymbol{\mu})} c_i d(\mu_i, \lambda_i) \right] \tag{1.7}$$

and rewrite

$$C_{\mathcal{S}}(\boldsymbol{\mu}) = \inf \left\{ \sum_{a \notin a_\star(\boldsymbol{\mu})} (\mu_\star - \mu_a) c_a \;\middle|\; \begin{array}{c} \forall a \notin a_\star(\boldsymbol{\mu}) \;\; c_a \geq 0, \\ \forall a \notin a_\star(\boldsymbol{\mu}) \;\; \sum_{i \notin a_\star(\boldsymbol{\mu})} d\left(\mu_i, \left(\lambda^a_{\mathcal{S}}(\boldsymbol{\mu}, \boldsymbol{c})\right)_i\right) c_i \geq 1 \end{array} \right\}. \tag{1.8}$$

This rewriting may look a bit artificial as solving the optimization problem (1.7) and therefore computing the constraints in (1.8) may be very difficult for some structures.

However, if $\mathcal{S}$ is such that $\lambda^a_{\mathcal{S}}(\boldsymbol{\mu}, \boldsymbol{c})$ is easy to compute and independent of $\boldsymbol{c}$, the optimization problem (1.8) becomes a Linear Program for which standard optimization techniques can be used. This is the case for Lipschitz bandits, for which Magureanu et al. (2014) show that

$$\forall \boldsymbol{c}, \forall i, \; \left(\lambda^a_{\mathcal{S}_{\text{Lipschitz}}}(\boldsymbol{\mu}, \boldsymbol{c})\right)_i = \max \left\{ \mu_i, \mu_\star - L|x_a - x_i| \right\}.$$

On the contrary, for unimodal bandits, solving (1.7) happens to be harder than computing (1.4) directly and Combes and Proutière (2014a) exhibit the following closed-form expression:

$$C_{\mathcal{S}_{\text{unimodal}}}(\boldsymbol{\mu}) = \sum_{a \in \mathcal{N}(a_\star)} \frac{1}{d(\mu_a, \mu_\star)}.$$

This quantity looks like the complexity term in the un-structured case $C_{\mathcal{S}_{\text{classical}}}(\boldsymbol{\mu}) = \sum_{a \neq a_\star} \frac{1}{d(\mu_a, \mu_\star)}$ with the notable difference that the sum is restricted to arms in the set $\mathcal{N}(a_\star)$, defined as the neighboring arms of $a_\star$, which involves at most two arms. Hence the minimal regret one can achieve by exploiting the knowledge of a unimodal structure is much smaller than $C_{\mathcal{S}_{\text{classical}}}(\boldsymbol{\mu}) \log(T)$ when $K > 3$.

Through these two examples, we can guess that finding a universal algorithm that can efficiently compute $C_{\mathcal{S}}(\boldsymbol{\mu})$ for any structure $\mathcal{S}$ may be too much to ask. However, we note that so far our discussion was only based on the form (1.4) for the complexity term, and the form (1.5) may provide some insights too. Indeed, this saddle-point formulation allows to interpret $C_{\mathcal{S}}(\boldsymbol{\mu})$ as the value of some game between a player choosing a distribution over sub-optimal arms $\boldsymbol{x}$, and an adversary selecting a confusing alternative bandit model $\boldsymbol{\lambda} \in B_{\mathcal{S}}(\boldsymbol{\mu})$. This interpretation of lower bounds was recently popularized by Degenne et al. (2019) for pure exploration problems, and will be explained in more details in Chapter 4. For structured bandits, Degenne et al. (2020) then proposed a saddle-point view on a relaxation of the optimization problem that defines $C_{\mathcal{S}}(\boldsymbol{\mu})$, that they also leverage to design algorithms.

Linear bandits We now consider the important example of linear bandits with Gaussian rewards. Given a set of contexts $\{x_1, \dots, x_A\} \subseteq \mathbb{R}^d$, the reward upon choosing arm A_t is $r_t = x_{A_t}^\top \theta + \varepsilon_t$ where $\theta \in \mathbb{R}^d$ is an unknown regression parameter and $\varepsilon_t \sim \mathcal{N}\left(0, \sigma^2\right)$ is independent noise. A bandit instance in $\mathcal{L} := \mathcal{S}_{\text{linear}}$ is necessarily of the form $\boldsymbol{\lambda} = (x_a^\top \theta')_{a=1}^A$ for some θ' in $\mathbb{R}^d$ and the KL-divergence between the distribution of arm a under $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ is $d(\mu_a, \lambda_a) = \frac{(\mu_a - \lambda_a)^2}{2\sigma^2} = \frac{(x_a^\top(\theta - \theta'))^2}{2\sigma^2}$. Lattimore and Szepesvári (2017) prove a tight lower bound in this particular setting (matched by an algorithm): they show that a uniformly efficient algorithm satisfies $\liminf_{T\to\infty} \frac{\mathcal{R}_\theta(\mathcal{A},T)}{\log(T)} \geq \widetilde{C}_{\mathcal{L}}(\boldsymbol{\mu})$, where

$$\widetilde{C}_{\mathcal{L}}(\boldsymbol{\mu}) = \inf \left\{ \sum_{a \notin a_\star(\boldsymbol{\mu})} c_a(\mu_\star - \mu_a) \,\middle|\, \begin{array}{c} \forall a \in [A],\ c_a \geq 0, \\ \forall a \notin a_\star(\boldsymbol{\mu}),\ \|x_a\|^2_{H(c)^{-1}} \leq \frac{(\mu_\star - \mu_a)^2}{2\sigma^2} \end{array} \right\},$$

with $H(c) = \sum_{a=1}^A c_a x_a x_a^\top \in \mathbb{R}^{d\times d}$ and $\|x\|_M = \sqrt{x^\top M x}$ is the Mahalanobis norm associated to a symmetric and positive definite matrix M.

What is interesting here is that, to the best of my understanding, this (optimal) lower bound is *not* obtained as a consequence of Theorem 1.8, therefore we may have $\widetilde{C}_{\mathcal{L}}(\boldsymbol{\mu}) > C_{\mathcal{L}}(\boldsymbol{\mu})$. Indeed, the constraint $\|x_a\|^2_{H(c)^{-1}} \leq \frac{(\mu_\star - \mu_a)^2}{2\sigma^2}$ is obtained by a sophisticated reasoning still based on a change-of-measure argument but which considers alternative models in $\text{Alt}_{\mathcal{L}}(\boldsymbol{\mu})$ and not only in $B_{\mathcal{L}}(\boldsymbol{\mu})$. In the definitions

$$\begin{aligned} \text{Alt}_{\mathcal{L}}(\boldsymbol{\mu}) &= \left\{ \boldsymbol{\lambda} = (x_a^\top \theta')_{a=1}^A : \theta' \in \mathbb{R}^d, \exists a \notin a_\star(\boldsymbol{\mu}) : x_a^\top \theta' > x_{a_\star(\boldsymbol{\mu})}^\top \theta' \right\} \\ B_{\mathcal{L}}(\boldsymbol{\mu}) &= \left\{ \boldsymbol{\lambda} = (x_a^\top \theta')_{a=1}^A : \theta' \in \mathbb{R}^d, \forall a \in a_\star(\boldsymbol{\mu}), x_a^\top \theta' = \mu_a \text{ and } \exists a \neq a_\star(\boldsymbol{\mu}) : x_a^\top \theta' > x_{a_\star(\boldsymbol{\mu})}^\top \theta' \right\} \end{aligned}$$

the extra constraint on θ' (in orange) present in $B_{\mathcal{L}}(\boldsymbol{\mu})$ can be very restrictive. It can be checked that the constraint $\sum_{i \notin a_\star(\boldsymbol{\mu})} d\left(\mu_i, \left(\lambda_{\mathcal{S}}^a(\boldsymbol{\mu}, \boldsymbol{c})\right)_i\right) c_i \geq 1$ that appears in (1.8) and involves a minimization over $B_{\mathcal{L}}(\boldsymbol{\mu})$ is actually *not* equivalent to the constraint $\|x_a\|^2_{H(c)^{-1}} \leq \frac{(\mu_\star - \mu_a)^2}{2\sigma^2}$ present in $\widetilde{C}_{\mathcal{L}}(\boldsymbol{\mu})$.

An informal justification for $\widetilde{C}_{\mathcal{L}}(\boldsymbol{\mu})$, made much more rigorous by Lattimore and Szepesvári (2017), is the following. Letting θ such that $\boldsymbol{\mu} = (x_a^\top \theta)_{a=1}^A$ and $\overline{G}_T = \sum_{a=1}^A \mathbb{E}_\theta[N_a(T)] x_a x_a^\top$, it follows from Lemma 1.7 that for every $a \notin a_\star(\boldsymbol{\mu})$,

$$\inf_{\{\theta' : x_a^\top \theta' > x_{a_\star}^\top \theta'\}} \liminf_{T\to\infty} \frac{\frac{1}{2\sigma^2} \|\theta - \theta'\|^2_{\overline{G}_T}}{\log(T)} \geq 1.$$

Assuming we are allowed to invert the $\inf$ and the $\liminf$ yields

$$\liminf_{T\to\infty} \frac{\inf_{\{\theta' : x_a^\top \theta' > x_{a_\star}^\top \theta'\}} \frac{1}{2\sigma^2} \|\theta - \theta'\|^2_{\overline{G}_T}}{\log(T)} \geq 1.$$

Now the infimum in θ' can be computed exactly: the argmin is $\theta' = \theta + \frac{\mu_\star - \mu_a}{\|x_{a_\star} - x_a\|^2_{\overline{G}_T^{-1}}} \overline{G}_T^{-1}(x_a - x_{a_\star})$, which does not belong to $B_{\mathcal{L}}(\boldsymbol{\mu})$ in general. Computing the value of the minimization problems yields

$$\liminf_{T\to\infty} \frac{(\mu_\star - \mu_a)^2}{2\sigma^2 \|x_{a_\star} - x_a\|^2_{\overline{G}_T^{-1}} \log(T)} \geq 1.$$

This inequality is proved rigorously in Theorem 1 of Lattimore and Szepesvári (2017). It is the cornerstone for the lower bound, as $\|x_{a_\star} - x_a\|^2_{\overline{G}_T^{-1}} \simeq \|x_a\|^2_{\overline{G}_T^{-1}}$ due to the optimal arm being selected a lot.

Towards universal algorithms for structured bandits The reason why it is so important to be able to compute the lower bound is that it is possible to build asymptotically optimal algorithm for structured bandits from an oracle which can return the *optimal allocation* for any bandit instance $\boldsymbol{\mu} \in \mathcal{S}$.

Indeed, more than the value $C_{\mathcal{S}}(\boldsymbol{\mu})$ itself, the interesting thing to compute is the vector $\boldsymbol{c}$ which attains the minimum in (1.4), that we denote by $\boldsymbol{c}_{\mathcal{S}}(\boldsymbol{\mu})$ and which satisfies

$$C_{\mathcal{S}}(\boldsymbol{\mu}) = \sum_{a \notin a_\star(\boldsymbol{\mu})} (\mu_\star - \mu_a) \left(\boldsymbol{c}_{\mathcal{S}}(\boldsymbol{\mu})\right)_a .$$

From the proof of Theorem 1.8, an algorithm matching the lower bound should explore each sub-optimal arm a in such a way that $\frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]}{\log(T)} \simeq \left(\boldsymbol{c}_{\mathcal{S}}(\boldsymbol{\mu})\right)_a$. This motivates a family of algorithms that enforce this condition by computing the vector $\boldsymbol{c}_{\mathcal{S}}(\hat{\boldsymbol{\mu}}(t))$ for the current empirical estimate of the arm means $\hat{\boldsymbol{\mu}}(t)$ and then forces with some tracking procedure the current amount of exploration $N_a(t)/\log(t)$ to be close to $\left(\boldsymbol{c}_{\mathcal{S}}(\hat{\boldsymbol{\mu}}(t))\right)_a$. Such an idea was first proposed by Magureanu et al. (2014) for Lipschitz bandits, and later generalized to other structures as the OSSB algorithm which stands for Optimal Sampling for Structured Bandits (Combes et al., 2017). The optimal algorithm proposed by Lattimore and Szepesvári (2017) for linear bandits is also of this flavor.

In classical bandits, the kl-UCB algorithm is asymptotically optimal and Cappé et al. (2013) also provide a nice and explicit finite-time upper bound on its regret. On the contrary, the regret analysis of the optimal algorithms for structured bandit of the above flavor is very asymptotic in nature. Also, to ensure that the "plug-in" estimate $\boldsymbol{c}_{\mathcal{S}}(\hat{\boldsymbol{\mu}}(t))$ converges to $\boldsymbol{c}_{\mathcal{S}}(\boldsymbol{\mu})$, some additional exploration needs to be added to make sure that $\hat{\boldsymbol{\mu}}(t)$ converges to $\boldsymbol{\mu}$. This forced exploration can hurt the practical performance of the algorithm, and may not be needed for all structures. So the natural question is: could other well known principles such as optimism in face-of-uncertainty or Thompson Sampling be used to perform the right amount of exploration in structured bandits?

Upper Confidence Bounds algorithms have indeed been proposed for structured bandits. Lattimore and Munos (2014) have proposed a first structured UCB algorithm which builds an individual confidence interval on each arm a, $\mathcal{I}_a(t)$ and defines the UCB on each arm a as the maximal value for the mean of arm a in a bandit model that is compatible with both the structure and the observations:

$$\mathrm{UCB}_a(t) = \max\left\{ \lambda_a \middle| \boldsymbol{\lambda} \in \mathcal{S} \cap \mathcal{I}_1(t) \times \cdots \times \mathcal{I}_A(t) \right\}.$$

Lattimore and Munos (2014) show that the UCB algorithm associated to this index achieves a constant regret for some $\mathcal{S}$ satisfying certain conditions (which are sufficient to prove that indeed $C_{\mathcal{S}}(\boldsymbol{\mu}) = 0$ and constant regret is possible). Magureanu et al. (2014) introduce an alternative structured UCB index for Lipschitz bandits, in which instead of using individual confidence intervals for each mean, a confidence set on $\boldsymbol{\mu}$ is built by aggregating samples from all arms. This index can be extended to any structure $\mathcal{S}$ and exponential family bandit model as

$$\mathrm{UCB}_a(t) = \max\left\{ \lambda_a \middle| \boldsymbol{\lambda} \in \mathcal{S}, \sum_{i=1}^{A} N_i(t) d\left(\hat{\mu}_i(t), \lambda_i\right) \leq f(t) \right\},$$

where the threshold function $f(t)$ is chosen such that the set $\{\boldsymbol{\lambda} : \sum_{i=1}^{A} N_i(t) d(\hat{\mu}_i(t), \lambda_i) \leq f(t)\}$ contains the true vector of means $\boldsymbol{\mu}$ with probability of order $1 - 1/t$[2]. For some structures, it may be shown that using such structured UCB algorithms can be optimal (see the examples in Lattimore and Munos (2014)) or at least better than a classical UCB for some instances $\boldsymbol{\mu}$ (this is essentially what Magureanu et al. (2014) prove for Lipschitz bandits, but CKL-UCB isn't proved to be asymptotically optimal). However, there is no hope for structured UCB to be optimal for any $\mathcal{S}$ as Lattimore and Szepesvári (2017) provide a counter-example in the linear case, thus predicting the "end of optimism". On the computational side, note that computing the above $\mathrm{UCB}_a(t)$ still requires to solve an optimization problem that may be numerically hard for some structure (such as unimodal bandits).

The same linear bandit instance for which Lattimore and Szepesvári (2017) shows that a Lin-UCB algorithm cannot be asymptotically optimal is also used to disqualify Thompson Sampling as an asymptotically optimal algorithm for any structure $\mathcal{S}$ and any instance $\boldsymbol{\mu} \in \mathcal{S}$. However, there is still hope that some *variants* of Thompson Sampling (or UCB), possibly also leveraging some properties of the lower bound for a given structure $\mathcal{S}$, may be asymptotically optimal. Finding a universal algorithm of this flavor that is asymptotically optimal for every structure $\mathcal{S}$, and alleviate the computational burden and bad practical performance of OSSB would be extremely interesting[3].

In the meantime, we study in the next section a variant that is asymptotically optimal for unimodal bandits: Unimodal Thompson Sampling.

1.3 Thompson Sampling for Rank-One Bandits

The motivation to study Unimodal Thompson Sampling came from an on-going project with Claire Vernarde and Richard Combes on rank-one bandits. Rank-one bandits are structured bandits motivated by the low-rank structure of a user-item matrix in recommender systems, and were first studied by Katariya et al. (2017b,a). For Bernoulli rewards, achieving asymptotic optimality in this model was left as an open question by Katariya et al. (2017a), that we have been trying to solve with Claire and Richard. When Richard realized that rank-one bandits were particular instances of graphical unimodal bandits, we changed our focus to proposing a new analysis for Unimodal Thompson Sampling, that would give us an asymptotically optimal algorithm for rank-one bandits. I supervised the master thesis of Cindy Trinh on this topic, which led to the joint publication Trinh et al. (2020).

Rank-one bandits are unimodal In a Bernoulli rank-one bandit model, there are $A = K \times L$ arms and the vector (or matrix) of means $\boldsymbol{\mu} \in [0,1]^{K \times L}$ belongs to the set

$$\mathcal{S}_{R1} = \left\{ \boldsymbol{\mu} = (\mu_{k,\ell})_{\substack{1 \leq k \leq K \\ 1 \leq \ell \leq L}} \middle| \exists \boldsymbol{u} \in [0,1]^K, \boldsymbol{v} \in [0,1]^L : \mu_{k,\ell} = u_k v_\ell \right\}$$

For each $\boldsymbol{\mu} \in \mathcal{S}_{R1}$, this matrix of means can be written of the form $\boldsymbol{\mu} = \boldsymbol{u}\boldsymbol{v}^\top$, and is therefore of rank one. Each arm (k, ℓ) could for example model the click of a user on an button with shape k and color ℓ. In this context, the rank-one structure means that the effect of the shape and color on the click are independent.

We now argue that any rank-one bandit model $\boldsymbol{\mu} = \boldsymbol{u}\boldsymbol{v}^\top$ such that $\boldsymbol{u} > 0$ or $\boldsymbol{v} > 0$ belongs to a family of graphical unimodal bandits. This definition, extracted from Combes and Proutière (2014a), generalizes the example presented in Section 1.2.

2. Theorem 3.1 in Chapter 3 allows to make this choice.
3. Degenne et al. (2020) proposed very recently the first asymptotically optimal algorithm that doesn't resort to costly oracle computations, yet its finite time performance could still be improved.

Definition 1.9. *Given an undirected graph $G = (V, E)$ with no self-edges, a vector $\boldsymbol{\mu} = (\mu_a)_{a\in V}$ is* unimodal with respect to *G if (i) there exists a unique $a_\star \in V$ such that $\mu_{a_\star} = \max_a \mu_a$ and (ii) from any $a \neq a_\star$, we can find an increasing path to the optimal arm: $\forall a \neq a_\star$, there exists a path $p = (a_1 = a, a_2, ..., a_{m_a} = a_\star)$ of length m_a, such that for all $i = 1, ..., m_a - 1$, $(a_i, a_{i+1}) \in E$, and $\mu_{a_i} < \mu_{a_{i+1}}$.*

We denote by $\mathcal{U}(G)$ the set of vectors $\boldsymbol{\mu}$ that are unimodal with respect to G. For $\boldsymbol{\mu} \in \mathcal{U}(G)$, for all $a \in V$ we denote by $\mathcal{N}_G(a) = \{b \in V : (a, b) \in E\}$ the set of neighbors of a and let $\mathcal{N}_G^+(a) := \mathcal{N}_G(a) \bigcup \{a\}$ be the extended neighborhood of a.

We define the undirected graph $G_1 = (V, E_1)$ as the graph with vertices $V = \{1, \dots, K\} \times \{1, \dots, L\}$ and such that $((i, j), (k, \ell)) \in E_1$ if and only if $(i, j) \neq (k, \ell)$ and ($i = k$ or $j = \ell$). In words, viewing the vertices as a $K \times L$ matrix, two distinct entries are neighbors if they belong to the same line or to the same column. It can be easily shown that rank-one bandit are unimodal with respect to G_1.

Proposition 1.10. *Let $\boldsymbol{u} = (u_1, u_2, ..u_K)$ and $\boldsymbol{v} = (v_1, v_2, ..v_L)$ be two nonzero vectors such that $\boldsymbol{u} > 0$ or $\boldsymbol{v} > 0$. A rank-one bandit instance parameterized by $\boldsymbol{u}, \boldsymbol{v}$ satisfies $\boldsymbol{\mu} \in \mathcal{U}(G_1)$.*

Note that the graph G_1 has diameter two. In particular, one can exhibit increasing paths of length at most two between any sub-optimal arm (k, ℓ) and the best arm $(k_\star, \ell_\star)$. An example is provided in Figure 1.4, in which we also illustrate the neighborhood of an arm.

$$\begin{bmatrix} (u_1v_1) & (u_1v_2) & \boldsymbol{(u_1v_3)} & (u_1v_4) \\ (u_2v_1) & (u_2v_2) & \boldsymbol{(u_2v_3)} & (u_2v_4) \\ \boldsymbol{(u_3v_1)} & \boldsymbol{(u_3v_2)} & \boxed{(u_3v_3)} & \boldsymbol{(u_3v_4)} \\ (u_4v_1) & (u_4v_2) & \boldsymbol{(u_4v_3)} & (u_4v_4) \end{bmatrix} \qquad \begin{bmatrix} \boldsymbol{(u_1v_1)} & (u_1v_2) & \boldsymbol{(u_1v_3)} & (u_1v_4) \\ (u_2v_1) & (u_2v_2) & (u_2v_3) & (u_2v_4) \\ (u_3v_1) & (u_3v_2) & \boldsymbol{(u_3v_3)} & (u_3v_4) \\ (u_4v_1) & (u_4v_2) & (u_4v_3) & (u_4v_4) \end{bmatrix}$$

Figure 1.4 – $\mathcal{N}_{G_1}((3,3))$ in bold (left). Increasing path from $(3,3)$ to $(k_\star = 1, \ell_\star = 1)$ (right).

Best achievable regret for Rank-One Bandits Combes and Proutière (2014a) provide the following closed-form expression of the lower bound of Theorem 1.8 for graphical unimodal bandits, that we state below for Bernoulli reward with the associated binary relative entropy $\mathrm{kl}(\mu, \mu')$.

Proposition 1.11. *Let $G = (V, E)$ be an undirected graph with no self-edges. Any algorithm $\mathcal{A}$ which is uniformly efficient on $\mathcal{U}(G)$ satisfies*

$$\forall \boldsymbol{\mu} \in \mathcal{U}(G), \quad \liminf_{T\to\infty} \frac{\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T)}{\log(T)} \geq \sum_{a\in\mathcal{N}_G(a_\star)} \frac{\mu_\star - \mu_a}{\mathrm{kl}\,(\mu_a, \mu_\star)}.$$

Interestingly, for rank-one bandits (that is, for $G = G_1$), this lower bound coincides with the lower bound independently proved by Katariya et al. (2017a). This lower bound shows that for graphical unimodal bandits the optimal allocation is supported on the neighborhood of the leader, $\mathcal{N}_G(a_\star)$. For rank-one bandits, it means that all arms (k, ℓ) that do not belong to the same row or the same column as the best arm $a_\star = (k_\star, \ell_\star)$ should be selected only a sub-logarithmic number of times.

Any algorithm that is matching the lower bound of Proposition 1.11 is therefore asymptotically optimal for rank-one bandits when G is particularized to G_1. We now present such an algorithm.

Unimodal Thompson Sampling The principle of Unimodal Thompson Sampling (UTS) for graphical unimodal bandits is very simple: given that an optimal algorithm should only focus on neighbors of the best arm, UTS performs in every round Thompson Sampling *restricted to the neighborhood of the*

empirical best arm. In order for this to work, we should be confident about this "leader" being the true best arm most of the time. To ensure that, the algorithm performs additional *forced exploration of the leader*, governed by the leader exploration parameter γ.

More precisely, after an initialization phase where each entry is pulled once, for each $t \geq A$, the algorithm computes the empirical leader $L(t) = \text{argmax}_{a\in[A]}\hat{\mu}_a(t)$. If the number of times $L(t)$ has been the leader is a multiple of γ, UTS selects $A_{t+1} = L(t)$. Otherwise, UTS draws a posterior sample for every entry in the extended neighborhood of the leader, $\mathcal{N}_G^+(L(t)) = \mathcal{N}_G(L(t)) \cup \{L(t)\}$ and selects the arm associated to the largest posterior sample. The pseudo code of UTS is given in Algorithm 3. In the particular case of rank-one bandits, $L(t)$ is the largest entry in the matrix $\hat{\boldsymbol{\mu}}(t) \in \mathbb{R}^{K\times L}$, and $\mathcal{N}_{G_1}^+(L(t))$ is the set of entries in the matrix of arms that are in the same row or column than $L(t)$.

Algorithm 3 Unimodal Thompson Sampling UTS(γ)

Input: Horizon T, neighborhoods $(\mathcal{N}_G(a))_{a\in[A]}$, leader exploration parameter γ
Initialization: for each $a \in [A]$, $N_a = 0$, $S_a = 0$, $L_a = 0$
for $t = 1 \dots T$ **do**
  Compute the leader $A_t = \text{argmax}_{a\in[A]} S_a/N_a$
  Update the leader count $L_{A_t} \leftarrow L_{A_t} + 1$
  if $L_{A_t} \notin 0[\gamma]$ **then**
    For each $a \in \mathcal{N}_G^+(A_t) = \mathcal{N}_G(A_t) \bigcup \{A_t\}$, $\theta_a \sim \text{Beta}(S_a + 1, N_a - S_a + 1)$
    $A_t \leftarrow \text{argmax}_{a\in\mathcal{N}_G^+(A_t)}\theta_a$
  end if
  Select arm A_t and get reward R_t
  Update the number of visits and the sum of rewards:
  $N_{A_t} \leftarrow N_{A_t} + 1$, $S_{A_t} \leftarrow S_{A_t} + R_t$
end for

The idea of using an optimal bandit algorithm in the neighborhood of the leader was first proposed by Combes and Proutière (2014a) who build on the kl-UCB algorithm to propose the OSUB algorithm. OSUB is analyzed for an exploration parameter γ which is equal to the maximal degree of a node in G (that is $\gamma = K + L - 1$ for rank-one bandits). Unimodal Thompson Sampling was first proposed by Paladino et al. (2017). In their version of the algorithm, the exploration parameter γ can vary in every round and is equal to the degree of the current leader (for rank-one bandits, one also get $\gamma_t = K + L - 1$).

We now present a new analysis of UTS which shows that if the leader exploration parameter is set to any integer larger or equal to 2, UTS is asymptotically optimal. The analysis follows the general decomposition introduced by Paladino et al. (2017) and then adapts some elements from both Agrawal and Goyal (2013a) and Kaufmann et al. (2012).

Theorem 1.12. *Let $\boldsymbol{\mu}$ be a unimodal bandit instance with respect to a graph G. For all $\gamma \geq 2$, UTS with parameter γ satisfies, for every $\varepsilon > 0$,*

$$\mathcal{R}_{\boldsymbol{\mu}}(\text{UTS}(\gamma), T) \leq (1+\varepsilon) \sum_{a\in\mathcal{N}_G(a_\star)} \frac{(\mu_\star - \mu_a)}{\text{kl}(\mu_a, \mu_\star)} \log(T) + C(\boldsymbol{\mu}, \gamma, \varepsilon),$$

where $C(\boldsymbol{\mu}, \gamma, \varepsilon)$ is some constant depending on the environment $\boldsymbol{\mu}$, on ε and on γ.

Sketch of proof. To ease notation we write $V = \{1, \dots, A\}$. Recall that $L(t) = \text{argmax}_a\ \hat{\mu}_a(t)$ is the leader after t rounds and let $\ell_a(t) = \sum_{s=1}^t \mathbb{1}\left(L(s) = a\right)$ be the number of times a was leader in the first t rounds. Observe that the leader exploration scheme ensures that

$$\forall a \in [A], \forall t \in \mathbb{N}^*,\ N_a(t) \geq \lfloor \ell_a(t)/\gamma \rfloor. \tag{1.9}$$

Just like in the analysis of Combes and Proutière (2014a); Paladino et al. (2017), we first decompose the regret in two terms, according to whether the current leader is the optimal arm.

$$\begin{aligned}\mathcal{R}_{\boldsymbol{\mu}}(\mathrm{UTS}(\gamma), T) &= \sum_{a \neq a_\star} \Delta_a \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=0}^{T-1} \mathbb{1}(A_{t+1} = a)\right] \\ &= \underbrace{\sum_{a \neq a_\star} \Delta_a \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=0}^{T-1} \mathbb{1}(A_{t+1} = a, L(t) = a_\star)\right]}_{\mathcal{R}_1(T)} + \underbrace{\sum_{a \neq a_\star} \Delta_a \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=0}^{T-1} \mathbb{1}(A_{t+1} = a, L(t) \neq a_\star)\right]}_{\mathcal{R}_2(T)} .\end{aligned}$$

To upper bound $\mathcal{R}_1(T)$, it can be noted that when $a_\star$ is the leader, the selected arm a is necessarily in the neighborhood of $a_\star$, hence the sum can be restricted to the neighborhood of $a_\star$. Therefore, we expect to upper bound $\mathcal{R}_1(T)$ by the same quantity which upper bounds the regret of Thompson Sampling restricted to $\mathcal{N}_G^+(a_\star)$. Note that a proper justification does need some care, as between two times the leader is $a_\star$, UTS may update the posterior of some arms in $\mathcal{N}_G^+(a_\star)$ for they belong to the neighborhoods of other potential leaders. A careful adaptation of the analysis of Agrawal and Goyal (2013a) yields the following lemma (see Appendix B in Trinh et al. (2020) for a proof).

Lemma 1.13. *For all $\varepsilon > 0$ and all $T \geq 1$,*

$$\mathcal{R}_1(T) \leq (1+\varepsilon) \sum_{a \in \mathcal{N}_G(a_\star)} \frac{\Delta_a}{\mathrm{kl}(\mu_a, \mu_\star)} \log(T) + \tilde{C}(\boldsymbol{\mu}, \varepsilon),$$

for some quantity $\tilde{C}(\boldsymbol{\mu}, \varepsilon)$ which depends on the means $\boldsymbol{\mu}$ and on ε but not on T.

Now $\mathcal{R}_2(T)$ can be related to the probability of choosing any given sub-optimal arm a as the leader:

$$\begin{aligned}\mathcal{R}_2(T) &\leq \sum_{b \neq a_\star} \sum_{a \neq a_\star} \Delta_a \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=0}^{T-1} \mathbb{1}(A_{t+1} = a, L(t) = b)\right] \\ &\leq \sum_{b \neq a_\star} \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\mu}}\left[\mathbb{1}(L(t) = b) \sum_{a \neq a_\star} \mathbb{1}(A_{t+1} = a)\right] = \sum_{a \neq a_\star} \sum_{t=0}^{T-1} \mathbb{P}_{\boldsymbol{\mu}}\left(L(t) = a\right) .\end{aligned}$$

For each $a \neq a_\star$, we define the set of best neighbors of a, $\mathcal{B}_a = \mathrm{argmax}_{b \in \mathcal{N}_G(a)} \mu_b$. Due to the unimodal structure, we know this set is nonempty because there exists at least one arm $b \in \mathcal{N}_G(a)$ such that $\mu_b > \mu_a$ on the increasing path from a to $a_\star$. All arms belonging to $\mathcal{B}_a$ have same mean, that we denote by $\mu_{a_2} = \max_{b \in \mathcal{N}_G(a)} \mu_b$. We also introduce $\tilde{B} = \max_{a \in [A] \setminus \{a_\star\}} |\mathcal{B}_a|$, the maximal number of best arms in the neighborhood of all sub-optimal arms, which is bounded by the maximum degree of the graph. With these notations, for any $\beta \in (0,1)$, one can write

$$\begin{aligned}\sum_{t=0}^{T-1} \mathbb{P}_{\boldsymbol{\mu}}\left(L(t) = a\right) &= \underbrace{\sum_{t=0}^{T-1} \mathbb{P}_{\boldsymbol{\mu}}\left(L(t) = a, \exists a_2 \in \mathcal{B}_a, N_{a_2}(t) > (\ell_a(t))^\beta\right)}_{\mathcal{T}_1^a(T)} \\ &\quad + \underbrace{\sum_{t=0}^{T-1} \mathbb{P}_{\boldsymbol{\mu}}\left(L(t) = a, \forall a_2 \in \mathcal{B}_{\mathcal{N}_G(a)}, N_{a_2}(t) \leq (\ell_a(t))^\beta\right)}_{\mathcal{T}_2^a(T)} .\end{aligned}$$

The first term can be easily upper bounded by using the fact that if both arm a and one of its best neighbors $a_2 \in \mathcal{B}_a$ are selected enough, it is unlikely that $\hat{\mu}_a(t) \geq \hat{\mu}_{a_2}(t)$ (which has to hold if $L(t) = a$).

Using the leader exploration (1.9) and introducing $\delta_a = \frac{\mu_{a_2} - \mu_a}{2}$, one can indeed upper bound

$$\mathcal{T}_1^a(T) \leq \sum_{t=0}^{T-1} \mathbb{P}_{\boldsymbol{\mu}}\left(L(t) = a, \hat{\mu}_a(t) > \mu_a + \delta_a, N_a(t) \geq \lfloor \ell_a(t)/\gamma \rfloor\right) \tag{1.10}$$

$$+ \sum_{t=0}^{T-1} \mathbb{P}_{\boldsymbol{\mu}}\left(L(t) = a, \exists a_2 \in \mathcal{B}_a, \hat{\mu}_{a_2}(t) \leq \mu_{a_2} - \delta_a, N_{a_2}(t) > (\ell_a(t))^\beta\right), \tag{1.11}$$

Both terms can be upper bounded in the same way, by introducing the sequence of stopping times $(\tau_i^a)_i$, where τ_i^a is the instant at which arm a is the leader for the i-th time (one can have $\tau_a^k > T$ or $\tau_i^a = +\infty$ if arm a would be the leader less than i time when UTS is run forever). For example

$$\begin{aligned}
(1.11) &\leq \sum_{a_2 \in \mathcal{B}_a} \sum_{i=1}^{T-1} \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\mu}}[\mathbb{1}(L(t) = a, \ell_a(t) = i, \hat{\mu}_{a_2}(t) \leq \mu_{a_2} - \delta_a, N_{a_2}(t) > i^\beta)] \\
&= \tilde{B} \sum_{i=1}^{T-1} \mathbb{P}_{\boldsymbol{\mu}}\left(\hat{\mu}_{a_2}(\tau_i^a) \leq \mu_{a_2} - \delta_a, N_{a_2}(\tau_i^a) > i^\beta, \tau_i^a \leq T\right) \\
&\leq \tilde{B} \sum_{i=1}^{T-1} \sum_{u=i^\beta}^{T} \mathbb{P}_{\boldsymbol{\mu}}\left(\hat{\mu}_{a_2,u} \leq \mu_{a_2} - \delta_a, N_{a_2}(\tau_i^a) = u\right) \\
&\leq \tilde{B} \sum_{i=1}^{\infty} \sum_{u=i^\beta}^{\infty} \exp(-2\delta_a^2 u) \leq \tilde{B} \sum_{i=1}^{\infty} \frac{\exp(-2\delta_a^2 i^\beta)}{1 - \exp(-2\delta_a^2)}.
\end{aligned}$$

The notation $\hat{\mu}_{a_2,u}$ denotes the empirical mean of the first u observations from arm k_2, which are i.i.d. Bernoulli random variables with mean μ_{a_2}. Proceeding similarly for (1.10) yields

$$\mathcal{T}_1^a(T) \leq \sum_{i=1}^{\infty} \frac{\exp(-2\delta_a^2 i^\beta)}{1 - \exp(-2\delta_a^2)} + \sum_{i=1}^{\infty} \frac{\exp(-2\delta_a^2 \lfloor i/\gamma \rfloor)}{1 - \exp(-2\delta_a^2)} < \infty.$$

To conclude the proof, we show that $\mathcal{T}_2^a(T)$ is also upper bounded by a constant for some well chosen value of $\beta \in (0,1)$. This follows from the following adaptation of Proposition 1 in Kaufmann et al. (2012), which says that for vanilla Thompson Sampling restricted to $\mathcal{N}_G^+(a_\star)$, the (unique) optimal arm a_2 cannot be drawn too many times. Observe that Lemma 1.14 permits to handle possible multiple optimal arms, and handles again the extra difficulty that arms in $\mathcal{N}_G^+(a_\star)$ are not only selected when a is the leader. Its proof can be found in Appendix C of Trinh et al. (2020).

Lemma 1.14. *When $\gamma \geq 2$, there exists $\beta \in (0,1)$ and a constant $D_a(\boldsymbol{\mu}, \beta, \gamma)$ such that*

$$\sum_{t=0}^{T-1} \mathbb{P}_{\boldsymbol{\mu}}\left(L(t) = a, \forall a_2 \in \mathcal{B}_a, N_{a_2}(t) \leq (\ell_a(t))^b\right) \leq D_a(\boldsymbol{\mu}, \beta, \gamma).$$

□

Solving rank-one bandits, and beyond We illustrate below the practical impact of using Unimodal Thompson Sampling for solving the rank-one bandit problem. In Figure 1.5, we see that on a large rank-one instance in which the previous state-of-the art algorithm RANK1ELIMKL (Katariya et al., 2017a) was shown to outperform kl-UCB, UTS largely outperforms RANK1ELIMKL. Moreover, unlike RANK1ELIMKL, UTS also outperforms kl-UCB on smaller instances, as illustrated in Figure 1.6.

In the experiments whose results are reported in these figures, UTS was run with the parameter $\gamma = 2$, which appeared to be consistently the best choice among the different values we tried on different scenarios. We also did some experiments with the tuning $\gamma = +\infty$, which corresponds to no forced

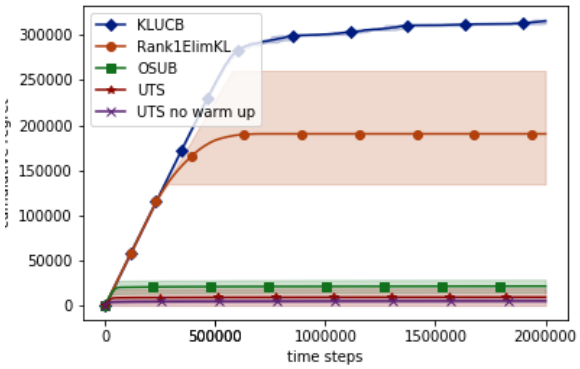


Figure 1.5 – Regret as a function of the horizon T for various algorithm on a 128×128 instance with $\boldsymbol{u} = \boldsymbol{v} = (0.75, 0.25, \dots, 0.25)$ (average over $N = 20$ simulations)

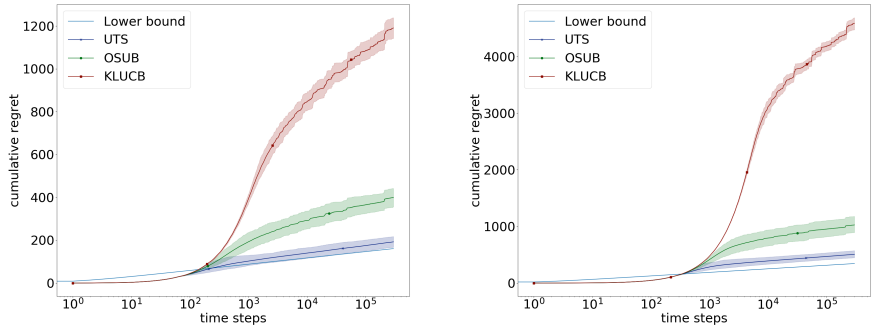


Figure 1.6 – Regret as a function of the horizon T (in log scale) for various algorithm on a 8×8 (left) and 16×16 (right) instance with $\boldsymbol{u} = \boldsymbol{v} = (0.75, 0.25, \dots, 0.25)$ (average over $N = 100$ simulations)

exploration of the leader. We conjecture that this choice, which is currently not supported by theory, also leads to asymptotically optimal regret. However, our experiments overall revealed that forcing the exploration of the leader (which actually introduces more *exploitation* in the algorithm) is actually helping in practice.

Finally, we can take a step back and think about the insights provided by this algorithm for the design of asymptotically optimal algorithms for more general structures. In this particular case, assuming that the optimal arm $a_\star$ is known, it is possible to design a variant of Thompson Sampling, namely Thompson Sampling restricted to $\mathcal{N}_G(a_\star)$, which achieves the optimal allocation among sub-optimal arms. This property, combined with the leader exploration scheme described above, easily yields asymptotic optimality. In future work we will investigate a more general combination of Thompson Sampling with some knowledge of the optimal allocation $\boldsymbol{c}_{\mathcal{S}}(\boldsymbol{\mu})$ achieving the optimal allocation. It would also be interesting to find out whether forcing the exploration of the leader remains empirically good for more general structures.

Chapter 2

Multi-Player Bandits

In the last ten years, researchers working on cognitive radio systems started to investigate the use of multi-armed bandit tools for adaptive channel selection (Jouini et al., 2009). In particular, in crowded communication networks *several* radio devices (agents) may try to communicate in the *same* set of channels (arms), which gives rise to an interesting *multi-agent* bandit problem. This chapter presents some contributions to the study of this problem. Some of them were obtained in collaboration with Lilian Besson, my first PhD student, whom I supervised with Christophe Moy (Université de Rennes).

2.1 Several Decentralized Bandit Problems

In a Multi-Player Multi-Armed Bandit model (MP-MAB), there are A arms and M agents, or players, with $M \leq A$. For each player $m \in [M]$, each arm a is associated to an i.i.d. reward stream $(X^m_{a,t})_{t\in\mathbb{N}}$ with mean μ^m_a. In each round t, each player m selects an arm A^m_t and receives a reward R^m_t with is equal to $X^m_{A^m_t,t}$ if she is the only one to select arm A^m_t and to zero reward otherwise [1]. Hence

$$R^m_t = X^m_{A^m_t,t}\left(1 - \mathbb{1}\left(\mathcal{C}^m_t\right)\right),$$

where $\mathcal{C}^m_t := \{\exists m' \neq m : A^m_t = A^{m'}_t\}$ is the event that a *collision* occurs for player m at time t. The goal is to find an arm selection strategy for each agent that maximizes the expected *total reward* of the system, $\mathbb{E}\left[\sum_{t=1}^{T}\sum_{m=1}^{M} R^m_t\right]$, without allowing explicit communications between agents.

This model is inspired by what may happen in a cognitive radio system in which M smart radio devices are allowed to use A distinct frequency bands (channels). The random variable $X^m_{a,t}$ models the quality of channel a for device m at time t. For example, $X^m_{a,t} = 1$ if the channel is free from other types of users, $X^m_{a,t} = 0$ otherwise. In each round, all devices select a channel for communication. If several devices try to communicate in the same channel none of their communications is successful, possibly due to some interference. From this story, one can think of two interesting settings:

- **the homogeneous setting** in which $\forall a \in [A], \mu^m_a = \mu_a$, that is the average quality of a channel a is the same for the M devices. This assumption makes sense for example when $X^m_{a,t} = X_{a,t}$ measures the availability of channel a, and all devices communicate successfully on a free channel.
- **the heterogeneous setting** in which one may have $\mu^m_a \neq \mu^{m'}_a$ for some channel a and two devices m and m'. This assumption makes sense when there are external reasons (such as configuration or position) that make some channels better for some devices.

1. Alternative models in which the reward is *reduced* in some other way when a collision occurs can be considered, see e.g., Liu and Zhao (2010).

The multi-player MAB was pioneered by Liu and Zhao (2010) and Anandkumar et al. (2010) who considered the homogeneous setting in which $X_{a,t}^m = X_{a,t}$, and we will first stick to this particular case, that already captures the interesting trade-off between exploration, exploitation... and collisions. We will go back to the heterogeneous setting in Section 2.4.

To understand the challenges of learning in a MP-MAB model, we now clarify what *information* each agent can use to guide its arm selection strategy.

Several feedback models We consider a *decentralized* setting in which agent m only observes her own reward R_t^m, and maybe a little bit more. Indeed, for some applications it makes sense that the agent also observes the *sensing information* $X_t^m := X_{A_t^m,t}$ and/or the *collision information* $C_t^m := \mathbb{1}(\mathcal{C}_t^m)$. With this notation, one can rewrite

$$R_t^m = X_t^m(1 - C_t^m).$$

One can consider four different feedback models that we present below and of which we explain the relevance for cognitive radio systems:

- **(I) Learning from rewards only**: player m only observes R_t^m at the end of round t. In the context of IoT networks, it is common that after sending a packet to a base station, a radio device receives an acknowledgment if the communication is successful. The smart device can then adjust its next channel selection using this reward signal. However when a communication fails, it is impossible to know whether it is because of the intrinsic bad channel quality at this time ($X_t^m = 0$) or because a collision with another smart device occurred ($C_t^m = 1$).
- **(II) Learning from sensing and rewards**: player m first observes X_t^m, then the reward R_t^m. This is the feedback model originally studied by Anandkumar et al. (2010), as it is relevant to model Opportunistic Spectrum Access (OSA). In OSA, there are two types of users of radio resources: primary users who pay to be guaranteed channel access when they want to communicate, and secondary users (our M players), who can communicate if the channel they select is free from primary users. When selecting a channel a, a secondary user has to perform *sensing* to detect the presence ($X_{a,t} = 0$) or absence ($X_{a,t} = 1$) of primary users. If the channel is free, the secondary user can try to communicate, and the communication is successful if and only if no collision occurs.
- **(III) Learning from collisions and rewards**: player m first observes C_t^m (that is, she always knows if she was the only one to select a channel), then the reward R_t^m. It would correspond to a system in which a radio device has a way to first detect the presence of the other devices using the same standard (the $M-1$ other players) on the selected channel.
- **(IV) Learning with full feedback**: player m observes X_t^m *and* C_t^m (and therefore R_t^m).

For some types of distributions for the arms, these four feedback models are actually equivalent. Indeed, if for each arm a, $\mathbb{P}(X_{a,1} = 0) = 0$ (which is the case if ν_a is a continuous distribution for example), observing R_t^m always permits to reconstruct both X_t^m and C_t^m. However, this assumption is not satisfied for Bernoulli distributions which are often used to model successful or failed communications. Hence, in the rest of the chapter, we will focus on the MP-MAB model with Bernoulli arms [2]. In that case, none of the four feedback models are equivalent: under **(II)** one can only reconstruct C_t^m when $X_t^m = 1$ whereas under **(III)** one can only reconstruct X_t^m when $C_t^m = 0$.

Given one of these observation models, a multi-player MAB strategy is a tuple $\mathcal{A}_M = (\mathcal{A}^1, \dots, \mathcal{A}^M)$ of strategies for each player such that under $\mathcal{A}^m$ the arm A_t^m chosen by player m at time t depends on the observations made by player m up to round $t-1$ (and possibly some independent external randomness).

2. All the algorithms that we will propose naturally extend to reward distributions that are bounded in $[0, 1]$.

For example, under the observation model **(II)**, A_t^m is $\mathcal{F}_{t-1}^m$-measurable where

$$\mathcal{F}_t^m = \sigma\left(U_0^m, A_1^m, X_1^m, R_1^m, U_1^m, \dots, A_t^m, R_t^m, X_t^m, U_t^m\right)$$

is the σ-algebra generated by the observations up to round t, where $U_t^m \sim \mathcal{U}([0,1])$ materializes the extra randomness possibly used at round $t+1$. Under other observation models, one can similarly define an appropriate σ-algebra $\mathcal{F}_t^m$.

Regret for multi-player bandits In the homogeneous setting, the oracle strategy that maximizes the total reward is simple: the M players should be assigned to the M arms with largest means. One can therefore define the regret incurred by algorithm $\mathcal{A}_M$ in T rounds as the difference between the total reward of the oracle strategy and that of the algorithm:

$$\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}_M, T) = T\left(\sum_{a=1}^{M} \mu_{[a]}\right) - \mathbb{E}\left[\sum_{t=1}^{T}\sum_{m=1}^{M} R_t^m\right],$$

where $\mu_{[a]}$ denotes the mean of the arm with a-th largest mean: $\mu_{[1]} \geq \mu_{[2]} \geq \cdots \geq \mu_{[A]}$. We denote by $\mathrm{TopM}(\boldsymbol{\mu})$ the set of M arm with largest mean, $\{[1], \dots, [M]\}$.

As in the classical MAB problem, one can express the regret in terms of the number of selections of each arm, but the rewriting also features the number of collisions that happen on each arm. Letting $N_a^m(t) = \sum_{s=1}^{t} \mathbb{1}\left(A_s^m = a\right)$ be the number of selections of arm a by player m in the first t rounds,

$$N_a(t) = \sum_{m=1}^{M} N_a^m(t) \quad \text{and} \quad C_a(t) = \sum_{s=1}^{t}\sum_{m=1}^{M} C_t^m \mathbb{1}\left(A_s^m = a\right)$$

respectively denote the total number of selections of arm a and the total number of collisions on arm a. In the paper Besson and Kaufmann (2018) we propose the following rewriting of the regret, which generalizes the usual regret decomposition used in the classical MAB.

Lemma 2.1. *For every $\boldsymbol{\mu}$ such that $\mu_{[M]} > \mu_{[M+1]}$ it holds that*

$$\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}_M, T) = \sum_{a=M+1}^{A} (\mu_{[M]} - \mu_{[a]})\mathbb{E}_{\boldsymbol{\mu}}[N_{[a]}(T)] + \sum_{a=1}^{M} (\mu_{[a]} - \mu_{[M]})\left(T - \mathbb{E}_{\boldsymbol{\mu}}[N_{[a]}(T)]\right) + \sum_{a=1}^{A} \mu_a \mathbb{E}_{\boldsymbol{\mu}}[C_a(T)].$$

The three terms in this regret decomposition suggest that a good multi-player algorithm $\mathcal{A}_M$

- ➜ selects each sub-optimal arm in $\{[M+1], \dots, [A]\}$ very little,
- ➜ selects each optimal arm, i.e. each arm in $\mathrm{TopM}(\boldsymbol{\mu})$, almost in every round,
- ➜ and experiences few collision on each arm.

For classical bandits ($M = 1$) the regret decomposition features only the first of these three terms, and upper bounds on the number of selections of each sub-optimal arm directly yield a regret bound. With $M > 1$, the second and third terms are the most difficult to upper bound, as we shall see.

This decomposition of the regret also applies to a "centralized" version of our problem, in which in each round a central controller selects a subset $(A_t^1, \dots, A_t^M)$ of M distinct arms with the same goal to maximize total reward. This so-called multiple-play bandit problem was introduced by Anantharam et al. (1987). For multiple-play bandits, the third term in Lemma 2.1 disappears as the selected arms are always distinct, however the second term remains and an asymptotically optimal algorithm in this setting actually needs to satisfy $(T - \mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]) = o(\log(T))$ for every arm a in $\mathrm{TopM}(\boldsymbol{\mu})$. Proving this is in general the tricky part of the analysis a multiple-play bandit algorithms.

Algorithms for the multiple-play problem are interesting as they can serve as a "centralized" (oracle) baseline for our decentralized problem. For example, the algorithms that select the arms with M largest Thompson samples or M largest kl-UCB indices were proved to be asymptotically optimal for the multiple-play bandit problem (Komiyama et al., 2015; Luedtke et al., 2019). Equipped with these baselines, we are now ready to investigate the cost of decentralization.

2.2 Algorithms for the Homogeneous Multi-Player MAB

During the PhD of Lilian, we worked together on algorithms for the Bernoulli MP-MAB under the observation models **(I)** and **(II)** that we found the most relevant for applications to communication.

Lilian's PhD had an emphasis on the design of smart IoT devices (e.g. connected watches, fridges, sensors...). The most common assumption for IoT communication it that each smart object would be able to learn how to communicate more efficiently only by receiving an acknowledgment (reward) from the base station after each successful communication. Hence we first investigated algorithms under observation model **(I)**, and came across the surprisingly good behavior of a "selfish" learning approach. However, the theoretical guarantees obtained for this algorithm are a bit disappointing, as we shall see. We then turned our attention to algorithms that could be used for the Opportunistic Spectrum Access problem, i.e. under observation model **(II)**. We proposed the MCTopM algorithm, which outperformed state-of-the-art algorithms both in theory and in practice when the paper Besson and Kaufmann (2018) was published.

2.2.1 The Selfish and MCTopM algorithms

Selfish learning from rewards only This idea of the Selfish heuristic is the following: each player m pretends she is alone and uses a classical MAB algorithm in order to select A_{t+1}^m based on the past selected arms $A_1^m, \dots, A_t^m$ and past observed rewards $R_1^m, \dots, R_t^m$. Due to the presence of other players, the distribution of R_t^m conditionally to the past observations is *not* $\mathcal{B}\left(\mu_{A_t^m}\right)$, hence player m is not interacting with a *stochastic* bandit model. Thus in principle, only the use of adversarial bandit algorithms such that EXP3 (Auer et al., 2002b) is legitimate. Quite surprisingly, we found that the performance is better when each player uses a *stochastic* MAB algorithm such as kl-UCB or Thompson Sampling. The pseudo-code for Selfish-kl-UCB is given in Algorithm 4 below.

Algorithm 4 Selfish-kl-UCB for player m under observation model **(I)**

Initialization: for each $a \in [A]$, $S_a \leftarrow 0$, $N_a \leftarrow 0$,
for $t = 1 \dots T$ **do**
  if $t \in [A]$ **then**
    $A_t^m = t$
  else
    $A_t^m \in \underset{a \in [A]}{\operatorname{argmax}} \left[\max \left\{ q \in [0,1] : N_a \mathrm{kl}\left(\frac{S_a}{N_a}, q\right) \leq \log(t) \right\} \right]$
  end if
  Select arm A_t^m and observe reward R_t^m.
  $S_{A_t^m} \leftarrow S_{A_t^m} + R_t^m$, $N_{A_t^m} \leftarrow N_{A_t^m} + 1$.
end for

In Bonnefoi et al. (2017), we present an empirical evaluation the Selfish heuristic in a model which is a bit more realistic for long-range communications in unlicensed bands which is the dominant approach for IoT networks (Centenaro et al., 2016): there is a very large number $M > A$ of radio devices, but in each time step, only a small fraction of the devices needs to communicate. We then investigated the theoretical properties of Selfish in the simpler MP-MAB and our conclusions are a bit disappointing: we show that Selfish has actually a small probability to "get stuck" on some MP-MAB instances in which the players may end up colliding at every round (see Appendix E of Besson and Kaufmann (2018)). The work of Boursier and Perchet (2019) sheds light on the same phenomenon.

Still, our experiments showed that *most of the time*, Selfish does perform extremely well empirically, sometimes even better than algorithms using more feedback (sensing or collisions). Yet, this algorithm

does not have a logarithmic regret. Since we made this observation, algorithms with logarithmic regret in the observation model **(I)** were proposed by Lugosi and Mehrabian (2018) and Boursier and Perchet (2019). However, the practical performance of these algorithms is not clear, and some of the proposed algorithms require the knowledge of a lower bound on $\mu_{[M]}$ or $\mu_{[A]}$. Therefore, a multi-player algorithm that learns from rewards only with minimal regret and good finite-time performance is still to be found.

MCTopM for learning based on rewards and sensing Existing algorithms for observation model **(II)** such as TDFS (Liu and Zhao, 2010) or RhoRand (Anandkumar et al., 2010) combine a classical bandit algorithm with an *orthogonalization mechanism*. We propose a new such combination, called MCTopM, which builds on kl-UCB and a new orthogonalization mechanism inspired by the Musical Chair algorithm of Rosenski et al. (2016).

The first ingredient to describe our algorithm (and other related ones) is the set of candidate "Top M" arms, taken to be the M arms with highest Upper Confidence Bounds for each player m:

$$\widehat{M^m}(t) = \left\{[1], \dots, [M], \text{ where } \mathrm{UCB}^m_{[1]}(t) \geq \cdots \geq \mathrm{UCB}^m_{[A]}(t)\right\},$$

where $\mathrm{UCB}^m_a(t)$ is the kl-UCB index of arm a for player m:

$$\mathrm{UCB}^m_a(t) = \max\left\{q : N^m_a(t)\mathrm{kl}(\hat{\mu}^m_a(t), q) \leq \log(t) + c\log\log(t)\right\}, \tag{2.1}$$

with $\hat{\mu}^m_a(t) = \frac{1}{N^m_a(t)}\sum_{s=1}^t X^m_s \mathbb{1}(A_s = a)$ if $N^m_a(t) > 1$. Note that as the sensing information X^m_t is always observed, the UCB of the chosen arm will always be refined, regardless of collisions.

In MCTopM and other algorithms, player m selects at round t one *well-chosen arm* in $\widehat{M^m}(t-1)$. To choose this arm, the TDFS algorithm of Liu and Zhao (2010) relies on a pre-agreement between players on the time-steps at which they will target the arm in $\widehat{M^m}(t-1)$ with a-th largest UCB, for $a \in [M]$. The RhoRand algorithm of Anandkumar et al. (2010) instead assigns a rank $R^m_t \in [M]$ to each player m at each round t so that the selected arm A^m_t is the arm with R^m_t-th largest UCB in $\widehat{M^m}(t-1)$. If a collision is observed after this selection, a new rank R^m_{t+1} is assigned to player m uniformly at random.

Our algorithm, called MCTopM, doesn't rely on such a rank but instead tries to enforce a minimal number of arm switches in the following way. The first time player m is the only one to select some arm A^m_t in $\widehat{M^m}(t-1)$, she *fixes* herself on that arm for the next rounds ($s^m_{t+1} = 1$), and keeps selecting it, regardless of future collisions, until some future time t' in which this arm is no longer in $\widehat{M^m}(t'-1)$. When this happens, a new arm is selected uniformly at random *among arms with smaller UCB:* [3]

$$A^m_{t'} \sim \mathcal{U}\left(\widehat{M^m}(t'-1) \bigcap \left\{a : \mathrm{UCB}_a(t'-2) \leq \mathrm{UCB}_{A^m_{t'-1}}(t'-2)\right\}\right), \tag{2.2}$$

and player m *un-fixes* herself ($s^m_{t'} = 0$). When an un-fixed player experiences a collision in round t, she remains un-fixed for next round ($s^m_{t+1} = 1$) and selects a new arm $A^m_{t+1} \sim \widehat{M^m}(t)$. This orthogonalization strategy is inspired by the Musical Chair protocol used in the work of Rosenski et al. (2016): the first time a player finds an empty chair (i.e. an arm free from other players) she get "seated" on it. In MCTopM, player remain seated until their arm does not look good anymore.

In Besson and Kaufmann (2018), we present and analyze the MCTopM algorithm under observation model **(IV)**, which is slightly easier as collision are always observed. The pseudo-code of this version of the algorithm is given in Algorithm 5. However, it is possible to define MCTopM for observation model **(II)**, as described in Algorithm 6. When only sensing and rewards are observed, an un-fixed player m knows that she experienced a collision only when $(X^m_{t-1} = 1) \cap (R^m_{t-1} = 0)$, hence the modification in line 6 of the algorithm (highlighted in orange). Similarly, the only way player m can be sure no collisions occurred before fixing is when observing a reward ($R^m_{t-1} = 1$), hence the updated lines 12-15.

3. We know that there exists at least one such arm as $A^m_{t'-1} \in \widehat{M^m}(t'-2) \backslash \widehat{M^m}(t'-1)$, hence this arm must have been replaced by one arm that was not in $\widehat{M^m}(t'-2)$, whose UCB is therefore smaller than that of $A^m_{t'-1}$.

Algorithm 5 MCTopM for player m under observation model **(IV)**

Initialization: Let $A_0^m \sim \mathcal{U}(\{1,\dots,A\})$, $s_1^m = 0$
for $t = 1 \dots T$ **do**
  if $A_{t-1}^m \notin \widehat{M^m}(t-1)$ **then**
    `# switch if the previous arm looks bad`
    $A_t^m \sim \mathcal{U}\left(\widehat{M^m}(t-1) \cap \left\{a : \mathrm{UCB}_a(t-2) \le \mathrm{UCB}_{A_{t-1}^m}(t-2)\right\}\right)$; $s_t^m = 0$
  else if $(s_{t-1}^m = 0)$ and $(C_{t-1}^m = 1)$ **then**
    `# unfixed player experiencing a collision switches`
    $A_t^m \sim \mathcal{U}(\widehat{M^m}(t-1))$; $s_t^m = 0$
  else
    `# play the previous arm`
    $A_t^m = A_{t-1}^m$; $s_t^m = 1$
  end if
  Select arm A_t^m. Observe X_t^m and R_t^m.
  Compute $(\mathrm{UCB}_a^m(t))_{a=1}^A$ and the set $\widehat{M^m}(t)$.
end for

Algorithm 6 MCTopM for player m under observation model **(II)**

Initialization: Let $A_0^m \sim \mathcal{U}(\{1,\dots,A\})$, $s_1^m = 0$
for $t = 1 \dots T$ **do**
  if $A_{t-1}^m \notin \widehat{M^m}(t-1)$ **then**
    `# switch if the previous arm looks bad`
    $A_t^m \sim \mathcal{U}\left(\widehat{M^m}(t-1) \cap \left\{a : \mathrm{UCB}_a(t-2) \le \mathrm{UCB}_{A_{t-1}^m}(t-2)\right\}\right)$; $s_t^m = 0$
  else if $(s_{t-1}^m = 0)$ and $(X_{t-1}^m = 1) \cap (R_{t-1}^m = 0)$ **then**
    `# unfixed player experiencing a collision switches`
    $A_t^m \sim \mathcal{U}(\widehat{M^m}(t-1))$; $s_t^m = 0$
  else
    `# play the previous arm`
    $A_t^m = A_{t-1}^m$
    if $(s_{t-1}^m = 1) \bigcup (R_{t-1}^m = 1)$ **then**
      `# stay fixed or get fixed after a first reward`
      $s_t^m = 1$
    end if
  end if
  Select arm A_t^m. Observe X_t^m and R_t^m.
  Compute $(\mathrm{UCB}_a^m(t))_{a=1}^A$ and the set $\widehat{M^m}(t)$.
end for

2.2.2 Elements of analysis of MCTopM

To ease the notation, we assume that the arms' means satisfy $\mu_1 \ge \mu_2 \ge \mu_M > \mu_{M+1} \ge \dots \mu_A$. In this section, we present an upper bound the regret of MCTopM under the extra assumption that all these means are distinct (which is necessary for the upper bound on the collisions given in Theorem 2.3). The regret decomposition of Lemma 2.1 features two important quantities: the number of sub-optimal selections, that is $\mathbb{E}[N_a(T)]$ for each arm $a \in \{M+1,\dots,A\}$, and the number of collisions that happen on a given arm a, $\mathbb{E}[C_a(T)]$.

Upper bound on the number of sub-optimal selections The first part of our analysis exploits the definition of the "Top-M" set $\widehat{M^m}(t)$ in order to get an upper bound on $\mathbb{E}[N_a^m(T)]$ for each player m. We highlight that this result is not specific to the MCTopM algorithm but applies for any algorithm for which $A^m(t) \in \widehat{M^m}(t-1)$, such as variants of RhoRand and TDFS based on kl-UCB indices.

Theorem 2.2. *For each player m, under any algorithm for which $A^m(t) \in \widehat{M^m}(t-1)$ for all $t \geq 1$, it holds that for all $a \in \{M+1, \dots, A\}$,*

$$\mathbb{E}_{\boldsymbol{\mu}}[N_a^m(T)] \leq \frac{\log(T)}{\mathrm{kl}(\mu_a, \mu_M)} + \mathcal{O}_{\boldsymbol{\mu}}(\sqrt{\log(T)}) \,,$$

for c in the definition of the kl-UCB *indices* (2.1) *that satisfies $c \geq 3$.*

Proof. The key observation is that if $(A_t^m = a)$ for some sub-optimal arm a, as $A_t^m \in \widehat{M^m}(t-1)$, there must exists one arm $b \in [M]$ that is not in $\widehat{M^m}(t-1)$, and whose UCB is therefore smaller than that of arm a:

$$(A_t^m = a) = (A_t^m = a, \exists b \in [M] : \mathrm{UCB}_b^m(t) < \mathrm{UCB}_a^m(t)) \,.$$

This yields

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu}}[N_a^m(T)] &= \sum_{t=1}^{T} \mathbb{P}_{\boldsymbol{\mu}}\left(A_t^m = a,\ \exists b \in [M] :\ \mathrm{UCB}_b^m(t-1) < \mathrm{UCB}_a^m(t-1)\right) \\
&\leq \sum_{t=1}^{T} \mathbb{P}_{\boldsymbol{\mu}}\left(A_t^m = a,\ \exists b \in [M] :\ \mathrm{UCB}_b^m(t-1) \leq \mathrm{UCB}_a^m(t-1), \forall b' \in [M] :\ \mathrm{UCB}_{b'}^m(t-1) \geq \mu_{b'}\right) \\
&\quad + \sum_{t=1}^{T} \mathbb{P}_{\boldsymbol{\mu}}\left(\exists b' \in [M] :\ \mathrm{UCB}_{b'}^m(t-1) < \mu_{b'}\right) \\
&\leq \sum_{t=1}^{T} \mathbb{P}_{\boldsymbol{\mu}}\left(A_t^m = a,\ \exists b \in [M] :\ \mu_b \leq \mathrm{UCB}_a^m(t-1)\right) + \sum_{b'=1}^{M} \sum_{t=1}^{T} \mathbb{P}_{\boldsymbol{\mu}}\left(\mathrm{UCB}_{b'}^m(t-1) < \mu_b\right) \\
&\leq \sum_{t=1}^{T} \mathbb{P}_{\boldsymbol{\mu}}\left(A_t^m = a,\ \mu_M \leq \mathrm{UCB}_a^m(t-1)\right) + \sum_{b'=1}^{M} \sum_{t=1}^{T} \mathbb{P}_{\boldsymbol{\mu}}\left(\mathrm{UCB}_{b'}^m(t-1) < \mu_b\right) ,
\end{aligned}$$

where the last inequality comes from the fact that μ_M is the smallest of the μ_b for $b \in [M]$. Now each term can be upper bounded using standard tools developed by Cappé et al. (2013) for the analysis of kl-UCB. For $c \geq 3$, the second term is in $\mathcal{O}(\log(\log(T)))$ using the concentration inequality of Lemma 0.3 stated in the Introduction, while as $\mu_a < \mu_M$ the first term can be upper bounded by

$$\frac{\log(T) + 3\log\log(T)}{\mathrm{kl}(\mu_a, \mu_M)} + \mathcal{O}_{\boldsymbol{\mu}}\left(\sqrt{\log(T)}\right) \,,$$

using the same technique as in Appendix A of Cappé et al. (2013).

□

Upper bound on the number of collisions The most intricate part of our analysis is the control of the total number of collisions $C(T) = \sum_{a=1}^{A} C_a(T)$. It is given in the following theorem.

Theorem 2.3. *Under observation model **(IV)**, the total number of collisions under* MCTopM *satisfies, if all the arm means are distinct,*

$$\mathbb{E}_{\boldsymbol{\mu}}[C(T)] \leq 2M^3 \left(\sum_{a,b:\mu_a<\mu_b} \frac{1}{\mathrm{kl}(\mu_a, \mu_b)} \right) \log(T) + o_{\boldsymbol{\mu}}(\log(T)) \,.$$

*Under observation model **(II)**, the upper bound is multiplied by $(\min_k \mu_k)^{-1}$.*

Sketch of proof. Under MCTopM, if all players are fixed ($s_t^m = 1$), it means that they found a configuration of different arms, therefore no collision can occur. Hence, if a collision occurs, it occurs at least for one non-fixed player: $(C_t^m = 1) \subseteq \bigcup_{j\in[M]} \left(C_t^j = 1, s_t^j = 0\right)$. Therefore

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu}}[C(T)] &= \sum_{m=1}^{M}\sum_{t=1}^{T} \mathbb{P}_{\boldsymbol{\mu}}(C_t^m = 1) \\
&\leq M \sum_{m=1}^{M}\sum_{t=1}^{T} \mathbb{P}_{\boldsymbol{\mu}}\left(C_t^m = 1, s_t^m = 0\right) \\
&\leq M \sum_{m=1}^{M}\sum_{t=1}^{T} \mathbb{P}_{\boldsymbol{\mu}}\left(A_{t-1}^m \notin \widehat{M^m}(t-1)\right) + M \sum_{m=1}^{M} \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} \mathbb{1}\left(A_{t-1}^m \in \widehat{M^m}(t-1), C_t^m = 1, s_t^m = 0\right)\right].
\end{aligned}$$

The first term can be upper bounded using the following lemma, which shows that MCTopM cannot switch arms too much. The proof of this result, that can be found in Appendix D.2 of Besson and Kaufmann (2018) crucially exploits the rule (2.2) for selecting a new arm when $A_{t-1}^m \notin \widehat{M^m}(t-1)$.

Lemma 2.4. *For any arm $a \in [A]$,* MCTopM *satisfies, if all the arm means are distinct,*

$$\sum_{t=1}^{T} \mathbb{P}_{\boldsymbol{\mu}}\left(A_t^m = a, a \notin \hat{M}^m(t)\right) = \left(\sum_{b:\mu_b<\mu_a} \frac{1}{\mathrm{kl}(\mu_a,\mu_b)} + \sum_{b:\mu_b>\mu_a} \frac{1}{\mathrm{kl}(\mu_b,\mu_a)}\right)\log(T) + o_{\boldsymbol{\mu}}(\log(T)) \,.$$

To control the second term, we note that as long as $A_{t-1}^m \in \widehat{M^m}(t-1)$ and the player is not fixed ($s_t^m = 0$), arm A_t^m is chosen at random from a pool of M arms, out of which as least one if free from other players. The probability to fix in the next round is therefore at least $\frac{1}{M}$ if collisions are observed and $\frac{1}{M}\left(\min_k \mu_k\right)$ if collisions are not observed (the sensing also needs to be 1 in that case). Hence, the average length of a sequence of *consecutive* time steps t in which $\left(A_{t-1}^m \in \widehat{M^m}(t-1), C_t^m = 1, s_t^m = 0\right)$ holds is upper bounded by M or $M/(\min_k \mu_k)$ when collisions are not observed. It remains to control the number of times in which such a sequence can *begin*, and to observe that if $\left(A_{t-1}^m \in \widehat{M^m}(t-1), C_t^m = 1, s_t^m = 0\right)$ holds at time t but not at time $t-1$, we must have $(A_{t-2}^m \notin \widehat{M^m}(t-2))$. Hence, one can again use Lemma 2.4 to upper bound the expected number of beginnings of such consecutive sequences.

□

Logarithmic regret Going back the the regret decomposition in Lemma 2.1, it remains to upper bound the middle term, that quantifies the under-selection of optimal arms. Fortunately, Lemma 7 in Besson and Kaufmann (2018) gives the following upper bound

$$\sum_{a=1}^{M}(\mu_a - \mu_M)(T - \mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]) \leq (\mu_\star - \mu_M)\left[\sum_{a=M+1}^{A} \mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] + \sum_{a=1}^{M} \mathbb{E}_{\boldsymbol{\mu}}[C_a(T)]\right]$$

which implies the following (crude) bound on the regret of any algorithm $\mathcal{A}$:

$$\mathcal{R}_{\boldsymbol{\mu}}\left(\mathcal{A}_M, T\right) \leq 2\left(\sum_{a=M+1}^{A} \mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] + \mathbb{E}_{\boldsymbol{\mu}}[C(T)]\right). \tag{2.3}$$

Hence, it follows from Theorem 2.2 (and a sum over players) and from Theorem 2.3 that the regret of MCTopM is logarithmic, under observation model **(II)** and observation model **(IV)**.

2.2.3 Empirical Evaluation

We report here a subset of the experimental results that can be found in Besson and Kaufmann (2018) to compare the performance of MCTopM and Selfish-kl-UCB with that of other algorithms. In Figure 2.1, we display the empirical regret as a function of time for several algorithms with provable sub-linear regret: RhoRand (Anandkumar et al., 2010) but also MEGA (Avner and Mannor, 2015) and Musical Chair (Rosenski et al., 2016). Musical Chair requires to specify the length T_0 of a uniform exploration phase in which each player estimates the set of M best arms from the sensing information. At the end of the exploration phase, players perform a musical chair protocol after which they end up on M distinct arms (with high probability). (Rosenski et al., 2016) propose a tuning of T_0 as a function of δ (and a lower bound on $\mu_M - \mu_{M+1}$) under which, w.p. $1-\delta$ the regret is in $\mathcal{O}(\log(1/\delta))$. Selecting $\delta = 1/T$ yields a provable logarithmic expected regret at time T. The three instances in the figure correspond to $\delta = 0.1, 0.5$ and $\delta = 1/T$.

We see that MCTopM and another variant that we propose in Besson and Kaufmann (2018), called RandTopM largely outperform MEGA and Musical Chair, and are also doing better than RhoRand. In order to measure the empirical cost of decentralized learning for our algorithms, we display the regret of a centralized (multiple-play) algorithm based on kl-UCB, which unsurprisingly performs better.

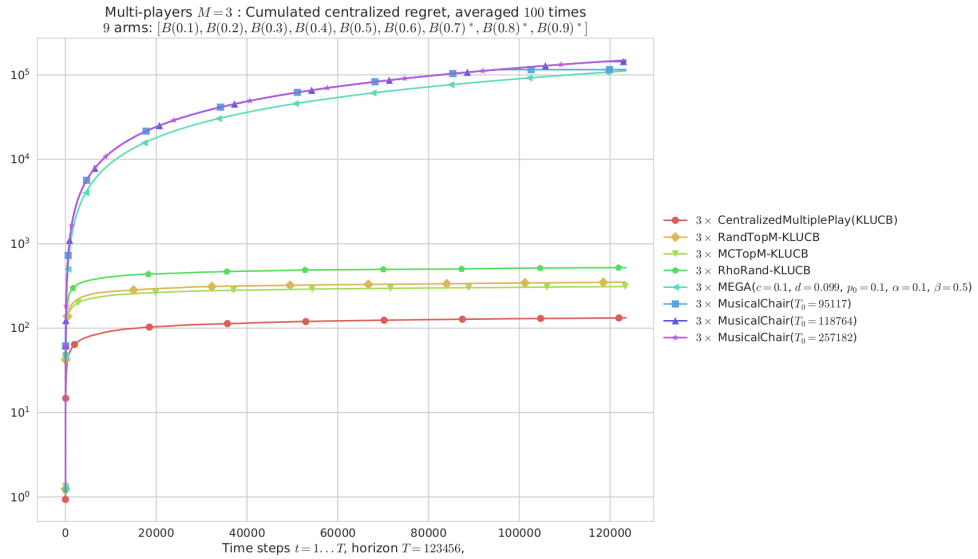


Figure 2.1 – Regret of several algorithms on a MP-MAB with $A = 9$ arms and $M = 6$ players estimated over $N = 100$ simulations

We then include the Selfish heuristic in our study. First, to measure its robustness we compare its regret to that of MCTopM, RandTopM and RhoRand averaged over $N = 500$ random Bernoulli MP-MAB instances with $A = 9$ arms and $M = 6$ players (which produces an estimate of the Bayesian regret, which is the regret averaged over a prior distribution on $\boldsymbol{\mu}$). The results, which are displayed in Figure 2.2 (left), show that Selfish is on average a bit worse than MCTopM and a bit better than RhoRand. This is promising as we recall that Selfish can also be used without the observation of the sensing information, which is crucial for other algorithms. However, we also put forward a drawback of Selfish: on some instances, it has a small probability to behave very badly. To illustrate this fact, in the right part of Figure 2.2 we display a histogram of the final regret pseudo-regret $R_T = T\sum_{a=1}^{M}\mu_{[a]} - \sum_{t=1}^{T}\sum_{m=1}^{M} R_t^m$ (such that $\mathcal{R}_{\boldsymbol{\mu}}(\text{Selfish-kl-UCB}, T) = \mathbb{E}_{\boldsymbol{\mu}}[R_T]$) for $T = 5000$ in a Bernoulli MP-MAB with $A = 3$ arms and $M = 2$ players: we see that in 17 out of 1000 simulations, the pseudo-regret of Selfish was of order 7000, whereas the pseudo-regret of other algorithms never exceeded 70.

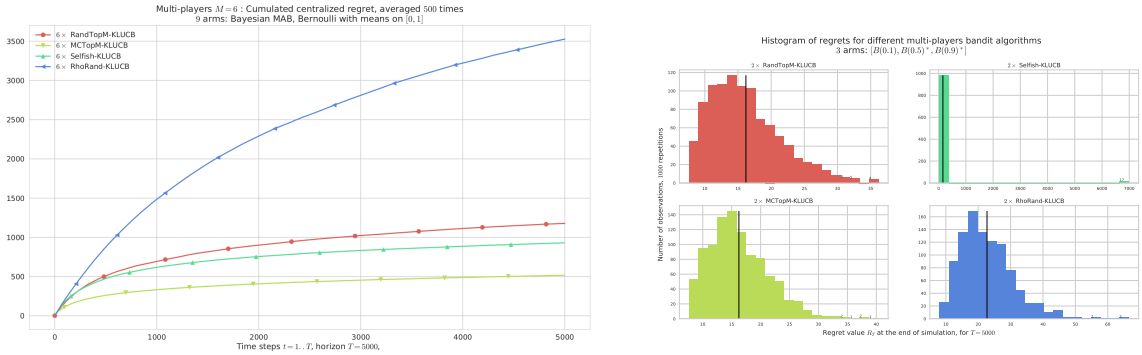


Figure 2.2 – Bayesian regret for $A = 9, M = 6$ estimated over $N = 500$ simulations (left) and histogram of R_T on a specific instance with $A = 3$ and $M = 2$ based on $N = 1000$ simulations (right)

2.3 Towards Optimal Multi-Player Algorithms

Our emphasis in this document is on *optimal* algorithms for different types of bandit problems. Hence, we naturally asked ourselves the question: is MCTopM optimal in some way?

Our final regret bound for MCTopM is actually quite loose, first because of the rough approximation (2.3) but mostly because of the large constant that multiplies $\log(T)$ in Theorem 2.3. Still, at the time our paper was published, this upper bound on the total number of collisions was improving upon the one available for RhoRand. But then Boursier and Perchet (2019) came up with an algorithm suffering from fewer collisions. Forgetting collisions, we thought for some time that MCTopM could achieve the minimal number of sub-optimal selections (the bound in Theorem 2.2 was also the smallest existing bound available). But this turned out not to be the case, as our claim was based on an erroneous lower bound. We explain below why one needs to be very careful with the usual change-of-distribution arguments in the context of a multi-player MAB, and what lower bound can still be derived using them.

Existing lower bounds A first observation is that the regret of a multi-player algorithm $\mathcal{A}_M$ is always larger than the regret of a centralized algorithm $\mathcal{A}'_M$ built from $\mathcal{A}_M$ by reassigning players that would experience a collision under $\mathcal{A}_M$ to a random subset of the arms that are still available. If $\mathcal{A}_M$ is uniformly efficient ($\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}_M, T) = o(T^\alpha)$ for all $\boldsymbol{\mu}$), so is $\mathcal{A}'_M$, hence the lower bound given by Anantharam et al. (1987) for the centralized multiple-play problem also applies for the MP-MAB.

Proposition 2.5. *Any uniformly efficient algorithm for the Bernoulli MP-MAB algorithm (under any of the observation models **(I)** - **(IV)**) satisfies*

$$\forall \boldsymbol{\mu},\ \liminf_{T\to\infty} \frac{\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}_M, T)}{\log(T)} \geq \sum_{a=M+1}^{A} \frac{1}{\mathrm{kl}(\mu_{[a]}, \mu_{[M]})}.$$

This lower bound may be too good to be achievable by a MP-MAB algorithm as there may be a *cost for decentralization*. At least this was believed until the work of Boursier and Perchet (2019), and was supported by two (wrong) larger lower bounds given by Liu and Zhao (2010) and by us in Besson and Kaufmann (2018). As usual, these lower bounds rely on change-of-measure arguments, but there is an extra difficulty in the multi-player model which makes the both aforementioned lower bounds wrong. We now explain what type of lower bounds can be obtained for multi-player algorithms when following our usual methodology.

Change-of-measure arguments for multi-player bandits We first observe that the regret is always lower bounded in terms of the number of sub-optimal selections:

$$\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}_M, T) \geq \sum_{a \in [A] \setminus \mathrm{TopM}} (\mu_{[M]} - \mu_a) \mathbb{E}_{\boldsymbol{\mu}}[N_a(T)].$$

This inequality is actually non-trivial, as one has to show that the sum of the second and third terms in the regret decomposition of Lemma 2.1 is always non-negative. This is done for example in Lemma 4 of Besson and Kaufmann (2018). Hence, a uniformly efficient algorithm also satisfies $\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] = o(T^\alpha)$ for all arms $a \notin \mathrm{TopM}(\boldsymbol{\mu})$, and thus $\mathbb{E}_{\boldsymbol{\mu}}[N_a^m(T)] = o(T^\alpha)$ for each player m.

Consider a multi-player algorithm $\mathcal{A}_M$ based on observation model **(IV)**. Using the vocabulary of the Introduction, the *information* available to each player m for selecting A_t^m under $\mathcal{A}_M$ is I_{t-1}^m where

$$I_t^m = (U_0^m, A_1^m, X_1^m, C_1^m, U_1^m \ldots, A_t^m, X_t^m, C_t^m, U_t^m).$$

with $U_t^m \sim \mathcal{U}([0,1])$ independent from all other variables. A_t^m is a deterministic function of I_{t-1}^m. Using similar tools as in the proof of Lemma 0.2 in the Introduction, one can prove that, for all alternative models $\boldsymbol{\lambda}$ such that $\mathrm{TopM}(\boldsymbol{\mu}) \neq \mathrm{TopM}(\boldsymbol{\lambda})$, for each player m,

$$\liminf_{T\to\infty} \frac{\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{I_T^m}, \mathbb{P}_{\boldsymbol{\lambda}}^{I_T^m}\right)}{\log(T)} \geq 1.$$

Now the difficulty comes from the computation of $\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{I_T^m}, \mathbb{P}_{\boldsymbol{\lambda}}^{I_T^m}\right)$ under the multi-player model in which the information features both sensing (which is random once the chosen arm is fixed) and collisions (which depend on other players). Using the chain rule, one can write

$$\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{I_T^m}, \mathbb{P}_{\boldsymbol{\lambda}}^{I_T^m}\right) = \mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{I_{T-1}^m}, \mathbb{P}_{\boldsymbol{\lambda}}^{I_{T-1}^m}\right) + \mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{X_T^m, C_T^m, U_T | I_{T-1}^m}, \mathbb{P}_{\boldsymbol{\lambda}}^{X_T^m, C_T^m, U_T | I_{T-1}^m}\right).$$

Now given I_{T-1}^m, the random variables X_t^m, C_t^m and U_t^m are independent and the distribution of U_t^m is the same under $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$. Hence one can write

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{I_T^m}, \mathbb{P}_{\boldsymbol{\lambda}}^{I_T^m}\right) &= \mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{I_{T-1}^m}, \mathbb{P}_{\boldsymbol{\lambda}}^{I_{T-1}^m}\right) + \mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{X_T^m | I_{T-1}^m}, \mathbb{P}_{\boldsymbol{\lambda}}^{X_T^m | I_{T-1}^m}\right) + \mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{C_T^m | I_{T-1}^m}, \mathbb{P}_{\boldsymbol{\lambda}}^{C_T^m | I_{T-1}^m}\right) \\
&= \mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{I_{T-1}^m}, \mathbb{P}_{\boldsymbol{\lambda}}^{I_{T-1}^m}\right) + \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{a=1}^{A} \mathbb{1}_{(A^m(T)=a)} \mathrm{kl}(\mu_a, \lambda_a)\right] + \mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{C_T^m | I_{T-1}^m}, \mathbb{P}_{\boldsymbol{\lambda}}^{C_T^m | I_{T-1}^m}\right) \\
&= \sum_{a=1}^{A} \mathbb{E}[N_a^m(T)] \mathrm{kl}(\mu_a, \lambda_a) + \underbrace{\sum_{t=1}^{T} \mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{C_T^m | I_{T-1}^m}, \mathbb{P}_{\boldsymbol{\lambda}}^{C_T^m | I_{T-1}^m}\right)}_{:= \mathcal{I}_{\boldsymbol{\mu},\boldsymbol{\lambda}}^m(\mathcal{A}, T)},
\end{aligned}$$

where the last line is obtained by induction. Compared to the standard bandit model, note the presence of an extra term that we denote by $\mathcal{I}_{\boldsymbol{\mu},\boldsymbol{\lambda}}^m(\mathcal{A}, T)$. This terms somehow quantifies the "collision information", and the lower bound given in Besson and Kaufmann (2018) relied on the wrong claim that $\mathcal{I}_{\boldsymbol{\mu},\boldsymbol{\lambda}}^m(\mathcal{A}, T) = 0$. However, $\mathbb{P}_{\boldsymbol{\mu}}(C_t^m = 1 | I_{t-1}^m)$ does depend on $\boldsymbol{\mu}$, as the probability that one of the $(M-1)$ other players selects arm A_t^m depends on their previous observation, which depends on the mean values.

Still, for an algorithm for which the collision information term is small, one can derive the following lower bound. Theorem 2.6 introduces a class of algorithms $\mathcal{C}$ that are doomed to have a regret at least M times larger than the regret of the best centralized algorithm.

Theorem 2.6. *Let $\mathcal{C}$ be the class of algorithms that satisfy $\mathcal{I}^m_{\boldsymbol{\mu},\boldsymbol{\lambda}}(\mathcal{A}_M, T) = o(\log(T))$ for all player m and all $\boldsymbol{\mu}, \boldsymbol{\lambda}$ with $\mathrm{TopM}(\boldsymbol{\mu}) \neq \mathrm{TopM}(\boldsymbol{\lambda})$. Any uniformly efficient algorithm $\mathcal{A}_M$ that belongs to this class $\mathcal{C}$ is such that for any player m and arm $a \in \overline{\mathrm{TopM}(\boldsymbol{\mu})}$*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a^m(T)]}{\log(T)} \geq \frac{1}{\mathrm{kl}(\mu_a, \mu_{[M]})},$$

which yields $\liminf_{T\to\infty} \frac{\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}_M,T)}{\log(T)} \geq M \times \left(\sum_{a=M+1}^{A} \frac{1}{\mathrm{kl}(\mu_{[a]},\mu_{[M]})}\right)$.

Proof. To fix the ideas, assume that $\boldsymbol{\mu}$ is such that $\mu_1 \geq \mu_2 \geq \ldots \mu_A$. For $\varepsilon > 0$, choosing $\boldsymbol{\lambda}$ such that

$$\begin{cases} \lambda_k &= \mu_k \quad \text{for all } k \neq a, \\ \lambda_a &= \mu_M + \varepsilon. \end{cases}$$

In this alternative model, $\mathrm{TopM}(\boldsymbol{\lambda}) = \{1, \ldots, M-1, a\} \neq \mathrm{TopM}(\boldsymbol{\lambda})$ and

$$\begin{aligned} \liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a^m(T)]\mathrm{kl}(\mu_a, \mu_M + \varepsilon)}{\log(T)} &= \liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a^m(T)]\mathrm{kl}(\mu_a, \mu_M + \varepsilon) + \mathcal{I}^m_{\boldsymbol{\mu},\boldsymbol{\lambda}}(\mathcal{A}, T)}{\log(T)} \\ &= \liminf_{T\to\infty} \frac{\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{I_T^m}, \mathbb{P}_{\boldsymbol{\lambda}}^{I_T^m}\right)}{\log(T)} \geq 1 \end{aligned}$$

Hence, for all $\varepsilon > 0$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a^m(T)]}{\log(T)} \geq \frac{1}{\mathrm{kl}(\mu_a, \mu_M + \varepsilon)}$$

and the conclusion follows by letting ε go to zero.

□

Closing the gap with algorithms exploiting collisions Under MCTopM, $\mathbb{E}_{\boldsymbol{\mu}}[N_a^m(T)]$ attains the lower bound given in Theorem 2.6, as proved in Theorem 2.2. However, we did not establish that MCTopM actually belongs to the class $\mathcal{C}$ of algorithms to which this lower bound applies, i.e. that $\mathcal{I}^m_{\boldsymbol{\mu},\boldsymbol{\lambda}}(\mathrm{MCTopM}, T) = o(\log(T))$. If it is the case, MCTopM would have the smallest possible number of sub-optimal selections among algorithms in the class $\mathcal{C}$.

Still, this is only a conjecture, and it actually rather motivates the search for algorithm outside the class $\mathcal{C}$, for which Theorem 2.6 does not hold and we have a hope to achieve a smaller regret and possibly match the lower bound of Proposition 2.5. An algorithm close to matching this lower bound was recently given by Boursier and Perchet (2019). The proposed algorithm, called SIC-MMAB relies on the interesting idea that the players may use *voluntary collisions as a means of implicit communication* (Synchronization Involves Communication). Players can transmit to each other sequences of bits using sequences of collisions/no collisions during some known time steps, which allows to share observation and speed up learning. We will explain this nice idea in more detail in the next section, as we build on it to propose a new algorithm for the *heterogeneous* MP-MAB problem.

Boursier and Perchet (2019) prove that under observation model **(III)** or **(IV)** (i.e. when the collision information is observed), the regret of SIC-MMAB is of order

$$\left(\sum_{a=M+1}^{A} \frac{1}{\mathrm{kl}(\mu_a, \mu_M)} + MA\right)\log(T) + o(\log(T)) \ .$$

Hence, it is matching the lower bound of Theorem 2.5 up to an extra $MA\log(T)$. This extra factor was recently removed by Proutière and Wang (2019), who propose a similar algorithm exploiting collisions that is asymptotically optimal, under observation models **(III)**-**(IV)**.

2.4 An Algorithm Exploiting Collisions for the Heterogeneous Case

The powerful idea of leveraging collisions for implicit communication can also be used for minimizing regret in the more challenging *heterogeneous* multi-player model, in which the mean reward of each arm varies across players. In this section, we present the M-ETC-Elim algorithm for this setting, which is based on a joint work with Abbas Mehrabian (who visited SequeL in 2018) and Etienne Boursier and Vianney Perchet who proposed some improvements to our initial algorithm (Boursier et al., 2020).

Regret in a heterogeneous MP-MAB model First, let us recall the heterogeneous version of the multi-player MAB. A stream of i.i.d. rewards $(X_{a,t}^m)_{t\in\mathbb{N}}$ which comes from some distribution with mean μ_a^m is associated to each arm a and each player m. In each round t, player m selects an arm A_t^m and observes a reward

$$R_t^m = X_m^t(1 - C_m^t)$$

with $X_m^t = X_{A^m,t}^m$ the sensing information and $C_m^t = \mathbb{1}\left(\exists m' \neq m : A_t^m = A_t^{m'}\right)$ the collision indicator. The goal is still to maximize the total reward $\mathbb{E}\left[\sum_{t=1}^T \sum_{m=1}^M R_t^m\right]$.

If the mean rewards μ_a^m were known and a central controller would assign arms to players, an oracle strategy would boil down to finding a maximum matching between players and arms. A *matching* is a one-to-one assignment of players to arms; formally, any one-to-one function $\pi : [M] \to [A]$ is a matching. The *utility* (or *weight*) of a matching π is defined as $U(\pi) := \sum_{m=1}^M \mu_{\pi(m)}^m$. We denote by $\mathcal{M}$ the set of all matchings and let $U^\star := \max_{\pi\in\mathcal{M}} U(\pi)$ denote the maximum attainable utility. A *maximum matching* (or *optimal matching)* is a matching with utility $U^\star$. The strategy maximizing the social utility of the players (i.e. the sum of all their rewards) would be to play according to a maximum matching in each round. The (expected) regret of an algorithm $\mathcal{A}_M$ with respect to that oracle is defined as

$$\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}_M, T) = TU^\star - \mathbb{E}\left[\sum_{t=1}^T \sum_{m=1}^M R_t^m\right].$$

We are interested in decentralized algorithms $\mathcal{A}_M = (\mathcal{A}^1, \dots, \mathcal{A}^M)$ that have a small regret. Under a decentralized algorithm, the arm selection strategy $\mathcal{A}^m$ for each player m can only leverage the past *observations* made by this player.

For observation model **(III)**, in which player m observes first the collision indicator C_t^m and then the reward R_t^m, we now propose an algorithm that has (quasi)-logarithmic regret.

The M-ETC-Elim algorithm Our algorithm operates under the extra assumption that rewards are bounded in $[0, 1]$, that is, we assume that $X_{a,t}^m \in [0, 1]$ for all a, m, t.

The algorithm relies on three ingredients, that are borrowed from Boursier and Perchet (2019):

- an **initialization phase**, after which players end up on M different arms, and get assigned M distinct ranks $\{1, \dots, M\}$. The distinct arms are the default *communication arms* of the players, while the rank determine in which order to perform communications.
- **exploration phases**, that are designed to be collision-free (provided that the initialization phase is successful), during which players select some arms on which they should gain information
- **communication phases**, that start simultaneously for all players and have a known pre-determined length L. In a communication phase, the default behavior of each player m is to pull her *communication arm* a_m during L time steps. It is crucial that the communication arms are all distinct. One player (say player i) will send another player (say player j) a sequence of bits of length L, $b_1, \dots, b_L$ using the following method: the receiving player, j, keeps selecting her communication arm a_j whereas the sending player, i, selects a_j if $b_i = 1$ and a_i if $b_i = 0$. By observing the

collisions that occurred on arm a_j, player j can reconstruct the sequence of 0's and 1's sent by player i, as no other player can have selected this arm. Figure 2.3 provides an illustration.

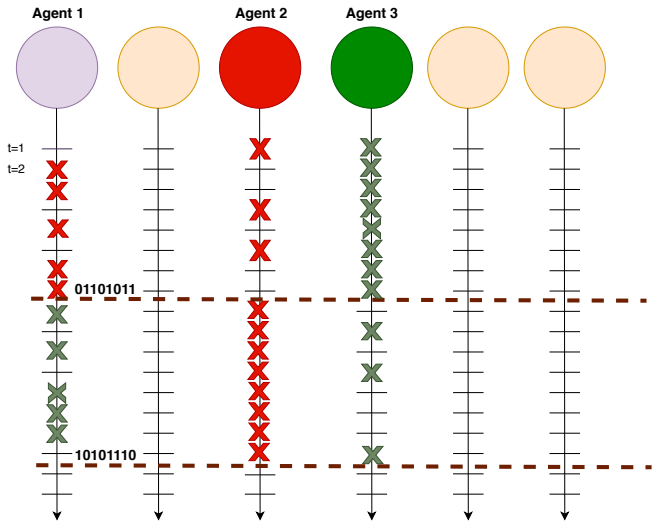


Figure 2.3 – Two successive communication phases of length $L = 8$ in a MP-MAB model with $M = 3$ agents and $A = 6$ arms: one from agent 2 (with communication arm 3) to agent 1 (with communication arm 1) followed by one from agent 3 (with communication arm 4) to agent 1

While in the original SIC-MMAB algorithm, designed for the homogeneous setting, each player sends her statistics for each arm to all the other players during each communication phase, we propose in M-ETC-Elim a leader/follower architecture in which the role of the players is asymmetric. The leader, i.e. the player that has rank 1 after the initialization phase, is responsible for aggregating all the estimates of the means (μ_a^m) and telling to the followers which arm to explore. These instructions are based on an efficient *matching elimination* procedure that we describe below.

Under M-ETC-Elim each player executes the algorithm described in Algorithm 7, which requires as input the number of arms A and the time horizon T.

Algorithm 7 M-ETC-Elim with parameter c

Input: Time horizon T, number of arms A
$\mathrm{R}, \mathrm{M} \leftarrow \textsc{Init}(A, 1/AT)$
if $\mathrm{R} = 1$ **then**
  LEADERALGORITHM(M,A,c)
else
  FOLLOWERALGORITHM(R,M,A,c)
end if

As an **initialization procedure** we use the one proposed by Boursier and Perchet (2019) which relies on a "musical chairs" phase after which the players end up on distinct arms, followed by a "sequential hopping" protocol that permits to know their ordering. Formally $\textsc{Init}(A, \delta_0)$ outputs for each player a rank $R \in [M]$ as well as the value of M, which is initially unknown to the players, with a probability of accuracy of that is at least $1 - \delta_0$.

Lemma 2.7. *With probability at least $1 - \delta_0$, if the M players run the $\textsc{Init}(A, \delta_0)$ procedure, which takes less than $A \log(A/\delta_0) + 2A - 2$ time steps, all players obtain a distinct ranking from 1 to M.*

The leader and follower algorithms, described below, rely on several **communication phases** as described above. To minimize regret, each player uses as a communication arm her arm in the best matching found so far. In the first communication phase, such an assignment is unknown and players simply use their ranking as communication arm. The rankings are also useful to know *in which order communications are performed*, as the leader successively communicates messages to the $M-1$ followers, and then the $M-1$ followers successively communicate messages to the leader.

Leader and Follower Algorithms Consider a bipartite graph with parts of size M and A, where the edge (m,a) has weight μ_a^m and associates player m to arm k. The weights μ_a^m are unknown to the players, but the leader maintains a set of *estimated* weights that are sent to her by the followers, and approximate the real weights. The goal of these algorithms is for the players to jointly explore the matchings in this graph, while gradually focusing on better and better matchings. For this purpose, the leader maintains a set of *candidate edges* $\mathcal{E}$, which is initially $[M]\times[A]$, that can be seen as edges that are potentially contained in optimal matchings, and gradually refines this set by performing eliminations, based on the information obtained from the exploration phases and shared during communication phases.

M-ETC-Elim proceeds in epochs whose length is parameterized by an integer $c \geq 1$. In epoch $p = 1,2,\dots$, the leader weights the edges using the estimated weights. Then for every edge $(m,a) \in \mathcal{E}$, the leader computes an associated matching $\widetilde{\pi}_p^{m,a}$ defined as the estimated maximum matching containing the edge (m,a). This computation can be done in polynomial time using, e.g., the Hungarian algorithm (Munkres, 1957). The leader then computes the utility of the maximum matching and eliminates from $\mathcal{E}$ any edge for which the weight of its associated matching is smaller by at least $4M\varepsilon_p$, where

$$\varepsilon_p := \sqrt{\frac{\log(2/\delta)}{2^{1+p^c}}}, \text{ with } \delta := \frac{1}{M^2AT^2}. \tag{2.4}$$

The leader then forms the set of associated candidate matchings $\mathcal{C} := \{\widetilde{\pi}_p^{m,a}, (m,a) \in \mathcal{E}\}$ and communicates to each follower the list of arms to explore in these matchings. Then exploration begins, in which for each candidate matching every player pulls its assigned arm 2^{p^c} times and records the received reward. Then another communication phase begins, during which each follower sends her observed estimated mean for the arms to the leader. More precisely, for each explored arm, the follower truncates the estimated mean (a number in $[0,1]$) and sends only the $\frac{p^c+1}{2}$ most significant bits of this number to the leader. The leader updates the estimated weights and everyone proceeds to the next epoch.

If at some point the list of candidate matchings $\mathcal{C}$ becomes a singleton, it means that (with high probability) the actual maximum matching is unique and has been found; so all players jointly pull that matching for the rest of the game, that is, the players enter an **exploitation phase**. Note that in the presence of several optimal matchings, the players will not enter the exploitation phase but will keep exploring several optimal matchings, which still ensures small regret. On the contrary, in the presence of a unique optimal matching, they are guaranteed to eventually enter the exploitation phase. Also, observe that the set $\mathcal{C}$ of candidate optimal matchings does not necessarily contain *all* potentially optimal matchings, but all the edges in those matchings remain in $\mathcal{E}$ and are guaranteed to be explored.

The pseudo-code of the Leader and Follower algorithms are given in Algorithm 8 and 9 respectively. In these pseudo-code, `comm` refers to a call to the communication protocol. In Algorithm 8, we specify for each call to the communication protocol the identity of the communicating players, as well as the length L of the message. In Algorithm 9, we omit to mention that each corresponding call to `comm` lasts $(M-1)\times L$ time steps, divided in $(M-1)$ communication slots, and the player with rank R is a communicating player in the $(R-1)$-th slot.

Algorithm 8 LEADERALGORITHM(M,A,c)

Input: Number of players M, number of arms A, parameter c
$\mathcal{E} \leftarrow [M] \times [A]$ `# list of candidate edges`
$\widetilde{\mu}_a^m \leftarrow 0$ for all $(m,a) \in [M] \times [A]$ `# empirical estimates for utilities`
for $p = 1, 2, \ldots$ **do**
  $\mathcal{C} \leftarrow \varnothing$ `# list of associated matchings`
  $\pi_1 \leftarrow \text{argmax} \left\{ \sum_{n=1}^{M} \widetilde{\mu}_{\pi(n)}^n : \pi \in \mathcal{M} \right\}$ `(Hungarian algorithm)`
  for $(m,a) \in \mathcal{E}$ **do**
    $\pi \leftarrow \text{argmax} \left\{ \sum_{n=1}^{M} \widetilde{\mu}_{\pi(n)}^n : \pi(m) = a \right\}$ `(Hungarian algorithm)`
    if $\sum_{n=1}^{M} \left\{ \widetilde{\mu}_{\pi_1(n)}^n - \widetilde{\mu}_{\pi(n)}^n \right\} \leq 4M\varepsilon_p$ **then**
      add π to $\mathcal{C}$
    else
      remove (m,a) from $\mathcal{E}$
    end if
  end for
  for $m = 2, \ldots, M$ **do**
    Send to player m the value of size($\mathcal{C}$) `(comm)` $1 \to m, L = \lceil \log_2(MA) \rceil$
    for $i = 1, 2, \ldots, \text{size}(\mathcal{C})$ **do**
      Send to player m the arm associated to player m in $\mathcal{C}[i]$ `(comm)` $1 \to m, L = \lceil \log_2(A) \rceil$
    end for
    Send to player m her communication arm $\pi_1(m)$ `(comm)` $1 \to m, L = \lceil \log_2(A) \rceil$
    Send to player m the communication arm of the leader $\pi_1(1)$ `(comm)` $1 \to m, L = \lceil \log_2(A) \rceil$
  end for
  if size($\mathcal{C}$) = 1 **then**
    `# enter the exploitation phase`
    pull for the rest of the game the arm assigned to player 1 in the unique matching in $\mathcal{C}$
  end if
  for $i = 1, 2, \ldots, \text{size}(\mathcal{C})$ **do**
    `# exploration`
    pull 2^{p^c} times the arm assigned to player 1 in the matching $\mathcal{C}[i]$
  end for
  for $a = 1, 2, \ldots, A$ **do**
    $\widetilde{\mu}_a^1 \leftarrow$ empirically estimated utility of arm a if it was pulled in this epoch, 0 otherwise
  end for
  for $m = 1, 2, \ldots, M$ **do**
    for $a = 1, 2, \ldots, A$ **do**
      Receive from player m the value $\widetilde{\mu}_a^m$ `(comm)` $m \to 1, L = \frac{p^c+1}{2}$
    end for
  end for
end for

Algorithm 9 FOLLOWERALGORITHM(R,M,A,c)

Input: Ranking R, number of players M, parameter c
for $p = 1, 2, \ldots$ **do**
  Receive the value of size($\mathcal{C}$) (comm)
  for $i = 1, 2, \ldots,$ size($\mathcal{C}$) **do**
    Receive the arm assigned to this player in $\mathcal{C}[i]$ (comm)
  end for
  Receive the new communication arm (comm)
  Receive the communication arm of the leader (comm)
  if size($\mathcal{C}$) = 1 **then**
    # enter exploitation phase
    Pull for the rest of the game the arm assigned to this player in the unique matching in $\mathcal{C}$
  end if
  for $i = 1, 2, \ldots,$ size($\mathcal{C}$) **do**
    # exploration
    Pull 2^{p^c} times the arm assigned to this player in the matching $\mathcal{C}[i]$
  end for
  for $a = 1, 2, \ldots, A$ **do**
    $\widehat{\mu}_a^R \longleftarrow$ empirically estimated utility of arm a if arm a has been pulled in this epoch, 0 otherwise
    Truncate $\widehat{\mu}_a^R$ to $\widetilde{\mu}_a^R$ using the $\frac{p^c+1}{2}$ most significant bits
  end for
  Send the values $\widetilde{\mu}_1^R, \widetilde{\mu}_2^R, \ldots, \widetilde{\mu}_A^R$ to the leader $A\times$ (comm)
end for

Regret analysis of M-ETC-Elim We now present a regret analysis of M-ETC-Elim which permits to derive problem-dependent regret bounds that feature the gap of each matching π, defined as $\Delta(\pi) := U^\star - U(\pi)$. In Theorem 2.8, we present the scaling of these problem-dependent bounds in terms of the smallest gap defined as $\Delta := \inf_{\pi:\Delta(\pi)>0} \Delta(\pi)$. Observe that $\Delta > 0$ even in the presence of several optimal matchings.

Theorem 2.8. *Assume that* $\Delta < \infty$ [4]. *For every integer parameter* $c \in \{1, 2, \ldots\}$, M-ETC-Elim *satisfies*

$$\mathcal{R}_{\boldsymbol{\mu}}(\text{M-ETC-Elim}, T) = \mathcal{O}\left(MA\left(\frac{M^2 \log(T)}{\Delta}\right)^{1+\frac{1}{c}}\right).$$

Furthermore, if the optimal matching is unique, M-ETC-Elim *with parameter* $c = 1$ *satisfies*

$$\mathcal{R}_{\boldsymbol{\mu}}(\text{M-ETC-Elim}, T) = \mathcal{O}\left(\frac{M^3 A \log(T)}{\Delta}\right).$$

The heterogeneous multi-player multi-armed bandit was first studied by Bistritz and Leshem (2018), who proposed the Game-Of-Thrones (GOT) algorithm and proved a $\mathcal{O}(\log^2(T))$ regret upper bound for it, leaving as an open question as to whether one can get closer to the $\Omega(\log(T))$ lower bound. The motivation of our work was to try to answer this question, and the second statement in Theorem 2.8 does answer it positively in the presence of a unique optimal matching. For multiple optimal matchings, we

4. This excludes the degenerate case in which all matchings have the same utility, for which we still propose bounds in Boursier et al. (2020).

prove that there exists an algorithm with a nearly-logarithmic regret: for every $\kappa > 0$, M-ETC-Elim run with parameter $c = \lfloor 1/\kappa \rfloor$ has a $\mathcal{O}(\log^{1+\kappa}(T))$ regret.

As can be seen in the proof sketch given below, the result for multiple optimal matchings is quite asymptotic in nature (it holds for T larger than some constant $T_0(c)$ which can be very large) and is therefore mostly of theoretical interest. However, the improved result obtained for a unique optimal matching (which comes from the fact that the players are guaranteed to enter an exploitation phase after a controlled number of epochs in that case) doesn't suffer from this drawback. For M-ETC-Elim with $c = 1$ we further provide in Boursier et al. (2020) a problem-independent $\mathcal{O}(M^3\sqrt{AT\log(T)})$ regret bound that holds whether or not the optimal matching is unique, and recommend the use of this parameter tuning in practice, which outperforms GOT in our experiments.

In parallel to our work, several authors also improved the $\log^2(T)$ upper bound of Bistritz and Leshem (2018). First, in an updated preprint, Bistritz and Leshem (2019) propose a new analysis of GOT (with slightly modified phase lengths) which also achieves $\mathcal{O}(\log^{1+\kappa}(T))$ for every $\kappa > 0$ (with a worse, less explicit scaling in Δ). Then, Tibrewal et al. (2019) independently studied a slightly different model in which each player in each round has the option of "observing whether a given arm has been pulled by someone," without actually pulling that arm (thus avoiding collision due to this "observation"). Due to the stronger feedback, communications do not need to be implicitly done through collisions and bits can be broadcast to other players via this operation. Still, algorithms for this alternative feedback model can be modified to obtain algorithms for the MP-MAB. The two algorithms proposed by Tibrewal et al. (2019) share similarities with M-ETC-Elim: they also have exploration, communication and exploitation phases, but they do not use eliminations. Tibrewal et al. (2019) also obtain logarithmic regret in the presence of a unique optimal matching (with a slightly worse dependency in $1/\Delta$). Moreover, we note that the idea of "implicit communication" is also present in this work, that was done independently from Boursier and Perchet (2019).

We conclude this discussion with a few words about *optimality* for the heterogeneous multi-player MAB. The $\Omega(\log(T))$ lower bound proven by Bistritz and Leshem (2018) hides the dependency in the problem parameter, hence it is hard to know whether the bounds in Theorem 2.8 have a good scaling in A, M and Δ. However, just like in the homogeneous case, these bounds can be compared to the existing bounds for an easier centralized version of the problem, in which a central controller would choose one matching from players to arms in every round. This setting is a particular instance of a combinatorial semi-bandit problem (Gai et al., 2012). Audibert et al. (2014) provide a minimax $\Omega(\sqrt{MAT})$ lower bound for combinatorial bandits, while Combes et al. (2015) give a problem dependent lower bound of the form $c(\boldsymbol{\mu}, M)\log(T)$ and show that $c(\boldsymbol{\mu}, M) = \Theta(A/\Delta)$ for many common combinatorial structures, including matchings. This tells us that the dependency in M in our bounds may not be optimal. However, we note that the popular CUCB algorithm for combinatorial bandits (Chen et al., 2013) has a regret upper bound that scales in $\mathcal{O}\left((M^2A/\Delta)\log(T)\right)$ (Kveton et al., 2015), which is only a factor M smaller the the bound obtained in Theorem 2.8 for a unique optimal matching.

Sketch of proof of Theorem 2.8. We let $\lg(T) = \log_2(T)$ to ease the notation.

Let $\mathcal{C}_p$ and $\mathcal{E}_p$ denote the set of candidate matchings and candidate edges used in epoch p, and for each matching π let $\tilde{U}_p(\pi)$ be the utility of π that the leader can estimate based on the information received by the end of epoch p. Let $\hat{p}_T$ be the total number of epochs before the (possible) start of the exploitation phase. As $2^{\hat{p}_T^c} \leq T$, we have $\hat{p}_T \leq \lg(T)$. Recall that a successful initialization means all players identify M and their ranks are distinct. Define the *good event*

$$\mathcal{G}_T := \left\{\text{INIT}(A, 1/AT) \text{ is successful and } \forall p \leq \hat{p}_T, \forall \pi \in \mathcal{C}_{p+1}, |\tilde{U}_p(\pi) - U(\pi)| \leq 2M\varepsilon_p\right\}.$$

During epoch p, for each candidate edge (m, a), player m has pulled arm a at least 2^{p^c} times and the

quantization error is smaller than ε_p. Hoeffding's inequality and a union bound over at most $\lg(T)$ epochs yield that $\mathcal{G}_T$ holds with large probability.

Lemma 2.9. $\mathbb{P}(\mathcal{G}_T) \geq 1 - \frac{2}{MT}$.

We now upper bound the pseudo-regret $R_T = \sum_{t=1}^{T} \left(U^\star - \sum_{m=1}^{M} \mu_{A_t^m}^m (1 - C_t^m) \right)$ when the event $\mathcal{G}_T$ holds, by separately upper bounding the contributions from the different phases of the algorithm. We also use sometimes that the pseudo-regret incurred in each time step t is (crudely) upper bounded by M.

- **initialization phase**: from Lemma 2.7, the initialization lasts $A\log(A^2T) + 2A - 2$ times steps, which contributes $R_{\text{INIT}} = \mathcal{O}(MA\log(A^2T))$ to the pseudo-regret.
- **communication phases**: in each epoch p, the leader first communicate to the $M-1$ other players 1. one message of length $\lg(MA)$ (the size of $\mathcal{C}_p$) 2. at most MA messages of length $\lg(A)$ (the list of arms to explore) and 3. two messages of length $\lg(A)$ (the communicating arms). This takes a total of $(M-1)\left[(MA+2)\lg(A) + \lg(MA)\right]$ time steps. Then each follower communicates to the leader A messages of length $(1+p^c)/2$, which uses a total of $(M-1)A(1+p^c)/2$ times steps. Summing over epochs yields a total contribution to the pseudo-regret of $R_{\text{COMM}} = \mathcal{O}(M^3 A \lg(A)\hat{p}_T + M^2 A(\hat{p}_T)^{c+1})$.
- **exploitation phase**: on the event $\mathcal{G}_T$, $\mathcal{C}_p$ always contains an optimal matching, hence if the players enter an exploitation phase, they suffer zero regret in it: $R_{\text{EXPLOIT}} = 0$.

During **exploration phases**, the players always play arms in a matching, hence no collision occurs. The contribution to the pseudo regret of the exploration phase in epoch p is

$$\sum_{\pi \in \mathcal{C}_p} \Delta(\pi) 2^{p^c} = \sum_{(m,a) \in \mathcal{E}_p} \widetilde{\Delta}_p^{m,a} 2^{p^c},$$

where $\widetilde{\Delta}_p^{m,a} = U^\star - U(\tilde{\pi}_p^{m,a})$ is the gap of the matching associated to the edge (m,a) in epoch p. Now for each edge (m,a), we let $\pi^{m,a}$ be the best sub-optimal matching containing that edge:

$$\pi^{m,a} := \text{argmax}\left\{U(\pi) : \pi(m) = a, U(\pi) < U^\star\right\}.$$

We further introduce, for any matching π, the quantity

$$P(\pi) := \inf\{p \in \mathbb{N}^\star : 8M\varepsilon_p < \Delta(\pi)\}$$

On $\mathcal{G}_T$, if $p > P(\pi^{m,a})$ either (m,a) does not belong to $\mathcal{E}_p$ or its associated matching $\tilde{\pi}_p^{m,a}$ is an optimal matching, i.e. $\widetilde{\Delta}_p^{m,a} = 0$. Moreover, if $(m,a) \in \mathcal{E}_p$ for $p \leq P(\pi^{m,a})$, it holds that $\widetilde{\Delta}_m^{m,a} \leq \frac{\varepsilon_{p-1}}{\varepsilon_{P(\pi^{m,a})}} \Delta(\pi^{m,a})$ (which follows by combining the fact that $\widetilde{\Delta}_m^{m,a} \leq 8M\epsilon_{p-1}$, otherwise $(m,a) \notin \mathcal{E}_p$ and $\Delta(\pi^{m,a}) > 8M\epsilon_{P(\pi^{m,a})}$). Finally, the total pseudo-regret due to exploration phases is upper bounded by

$$R_{\text{EXPLORE}} = \sum_{(m,a) \in [M]\times[A]} \sum_{p=1}^{P(\pi^{m,a})} \frac{\varepsilon_{p-1}}{\varepsilon_{P(\pi^{m,a})}} \Delta\left(\pi^{m,a}\right) 2^{p^c}.$$

The regret of M-ETC-Elim is upper bounded as

$$\begin{aligned} \mathcal{R}_{\boldsymbol{\mu}}(\text{M-ETC-Elim}, T) &\leq \mathbb{E}\left[R_T \mathbb{1}(G_T)\right] + MT\mathbb{P}\left(\overline{\mathcal{G}_T}\right) \\ &\leq R_{\text{INIT}} + R_{\text{EXPLORE}} + R_{\text{COMM}} + R_{\text{EXPLOIT}} + 2 \end{aligned}$$

From there, the conclusion is mostly technical and follows from upper bounds on R_{COMM} and R_{EXPLORE}, which respectively require to upper bound $\hat{p}_T$ and the sum of $\Delta\left(\pi^{m,a}\right) 2^{p^c}$. An instrumental result for this purpose is Lemma 6 from Boursier et al. (2020), that we re-state below.

Lemma 2.10. *If T is larger than $T_0(c) := \exp\left(2^{\frac{c^c}{\log^c(1+\frac{1}{2c})}}\right)$, for any matching π,*

$$\begin{aligned}\Delta(\pi)2^{P(\pi)^c} &\leq \left(32M^2\log(2M^2AT^2)/\Delta(\pi)\right)^{1+\frac{1}{c}} \text{ if } c>1,\\ \Delta(\pi)2^{P(\pi)} &\leq 64M^2\log(2M^2AT^2)/\Delta(\pi) \text{ if } c=1.\end{aligned}$$

Moreover, $2^c \leq 2\lg(\log(T))$, $\hat{p}_T \leq 2(\lg T)^{1/c}$ and $(\hat{p}_T)^c \leq e\lg T$.

The improved result for $c = 1$ in the presence of a unique optimal matching follows the improved bound in Lemma 2.10 for $c = 1$ but also from a tighter bound on $\hat{p}_T$. Indeed, when the optimal matching is unique, on $\mathcal{G}_T$ one is guaranteed to enter an exploration phase when the second best matching π_2 satisfies $\Delta(\pi_2) > 8M\epsilon_p$, which leads to

$$\hat{p}_T \leq \lg\left(\frac{64M^2\log(2M^2AT^2)}{\Delta^2}\right).$$

□

To summarize, we presented an analysis of two different algorithms, MCTopM and M-ETC-Elim for the homogeneous and heterogeneous multi-player multi-armed bandit, respectively. The two algorithms also rely on different observation models: MCTopM requires the observation of the sensing information followed by the reward (observation model **(II)**) while M-ETC-Elim requires the observation of the collision information followed by the reward (observation model **(III)**). While the two algorithms achieve (nearly) logarithmic regret, they have not been proved to be optimal, that is, they do not exactly match existing lower bounds.

In the homogeneous setting, we explained that when the collisions are observed (under models **(III)-(IV)**) an asymptotically optimal algorithm has been found (Proutière and Wang, 2019), whose regret is similar to that of the best centralized algorithm. However, to the best of my knowledge, there is no algorithm matching the lower bound of Proposition 2.5 under observation models **(I)-(II)**, despite the fact that they are particularly relevant for applications to IoT communications or Opportunistic Spectrum Access. For learning based on rewards only (i.e., under model **(I)**), the best known logarithmic regret bounds scale in $(MA)\log(T)/(\mu_{[M]} - \mu_{[M+1]})^2$ (Lugosi and Mehrabian, 2018) and $\left(M\sum_{a>M}(\mu_{[M]} - \mu_{[a]})^{-1} + MA^2/\mu_{[A]}\right)\log(T)$ (Boursier and Perchet, 2019). Under observation model **(II)**, to the best of my knowledge MCTopM enjoys the smallest regret bound among algorithms analyzed in this setting, but is probably far from optimal. Hence, investigating the minimal regret under observation **(I)-(II)** remains a crucial open question.

As for the heterogeneous setting, it seems algorithms have only been proposed under the assumption that collisions can be observed, and it would be interesting to study the heterogeneous MP-MAB model under more challenging observation models.

Part II

Active Identification Problems

Chapter 3

A Universal Stopping Rule for Active Identification

Over the past years, I have worked on different examples of *pure exploration problems*, in which the samples collected by the agent in a bandit model are no longer perceived as rewards, but instead some *decision* has to be made as quickly as possible.

In this chapter, I introduce a general framework called Active Identification that encompasses these different problems, and present a single stopping rule, which leads to correct decision with high probability, whatever the sampling rule. This stopping rule was introduced under different names in different papers, but in this document I settle for the Parallel GLRT test. After presenting this stopping rule, I also highlight the technical tools needed to prove its correctness, namely time-uniform deviation inequalities.

3.1 Active Identification in a Bandit Model

As in the rest of this document, we consider a bandit model with A arms, parameterized by its vector of means $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_A)$, typically an exponential family bandit model. The set of possible means $\boldsymbol{\mu}$ belongs to some known region $\mathcal{R} \subseteq \mathcal{I}^A$, a learner seeks to *discover something* about the unknown vector $\boldsymbol{\mu}$, as quickly and accurately as possible.

More precisely, given I regions $\mathcal{R}_1, \dots, \mathcal{R}_I$ that form a covering of the region $\mathcal{R}$, i.e., such that $\mathcal{R} \subseteq \bigcup_{k=1}^{I} \mathcal{R}_i$, an agent wants to identify a region that contains $\boldsymbol{\mu}$. In order to gain information to identify such a region, she interacts with the bandit model as usual, by selecting an arm A_t at time t and observing an independent *sample* X_t from the distribution with mean μ_{A_t}. Note that this sample is no longer called a reward. After enough observations, the agent can make a confident *recommendation* of a region.

A strategy for active identification in a bandit model consists of three components. Letting $(\mathcal{F}_t)_{t\in\mathbb{N}^*}$ be the filtration generated by the observations collected by the agent up to time t (and some possible exogenous randomness), a strategy is composed of

- a *sampling rule* A_t which is $\mathcal{F}_{t-1}$ measurable,
- a *stopping rule* τ which is a stopping time with respect to $\mathcal{F}_t$ that indicates when the agent stops collecting observations
- a *recommendation rule* $\hat{\imath}_\tau$ which outputs a guess for a region to which $\boldsymbol{\mu}$ belongs

In words, the objective of the agent is to make a confident recommendation (i.e. that is correct with high probability) after seeing as few samples as possible.

After presenting a few motivating examples, we will introduce several mathematical formalizations of this objective that have been studied, notably the *fixed-confidence setting* that will be our focus.

3.1.1 Examples

In the bandit literature, the first and most studied active identification problem is (ε)-Best Arm Identification (BAI), in which the goal is to identify the arm with largest mean.

(Exact) Best Arm Identification In this context, the region $\mathcal{R}$ is either the set of all possible mean vectors $\mathcal{R} = \mathcal{I}^A$ or may be restricted to the set of vectors that have a unique optimal arm

$$\mathcal{R} = \left\{ \boldsymbol{\mu} \in \mathcal{I}^A : \exists a \in [A] : \mu_a > \max_{b \neq a} \mu_b \right\}.$$

There are $I = A$ regions with $\mathcal{R}_i$ being the set of models in which arm i is optimal:

$$\mathcal{R}_i = \left\{ \boldsymbol{\mu} \in \mathcal{R} : \mu_i > \max_{a \neq i} \mu_a \right\}.$$

If $\mathcal{R}$ is the set of vectors with a unique optimal arm, note that $\mathcal{R}_1, \dots, \mathcal{R}_A$ form a partition of $\mathcal{R}$ and one seeks to identify to which fold of this partition the true vector of means $\boldsymbol{\mu}$ belongs.

ε - Best Arm Identification The following relaxation of the BAI problem, that depends on a parameter $\varepsilon \in (0, 1)$ has also been studied a lot: we set $\mathcal{R} = \mathcal{I}^A$ and for all $i \in [A]$,

$$\mathcal{R}_i = \left\{ \boldsymbol{\mu} \in \mathcal{R} : \mu_i \geq \max_{a \neq i} \mu_a - \varepsilon \right\}.$$

In this setting, the goal is to identify one arm whose mean is at most ε away from that of the optimal arm. In that case, note that the $\mathcal{R}_i$ no longer form a partition of $\mathcal{R}$.

Why BAI? The question of finding the distribution with largest mean among a pool of distributions (possibly with some kind of adaptive sampling) is an old question in statistics, studied since the 1950s under the name ranking and selection (Bechhofer, 1954; Bechhofer et al., 1968). It was revisited in the bandit literature since the works of Even-Dar et al. (2006); Audibert et al. (2010); Bubeck et al. (2011).

A typical application of Best Arm Identification problems is A/B(/C) testing, which is a process often used in e-commerce in which the goal is to assess the impact that different versions of the same webpage have (e.g. different layout as in the naive example of Figure 3.1) on the *conversion probability*. A conversion is some target event the company wants to enforce such as visitors buying products, spending time on the website, creating an account, subscribing to a mailing list, etc.

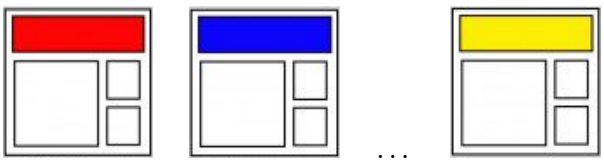


Figure 3.1 – Conversion probabilities for different versions of a website

In order to identify which version of its webpage has the largest conversion probability, the company implements a *testing phase* in which the different versions are shown to visitors and the conversion indicators are observed. This testing phase is called A/B Testing when two versions are compared, and

A/B/C Testing when the test includes 3 versions or more. A default strategy for the testing phase is to allocate each version to n users, when n is chosen in advance as a function of the minimum improvement we want with respect to a known baseline. Instead, viewing A/B testing as a best arm identification problem allows for fully-sequential A/B testing: for each visitor t, the company may look at the past data to decide which version A_t should be displayed, observe a possible conversion X_t and may also decide if the test should be stopped, without specifying a sample size or a fraction of allocation in advance. As the conversion probabilities are typically small and may be quite close to each other, it may be relevant to consider the ε-relaxation of the BAI problem to speed-up the test.

If the objective of a company can be phrased as maximizing its total number of conversions (e.g. when the conversions correspond to sales), one can also discuss the relevance of such a testing phase, in which rewards (conversions) are ignored. Indeed, regret minimization algorithms are tailored to maximize the total number of conversions, and could be more suitable in this setting. So why not using them? A possible answer could be practical: using (e.g.) a UCB algorithm would require the company to keep alternating between a pool of versions of its website, which can be costly (a computation needs to be performed for each visitor to decide which version should be displayed). Performing a short exploration (testing) phase followed by an exploitation phase in which the best version is displayed for a long time is more convenient. In the paper Garivier et al. (2016b), we investigate how sub-optimal this approach that decouples exploration and exploitation can be, if the overall goal is to maximize rewards.

Beyond Best Arm Identification In the past few years, there has been some interest in more complex pure-exploration problems that share with the best arm identification problem the absence of an incentive to maximize rewards, but in which one should answer a more complex question about the means $\boldsymbol{\mu}$ than "what is the largest component in $\boldsymbol{\mu}$?".

The general framework presented in this document is close to the adaptive sequential testing framework that we introduced in Kaufmann and Koolen (2018), which is similar to the General-Samp problem of Chen et al. (2017) and to the best partition identification problem of Juneja and Krishnasamy (2019). In these three works, the regions $\mathcal{R}_i$ are supposed to form a partition of $\mathcal{R}$, whereas Degenne and Koolen (2019) consider possible overlapping regions. Finally, the structured best arm identification framework of Huang et al. (2017) is a special case of this framework.

To mention a few concrete examples one can first go back to the **dose finding** problem discussed in Chapter 1, in which each arm a models the toxicity of a dose a, and the goal is to find the dose whose toxicity is the closest to some threshold θ (the Maximum Tolerated Dose). Assuming increasing toxicities leads us to define

$$\mathcal{R} = \left\{\boldsymbol{\mu} \in [0,1]^A : \forall a \in [A-1], \mu_a \leq \mu_{a+1}\right\}$$

and the set of means for which a particular dose i is the MTD is given by

$$\mathcal{R}_i = \left\{\boldsymbol{\mu} \in \mathcal{R} : |\mu_i - \theta| = \min_a |\mu_a - \theta|\right\}.$$

This particular active identification problem was studied by Garivier et al. (2019a).

As another example we can mention **anomaly detection**: assume that we are monitoring A different processes each having a probability μ_a to have a regular behavior. On may decide to trigger an intervention if one is convinced that there exists a process with mean μ_a that is smaller than some (small) threshold γ. The corresponding active identification problem sets $\mathcal{R} = [0,1]^A$ and

$$\mathcal{R}_1 = \left\{\boldsymbol{\mu} \in \mathcal{R} : \min_a \mu_a < \gamma\right\} \text{ and } \mathcal{R}_2 = \left\{\boldsymbol{\mu} \in \mathcal{R} : \min_a \mu_a \geq \gamma\right\}.$$

We will discuss this example in Chapter 5, in which it is motivated by a different story, related to the design of Monte-Carlo Tree Search algorithms.

3.1.2 Several mathematical frameworks

Several performance measures have been introduced in the bandit literature for assessing the quality of a best arm identification strategy. Most of them can be naturally extended to strategies for active identification (also called algorithms), that we recall consist of a triple $(A_t, \tau, \hat{\imath}_\tau)$. They seek to achieve different trade-offs between the number of samples needed before making a recommendation, τ, and the probability of error of this recommendation, $\mathbb{P}_{\boldsymbol{\mu}}(\boldsymbol{\mu} \notin \mathcal{R}_{\hat{\imath}_\tau})$.

In the *fixed-budget setting*, the total number of samples τ cannot exceed a given budget T (i.e. $\tau = T$), and the goal is to minimize the error probability $\mathbb{P}_{\boldsymbol{\mu}}(\boldsymbol{\mu} \notin \mathcal{R}_{\hat{\imath}_\tau})$. Fixed-budget strategies are allowed to use the knowledge of T.

In the *fixed-confidence setting*, given a risk parameter $\delta \in (0,1)$, the error probability $\mathbb{P}_{\boldsymbol{\mu}}(\boldsymbol{\mu} \notin \mathcal{R}_{\hat{\imath}_\tau})$ has to remain smaller than δ for every $\boldsymbol{\mu} \in \mathcal{R}$, and the goal is to minimize the (random) number of samples τ needed to make the recommendation, either in expectation or with high probability. Fixed-confidence strategies use the knowledge of δ in the stopping rule and possibly in the sampling rule.

In the *anytime exploration* framework, instead of issuing a recommendation after enough exploration, a recommendation $I_t \in [I]$ has to be made at the end of every round (I_t is $\mathcal{F}_t$-measurable), which can be seen as the best guess for an hypothesis if exploration were to be stopped after t rounds. The goal is to minimize, *for all* t, the error probability $\mathbb{P}_{\boldsymbol{\mu}}(\boldsymbol{\mu} \notin \mathcal{R}_{I_t})$. Unlike in the previous frameworks, a strategy for anytime exploration does not depend on the knowledge of a maximal exploration budget (T) or a maximal tolerated error probability (δ) and is therefore called *anytime*.

The three different objectives are summarized in the following table.

	Fixed-budget	Fixed-confidence	Anytime Exploration
Parameter	budget T	risk δ	none
Constraint to satisfy	$\tau \leq T$	$\mathbb{P}_{\boldsymbol{\mu}}(\boldsymbol{\mu} \notin \mathcal{R}_{\hat{\imath}_\tau}) \leq \delta$	none
Objective to minimize	$\mathbb{P}_{\boldsymbol{\mu}}(\boldsymbol{\mu} \notin \mathcal{R}_{\hat{\imath}_\tau})$	$\mathbb{E}_{\boldsymbol{\mu}}[\tau]$	$\mathbb{P}_{\boldsymbol{\mu}}(\boldsymbol{\mu} \notin \mathcal{R}_{I_t})$, for all t

The fixed-budget setting was first introduced for the best arm identification problem by Audibert et al. (2010) and the fixed confidence setting was first studied for ε-best arm identification by Even-Dar et al. (2006). The anytime exploration framework was studied by Bubeck et al. (2011) and Jun and Nowak (2016) for best arm identification and top-m arms identification respectively. In the context of BAI specifically, instead of minimizing the probability of error, an alternative goal can be minimizing the so-called *simple regret* (Bubeck et al., 2011), $r_t = \mathbb{E}[\mu_\star - \mu_{I_t}]$, which quantifies how far from optimal the proposed candidate for the best arm is. Finally, we will mention in Chapter 4 another Bayesian performance measure, recently introduced by Russo (2016).

After describing these three different frameworks, a natural question is whether one can *convert* algorithms from one framework to another. A strategy for anytime exploration can naturally be used in the fixed-budget setting (setting $\hat{\imath}_\tau = I_\tau$), but it may not attain the minimal possible error for a particular budget T. Similarly, such a strategy may be used in the fixed-confidence setting when coupled with a good stopping rule, but upper bounding its sample complexity τ may be tricky. In the other direction, the sampling rule A_t of a fixed-budget or fixed-confidence algorithm may be used directly for anytime exploration (when coupled with a good recommendation rule I_t) only if it is *anytime*, that is independent of the budget T or the risk parameter δ. Still, its analysis may be tricky in the anytime exploration setting. As for "conversions" from the fixed-budget to the fixed-confidence setting (discussed for example in Gabillon et al. (2012); Kaufmann and Kalyanakrishnan (2013)), they are only possible when some complexity constant characterizing the bandit instance (depending on the unknown means) is known.

In part II of this document, we will present active identification strategies for the fixed-confidence setting. In the rest of this chapter, we will first present a general stopping rule, that may be used for any

active identification problem. Then in Chapter 4 and 5, we will further study the *sample complexity* τ of some active identification algorithms. We will in particular propose *anytime* sampling rules A_t that, used in conjunction with our stopping rule, lead to an *asymptotically optimal* sample complexity.

3.2 The Parallel GLRT Stopping Rule

In this section, we propose a general stopping rule for active identification in a bandit model in the fixed confidence setting. Recall that given I regions

$$\mathcal{R}_1 \quad \mathcal{R}_2 \quad \dots \quad \mathcal{R}_I,$$

the goal is to identify *one* region to which $\boldsymbol{\mu} \in \mathcal{R}$ belongs. For this purpose, we use a sampling rule $(A_t)_{t\in\mathbb{N}}$ to collect data from the different arms: at each time step t, we collect $X_t \sim \nu_{\mu_{A_t}}$. We stop collecting information after a random number of samples τ and output a recommendation $\hat{\imath}_\tau \in [I]$. This recommendation should satisfy

$$\forall \boldsymbol{\mu} \in \mathcal{R}, \quad \mathbb{P}_{\boldsymbol{\mu}}\left(\tau < \infty, \boldsymbol{\mu} \notin \mathcal{R}_{\hat{\imath}_\tau}\right) \leq \delta.$$

A strategy satisfying this property is called δ-*correct*.

Assuming that the sampling rule $(A_t)_{t\in\mathbb{N}}$ is given and that we do not try to optimize it, the two other components of the identification strategy, $(\tau, \hat{\imath}_\tau)$ can be viewed as a *sequential test* of multiple, composite hypotheses $\mathcal{H}_1 : (\boldsymbol{\mu} \in \mathcal{R}_1), \dots, \mathcal{H}_I : (\boldsymbol{\mu} \in \mathcal{R}_I)$. For the resulting strategy to be δ-correct, the different types of testing error should be uniformly controlled.

Sequential testing was first studied by Wald (1945), who proposed the Sequential Probability Ratio Test (SPRT) for two simple hypotheses, e.g. $\mathcal{H}_1 : (\mu = \mu_1)$ and $\mathcal{H}_2 : (\mu = \mu_2)$ when collecting i.i.d. samples of a distribution with mean μ. Among the sequential tests with prescribed type I and type II errors, the SPRT is proved to have the smallest average duration $\mathbb{E}_{\mu_i}[\tau]$ for $i \in \{1, 2\}$. Later, particular examples of sequential tests of *composite* hypotheses (in which $\mathcal{R}_1$ and $\mathcal{R}_2$ are not reduced to a singleton) have also been studied, see e.g. Robbins and Siegmund (1974); Lai (1988). Our framework is more general as we allow for more than two hypotheses, that are possibly *overlapping*, when the regions $\mathcal{R}_i$ do not form a partition of $\mathcal{R}$.

3.2.1 Definition of the Parallel GLRT

The idea of the Parallel Generalized Likelihood Ratio Test (GLRT) is to run in parallel I sequential tests of the following two non-overlapping hypotheses

$$\widetilde{\mathcal{H}}_0 : (\mu \in \mathcal{R}\backslash\mathcal{R}_i) \quad \text{against} \quad \widetilde{\mathcal{H}}_1 : (\mu \in \mathcal{R}_i),$$

for each $i \in \{1, \dots, I\}$. The Parallel GLRT stops when one of these tests rejects $\widetilde{\mathcal{H}}_0$. The i-th test causing to stop means that μ is believed to belong to $\mathcal{R}_i$ (this does not exclude that it may also belong to other regions), in which case we set $\hat{\imath}_\tau = i$.

To test $\widetilde{\mathcal{H}}_0$ against $\widetilde{\mathcal{H}}_1$, we propose to use a Generalized Likelihood Ratio Test, which is a well-known extension of the standard Likelihood Ratio Test used for simple hypotheses (Wilks, 1938). We denote by $\ell(X_1, \dots, X_t; \boldsymbol{\lambda})$ the likelihood of the first t observation that are collected in a bandit model parameterized by $\boldsymbol{\lambda} \in \mathcal{R}$, under some sampling rule A_t (that is, $X_t \sim \nu_{\lambda_{A_t}}$). The Generalized Likelihood Ratio statistic based on t samples is defined as

$$\frac{\max_{\boldsymbol{\lambda}\in\mathcal{R}} \ell(X_1, \dots, X_t; \boldsymbol{\lambda})}{\max_{\boldsymbol{\lambda}\in\mathcal{R}\backslash\mathcal{R}_i} \ell(X_1, \dots, X_t; \boldsymbol{\lambda})} = \inf_{\boldsymbol{\lambda}\in\mathcal{R}\backslash\mathcal{R}_i} \frac{\ell(X_1, \dots, X_t; \hat{\boldsymbol{\mu}}(t))}{\ell(X_1, \dots, X_t; \boldsymbol{\lambda})},$$

where $\hat{\boldsymbol{\mu}}(t)$ is the maximum likelihood estimator (in $\mathcal{R}$). Large values of this GLR statistic tend to reject $\widetilde{\mathcal{H}}_0$. Calibrating the rejection threshold for a GLR based on a fixed sample size is often done by resorting to asymptotic arguments (like Wilks' phenomenon (Wilks, 1938)) describing the limit distribution of the GLR under the null hypothesis; this is however not useful for the finite-confidence bandit analysis that follows. We propose to use a threshold function $\beta(t,\delta)$ that depends on the current number of samples t and on the risk parameter δ. In Section 3.3 we will provide a possible choice for the threshold $\beta(t,\delta)$ that guarantees the δ-correctness of the corresponding test.

The parallel GLRT using the threshold function $\beta(t,\delta)$ is formally defined in the following way. Given $\delta \in (0,1)$, the stopping rule is

$$\tau_\delta = \inf\left\{ t \in \mathbb{N} : \max_{i=1,\dots,I} \inf_{\boldsymbol{\lambda}\in\mathcal{R}\backslash\mathcal{R}_i} \log \frac{\ell(X_1,\dots,X_t;\hat{\boldsymbol{\mu}}(t))}{\ell(X_1,\dots,X_t;\boldsymbol{\lambda})} > \beta(t,\delta) \right\} \tag{3.1}$$

and the decision rule is

$$\hat{\imath}_{\tau_\delta} \in \operatorname*{argmax}_{i=1,\dots,I} \inf_{\boldsymbol{\lambda}\in\mathcal{R}\backslash\mathcal{R}_i} \log \frac{\ell(X_1,\dots,X_{\tau_\delta};\hat{\boldsymbol{\mu}}(\tau_\delta))}{\ell(X_1,\dots,X_{\tau_\delta};\boldsymbol{\lambda})} \tag{3.2}$$

(ties can be resolved arbitrarily). Note that the maximum over $i \in \{1,\dots,I\}$ in the definition of τ_δ and $\hat{\imath}_{\tau_\delta}$ can be reduced to the set of hypotheses to which $\hat{\boldsymbol{\mu}}(t)$ belongs.

3.2.2 Simple Examples

Exponential family bandit model If we go back to our recurrent example of exponential family bandit models, the likelihood of the observation under a vector of means $\boldsymbol{\lambda}$ can be written

$$\ell(X_1,\dots,X_t;\boldsymbol{\lambda}) \propto \prod_{s=1}^{t} \exp\left(\dot{b}^{-1}(\lambda_{A_s})X_s - b(\dot{b}^{-1}(\lambda_{A_s}))\right),$$

where b is the log-partition function of the exponential family (see the Introduction or Notation Index for the notation related to exponential families). Letting $\hat{\boldsymbol{\mu}}(t) = (\hat{\mu}_1(t),\dots,\hat{\mu}_K(t))$ be the vector of empirical means of the arms, for every $\boldsymbol{\lambda} = (\lambda_1,\dots,\lambda_K) \in \mathcal{R}$ the log-likelihood ratio can be written

$$\begin{aligned}\log \frac{\ell(X_1,\dots,X_t;\hat{\boldsymbol{\mu}}(t))}{\ell(X_1,\dots,X_t;\boldsymbol{\lambda})} &= \sum_{a=1}^{A} N_a(t)\Big[\hat{\mu}_a(t)\left(\dot{b}^{-1}(\hat{\mu}_a(t)) - \dot{b}^{-1}(\lambda_a)\right) - b\left(\dot{b}^{-1}(\hat{\mu}_a(t))\right) + b\left(\dot{b}^{-1}(\lambda_a)\right)\Big] \\ &= \sum_{a=1}^{A} N_a(t) d\left(\hat{\mu}_a(t),\lambda_a\right) \ , \end{aligned} \tag{3.3}$$

where we use the closed-form expression of the Kullback-Leibler divergence in an exponential family that can be found in Equation (1) in the Introduction.

Hence, the Parallel GLRT can be expressed as

$$\begin{aligned}\tau_\delta &= \inf\left\{ t \in \mathbb{N} : \max_{i\in[I]} \inf_{\lambda\in\mathcal{R}\backslash\mathcal{R}_i} \sum_{a=1}^{A} N_a(t) d\left(\hat{\mu}_a(t),\lambda_a\right) > \beta(t,\delta) \right\} \ , \\ \hat{\imath}_{\tau_\delta} &= \operatorname*{argmax}_{i\in[I]} \inf_{\lambda\in\mathcal{R}\backslash\mathcal{R}_i} \sum_{a=1}^{A} N_a(\tau_\delta) d\left(\hat{\mu}_a(\tau_\delta),\lambda_a\right) \ . \end{aligned} \tag{3.4}$$

This expression still requires to compute a minimizer over $\mathcal{R}\backslash\mathcal{R}_i$ which can be solved in closed form for some particular choice of regions, as can be seen in the next example or in those of the next chapters.

Testing two overlapping hypotheses We now present a simple example of a sequential test between two overlapping hypotheses. Given a stream of independent samples $X_1, X_2, \dots$ that follow a $\mathcal{N}(\mu, \sigma^2)$ distribution, with a known variance σ^2, the statistical problem is to determine whether μ is positive or negative, but with an indifference region of width 2ε. We consider the two hypotheses

$$\mathcal{H}_1 : (\mu < \varepsilon) \quad \text{and} \quad \mathcal{H}_2 : (\mu > -\varepsilon) \, .$$

The hypotheses $\mathcal{H}_1$ and $\mathcal{H}_2$ are not mutually exclusive. In this particular example, one aims at building a stopping rule τ_δ and a recommendation rule $\hat{\imath}_{\tau_\delta}$ such that

$$\forall \mu \leq -\varepsilon, \ \mathbb{P}_\mu(\hat{\imath}_{\tau_\delta} = 2) \leq \delta \quad \text{and} \quad \forall \mu \geq \varepsilon, \ \mathbb{P}_\mu(\hat{\imath}_{\tau_\delta} = 1) \leq \delta \, , \tag{3.5}$$

but any answer $\hat{\imath}_{\tau_\delta}$ is considered correct when $\mu \in \mathcal{R}_1 \cap \mathcal{R}_2 = (-\varepsilon, \varepsilon)$.

This testing problem can be seen as a particular case of active identification in an exponential family bandit model with a single, Gaussian arm with mean $\boldsymbol{\mu} = \mu$, in which the two regions are $\mathcal{R}_1 = (-\infty, \varepsilon)$ and $\mathcal{R}_2 = (-\varepsilon; +\infty)$. With $d(\mu, \mu') = (\mu - \mu')^2/(2\sigma^2)$ and this particular choice of regions, the expression in (3.4) can be made more explicit and the Parallel GLRT test becomes

$$\begin{aligned} \tau_\delta &= \inf\left\{ t \in \mathbb{N} : \frac{t(|\hat{\mu}_t| + \varepsilon)^2}{2\sigma^2} > \beta(t, \delta) \right\} \, , \\ \hat{\imath}_{\tau_\delta} &= 2 \ \text{ if and only if } (\hat{\mu}_{\tau_\delta} > 0), \end{aligned}$$

where $\hat{\mu}_t = \frac{1}{t}\sum_{s=1}^t X_s$ is the empirical mean of the observation. The optimality of this test in discussed in Garivier and Kaufmann (2019).

We now explain how to choose the threshold $\beta(t, \delta)$ to guarantee δ-correctness for general active identification problems in an exponential family bandit model.

3.3 Correctness of the Parallel GLRT stopping rule

In this section we introduce a new deviation inequality to prove the δ-correctness of the Parallel GLRT stopping rule for active identification in an exponential family bandit model, for *any* sampling rule. We then provide a proof of this result, which sheds light on some interesting martingale tools.

3.3.1 A New Deviation Inequality

An expression of the Parallel GLRT $(\tau_\delta, \hat{\imath}_{\tau_\delta})$ for active identification in an exponential family bandit model was given in (3.4). Based on this expression, a generic argument to upper bound the error probability is the following:

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\mu}}\left(\tau_\delta < \infty, \boldsymbol{\mu} \notin \mathcal{R}_{\hat{\imath}_{\tau_\delta}}\right) &\leq \mathbb{P}\left(\exists t \in \mathbb{N}^*, \exists i : \boldsymbol{\mu} \notin \mathcal{R}_i, \inf_{\boldsymbol{\lambda} \in \mathcal{R} \backslash \mathcal{R}_i} \sum_{a=1}^A N_a(t) d(\hat{\mu}_a(t), \lambda_a) > \beta(t, \delta) \right) \\ &\leq \mathbb{P}\left(\exists t \in \mathbb{N}^*, \exists i : \boldsymbol{\mu} \in \mathcal{R} \backslash \mathcal{R}_i, \sum_{a=1}^A N_a(t) d(\hat{\mu}_a(t), \mu_a) > \beta(t, \delta) \right) \\ &\leq \mathbb{P}\left(\exists t \in \mathbb{N}^*, \sum_{a=1}^A N_a(t) d(\hat{\mu}_a(t), \mu_a) > \beta(t, \delta) \right). \end{aligned} \tag{3.6}$$

To upper bound this last probability, one needs a concentration inequality in which:

- the deviation of the empirical means from their true values are *measured with the KL-divergence* $d(\cdot, \cdot)$ (just like in the inequality of Lemma 0.3 in the Introduction which is instrumental for the analysis of kl-UCB)

- the deviations are *uniform over time* ($t \in \mathbb{N}$)
- the deviations are measured *simultaneously for the A arms* (the *sum* of $N_a(t)d(\hat{\mu}_a(t), \mu_a)$'s need to be controlled)

In a joint work with Wouter Koolen (Kaufmann and Koolen, 2018), we propose a new concentration inequality which fulfills these three requirements, which is stated in Theorem 3.1 below. It features a non-explicit function $\mathcal{T}$ defined by

$$\mathcal{T}(x) = 2\tilde{h}\left(\frac{h^{-1}(1+x) + \log(2\zeta(2))}{2}\right) \tag{3.7}$$

where ζ is the Riemann zeta function and the functions h and $\tilde{h}$ are defined as follows. For $u \geq 1$, we let $h(u) = u - \log u$. As h is a one-to-one mapping from $[1, +\infty)$ to $[1, +\infty)$ its inverse function h^{-1} is well defined and for any $x \geq 0$, we let

$$\tilde{h}(x) = \begin{cases} e^{1/h^{-1}(x)} h^{-1}(x) & \text{if } x \geq h^{-1}(1/\log(3/2)), \\ (3/2)(x - \log\log(3/2)) & \text{otherwise.} \end{cases} \tag{3.8}$$

The function $\mathcal{T}$ is easy to compute numerically and we show in Kaufmann and Koolen (2018) that it satisfies $\mathcal{T}(x) \simeq x + 4\log\left(1 + x + \sqrt{2x}\right)$ for $x \geq 5$ and $\mathcal{T}(x) \sim x$ when x is large.

Theorem 3.1. *Let $\boldsymbol{\mu}$ be an exponential family bandit model. Under* any *sampling rule (A_t), for every subset $\mathcal{S} \subseteq [A]$,*

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : \sum_{a \in \mathcal{S}} N_a(t) d(\hat{\mu}_a(t), \mu_a) \geq 3 \sum_{a \in \mathcal{S}} \log(1 + \log N_a(t)) + |\mathcal{S}| \mathcal{T}\left(\frac{x}{|\mathcal{S}|}\right)\right) \leq e^{-x} .$$

The proof of this inequality, given below, relies on the construction of a particular *mixture martingale*. For a subset $\mathcal{S}$ of size 1 and Gaussian distributions, this method can be traced back to the work of Robbins (1970) and similar inequalities were also given by Jamieson et al. (2014); Kaufmann et al. (2016). In this particular case, observe that the presence of $\log(1 + \log(t))$ in the right-hand side of the inequality is essentially tight due to the law of iterated logarithm. For $|\mathcal{S}| > 1$ and general exponential families, the only other inequality of this flavor is the one proposed by Magureanu et al. (2014), which holds for t in some times range $\{1, \dots, n\}$ instead of $t \in \mathbb{N}^*$, and whose proof does not rely on mixture martingales. The so-called method of mixtures for proving different types of deviations inequalities has been popularized by De La Pena et al. (2004) and used notably by Abbasi-Yadkori et al. (2011); Balsmubramani (2015).

Applying Theorem 3.1 Combining (3.6) and Theorem 3.1, one can see that with the threshold

$$\beta(t, \delta) = A\mathcal{T}\left(\frac{\log(1/\delta)}{A}\right) + 3A\log(1 + \log(t))$$

the Parallel GLRT satisfies $\mathbb{P}_{\boldsymbol{\mu}}\left(\tau_\delta < \infty, \boldsymbol{\mu} \notin \mathcal{R}_{\hat{\imath}_{\tau_\delta}}\right) \leq \delta$ for all $\boldsymbol{\mu} \in \mathcal{R}$ and is therefore δ-correct for any active identification problem in an exponential family bandit model. This threshold has the drawback of not having a simple closed-form, but it essentially scales in $\beta(t, \delta) \simeq \log(1/\delta) + 3A\log(1 + \log(t))$.

For some specific active identification problems, this general argument can be replaced by more specific arguments that may justify the use of different, possibly smaller thresholds. Typically, A in the expression of $\beta(t, \delta)$ may be replaced by a smaller quantity called the rank of the identification problem (see Section 6.2. of Kaufmann and Koolen (2018)). In words, the rank is R if $\mathcal{R} \backslash \mathcal{R}_i$ can be written as a finite union of sets that are each defined in terms of only $R < A$ arms. For example, for BAI,

$$\left\{\boldsymbol{\mu} : \mu_i \leq \max_a \mu_a\right\} = \bigcup_{a \neq i} \left\{\boldsymbol{\mu} : \mu_a \geq \mu_i\right\},$$

hence the best arm identification problem is of rank 2, which licentiates the use of the threshold $\beta(t,\delta) = 2\mathcal{T}\left(\frac{\log(1/\delta)}{2}\right) + 6\log(1+\log(t))$. Using different concentration arguments specific to Bernoulli distributions (see Theorem 10 in Garivier and Kaufmann (2016)), we also proved that in that case the more explicit threshold $\beta(t,\delta) = \log\left(\frac{2t(K-1)}{\delta}\right)$ can be used. Observe that for large values of t, the former threshold is guaranteed to be smaller than the latter, and it also justifies a threshold calibration often used in practice for Best Arm Identification: $\beta(t,\delta) = \log\left(\frac{\log(1+\log(t))}{\delta}\right)$.

3.3.2 Proof of Theorem 3.1: A Martingale Story

Concentration inequalities play a central role in the analysis of all kind of bandit algorithms, and in this proof we give an example of the use of *mixture martingales* to establish a time uniform, self-normalized inequality that furthermore aggregates information across arms.

Why martingales? We start by giving a quick recap about why martingales (or super-martingales) are a powerful tool for proving time-uniform deviation inequalities. The reason is due to the following property, that we refer to as the *maximal inequality* for super-martingales. This result is also known under the name *Ville's inequality* as according to Shafer et al. (2011) it was already proved by Ville (1939) for the particular case of martingales.

Lemma 3.2. *Let S_t be a super-martingale (i.e. a sequence of random variables adapted to a filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$ such that $\mathbb{E}[S_{t+1}|\mathcal{F}_t] \le S_t$) such that $S_t \ge 0$ and $\mathbb{E}[S_0] = 1$. For all $\delta \in (0,1)$,*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t > 1/\delta\right) \le \delta.$$

Proof. In probability textbooks, Doob's maximal inequality is often stated for martingales or *sub*-martingales, saying that for all $c > 0$,

$$\forall T \in \mathbb{N}^*, \quad \mathbb{P}\left(\sup_{t\in\{0,\dots,T\}} S_t^+ > c\right) \le \frac{\mathbb{E}[S_T^+]}{c}.$$

However, one can also prove a counterpart of this result if S_t is a non-negative *super*-martingale:

$$\forall T \in \mathbb{N}^*, \quad \mathbb{P}\left(\sup_{t\in\{0,\dots,T\}} S_t > c\right) \le \frac{\mathbb{E}[S_0]}{c}. \tag{3.9}$$

This inequality permits to prove the maximal inequality in Lemma 3.2 as

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t > 1/\delta\right) = \lim_{T\to\infty} \mathbb{P}\left(\sup_{t\in\{0,\dots,T\}} S_t > 1/\delta\right) \le \delta\mathbb{E}[S_0] = \delta.$$

We now prove (3.9). Introducing the stopping time $\tau = \inf\{t \in \mathbb{N} : S_t > c\}$, the sequence $(S_{\tau\wedge T})_{T\in\mathbb{N}^*}$ is also a super-martingale, in particular $\mathbb{E}[S_{\tau\wedge T}] \le \mathbb{E}[S_0]$. But it also holds that

$$\begin{aligned}
\mathbb{E}[S_{\tau\wedge T}] &= \mathbb{E}[S_\tau \mathbb{1}(\tau \le T)] + \mathbb{E}[S_T \mathbb{1}(\tau > T)] \\
&\ge \mathbb{E}[S_\tau \mathbb{1}(\tau \le T)] \\
&\ge c\mathbb{P}(\tau \le T)
\end{aligned}$$

which concludes the proof as $\{\tau \le T\} = \left\{\sup_{t\in\{0,\dots,T\}} S_t > c\right\}$.

□

Lemma 3.2 is very useful to build *sequential* tests or confidence intervals that are *uniform over time*. Indeed, assume that for a standard hypothesis testing problem with two hypotheses H_0 and H_1 one can find a sequence S_t which is a (super) martingale under H_0 satisfying the properties of Lemma 3.2, i.e. S_t is non-negative and $\mathbb{E}[S_0] = 1$. Then the sequential test which rejects H_0 at time

$$\tau = \inf\{t \in \mathbb{N} : S_t > 1/\delta\}$$

has a type I error which is upper bounded by δ. This explains why non-negative martingales with expectation 1 are sometimes called *test martingales* (Shafer et al., 2011). More broadly, in order to prove deviation inequalities, which allow to construct confidence intervals, a possible strategy consists in *upper-bounding* the process that we want to control *by a test (super) martingale*.

The proof of Theorem 3.1 follows this path, and combines this idea with the Cramér-Chernoff method. In Theorem 3.1, the process whose deviations should be controlled is

$$X_a(t) := N_a(t) d\left(\hat{\mu}_a(t), \mu_a\right) - 3\log(1 + \log(N_a(t))).$$

The technique that we propose to prove deviation inequalities for sums of these processes is summarized in the following lemma.

Lemma 3.3. *Let $g : \Lambda \to \mathbb{R}$ be a function defined on a non-empty interval $\Lambda \subseteq \mathbb{R}$. Assume that*

1. *For any arm a and $\lambda \in \Lambda$ there exists a test martingale $M_a^\lambda(t)$ such that*

$$\forall t \in \mathbb{N}, \quad M_a^\lambda(t) \geq e^{\lambda X_a(t) - g(\lambda)}. \tag{$*$}$$

2. *For any subset $\mathcal{S} \subseteq \{1, \dots, K\}$ and for any $\lambda \in \Lambda$, the product $\prod_{a \in \mathcal{S}} M_a^\lambda(t)$ is a martingale.*

Letting

$$C^g(x) := \min_{\lambda \in \Lambda} \frac{g(\lambda) + x}{\lambda},$$

for any subset $\mathcal{S}$, for all $x > 0$,

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a \in \mathcal{S}} X_a(t) \geq |\mathcal{S}| C^g\left(\frac{x}{|\mathcal{S}|}\right)\right) \leq e^{-x}.$$

Proof. Fix $\lambda \in \Lambda$. For any $u \in \mathbb{R}$, one can write

$$\begin{aligned}
\mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a \in \mathcal{S}} X_a(t) > u\right) &= \mathbb{P}\left(\exists t \in \mathbb{N} : e^{\lambda[\sum_{a \in \mathcal{S}} X_a(t)]} > e^{\lambda u}\right) \\
&\leq \mathbb{P}\left(\exists t \in \mathbb{N} : \prod_{a \in \mathcal{S}} M_a^\lambda(t) > e^{\lambda u - |\mathcal{S}| g(\lambda)}\right) \\
&\leq e^{-[\lambda u - |\mathcal{S}| g(\lambda)]},
\end{aligned}$$

where the last step uses the maximal inequality in Lemma 3.2 applied to the product $\prod_{a \in \mathcal{S}} M_a^\lambda(t)$ which is a martingale by assumption 2.. Equivalently, it also holds that for all $x > 0$, for all $\lambda \in \Lambda$,

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a \in \mathcal{S}} X_a(t) > \frac{|\mathcal{S}| g(\lambda) + x}{\lambda}\right) \leq e^{-x}.$$

The conclusion follows from selecting λ that yields the tightest possible inequality.

□

The proof of Theorem 3.1 consists in building test martingales that satisfy assumptions 1. and 2. in Lemma 3.3 for an adequate function g. More precisely, we prove the following.

Lemma 3.4. *Fix $\xi \in [0, 1/2]$ and define for all $\lambda \in \Lambda_\xi := [0, 1/(1+\xi))$,*

$$g_\xi(\lambda) = \lambda(1+\xi)\log\left(C(\xi)\right) - \log(1-\lambda(1+\xi)) \quad \textit{with} \quad C(\xi) = \frac{2\zeta(2)}{(\log(1+\xi))^2}.$$

There exists martingales satisfying assumptions 1. and 2. in Lemma 3.3 for $g_\xi : \Lambda_\xi \to \mathbb{R}$.

Before giving the proof of Lemma 3.4, we explain how to obtain Theorem 3.1 from it. From Lemma 3.3, we know that for every $\xi > 0$, for all $x > 0$,

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a \in \mathcal{S}} X_a(t) \geq |\mathcal{S}| C^{g_\xi}\left(\frac{x}{|\mathcal{S}|}\right)\right) \leq e^{-x}.$$

where $C^{g_\xi}(x) := \min_{\lambda \in [0,1/(1+\xi))} \frac{g_\xi(\lambda)+x}{\lambda}$. To obtain the result, it remains to optimize over the possible choice for ξ (which depends on x), that leads to

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a \in \mathcal{S}} X_a(t) \geq |\mathcal{S}| \mathcal{T}\left(\frac{x}{|\mathcal{S}|}\right)\right) \leq e^{-x},$$

with

$$\mathcal{T}(x) = \inf_{\substack{\xi \in [0,1/2] \\ \lambda \in [0,(1+\xi)^{-1})}} \frac{g_\xi(\lambda)+x}{\lambda}.$$

In order to make $\mathcal{T}$ explicit, we recall the definition of the function $h(u) = u - \log(u)$ for $u > 1$ and its inverse function h^{-1} and we write

$$\begin{aligned}
\mathcal{T}(x) &= \inf_{\substack{\xi \in [0,1/2] \\ \lambda \in [0,(1+\xi)^{-1})}} \left[(1+\xi)\log(C(\xi)) + \frac{x - \log(1-\lambda(1+\xi))}{\lambda}\right] \\
&= \inf_{\substack{z \in [1,3/2] \\ \lambda \in [0,1/z)}} \left[z\left(\log\left(\frac{2\zeta(2)}{\log^2(z)}\right) + \frac{x-\log(1-\lambda z)}{\lambda z}\right)\right] \\
&= \inf_{z \in [1,3/2]} \left[z\left(\log\left(\frac{2\zeta(2)}{\log^2(z)}\right) + \inf_{q \in [0,1)} \frac{x-\log(1-q)}{q}\right)\right] \\
&= \inf_{z \in [1,3/2]} \left[z\left(\log\left(\frac{2\zeta(2)}{\log^2(z)}\right) + h^{-1}(1+x)\right)\right] \\
&= 2 \inf_{z \in [1,3/2]} \left[z\left(\frac{\log(2\zeta(2)) + h^{-1}(1+x)}{2} - \log\log(z)\right)\right].
\end{aligned}$$

The expression in (3.7) follows by checking that the function $\tilde{h}$ defined in (3.8) satisfies

$$\tilde{h}(u) = \min_{z \in [1,3/2]} z(u - \log\log(z)).$$

Building the martingales: proof of Lemma 3.4 The interesting part of the proof is the construction of a martingale for each arm which together satisfy the conditions of Lemma 3.3.

A first natural candidate is the following, where we denote by $\phi_\mu(\eta) := \log \mathbb{E}_{X\sim\nu_\mu}\left[e^{\eta X}\right]$ the log moment generating function of the distribution that has mean $\mu \in \mathcal{I}$ and by $S_a(t) = \sum_{s=1}^t X_{a,s}\mathbb{1}(A_s = a)$ the sum of observations obtained from arm a in the first t rounds. For all $\eta \in \mathbb{R}$,

$$Z_a^\eta(t) = \exp\left(\eta S_a(t) - \phi_{\mu_a}(\eta)N_a(t)\right) \tag{3.10}$$

is a test martingale with respect to the filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$, for any sampling rule. Indeed, observing that $Z_a^\eta(t) = Z_a^\eta(t-1)\exp((\eta X_{a,t} - \phi_{\mu_a}(\eta))\mathbb{1}(A_t = a))$ with $X_{a,t}$ the observation made from arm a is it is selected at time t, we get $\mathbb{E}[Z_a^\eta(t)|\mathcal{F}_{t-1}] = Z_a^\eta(t-1)$ using the fact that A_t is $\mathcal{F}_{t-1}$-measurable and the definition of the moment generating function.

More generally, for any probability distribution π, the *mixture martingale*

$$\tilde{Z}_a^\pi(t) = \int Z_a^\eta(t)d\pi(\eta) \tag{3.11}$$

is also a test martingale, as can be seen by applying Tonelli's theorem

$$\mathbb{E}\left[\tilde{Z}_a^\pi(t)\middle|\mathcal{F}_{t-1}\right] = \int \underbrace{\mathbb{E}\left[\tilde{Z}_a^\eta(t)|\mathcal{F}_{t-1}\right]}_{=\tilde{Z}_a^\eta(t-1)} d\pi(\eta) = \tilde{Z}_a^\pi(t-1).$$

Given a family of priors $\boldsymbol{\pi} = (\pi_a)_{a=1}^K$, the *product martingale* $\tilde{Z}_{\mathcal{S}}^{\boldsymbol{\pi}}(t) = \prod_{a\in\mathcal{S}}\tilde{Z}_a^{\pi_a}(t)$ is also a test martingale with respect to $\mathcal{F}_t$, for any subset $\mathcal{S}$. Indeed, when $A_t \in \mathcal{S}$ we have

$$\mathbb{E}\left[\tilde{Z}_{\mathcal{S}}^{\boldsymbol{\pi}}(t)\middle|A_t, \mathcal{F}_{t-1}\right] = \tilde{Z}_{\mathcal{S}\setminus\{A_t\}}^{\boldsymbol{\pi}}(t-1)\underbrace{\mathbb{E}\left[\tilde{Z}_{A_t}^{\pi_{A_t}}(t)\middle|A_t, \mathcal{F}_{t-1}\right]}_{=\tilde{Z}_{A_t}^{\pi_{A_t}}(t-1)} = \tilde{Z}_{\mathcal{S}}^{\boldsymbol{\pi}}(t-1),$$

and the same result holds trivially when $A_t \notin \mathcal{S}$. The martingale property follows by the tower rule.

Hence, by associating one mixture martingale to each arm, assumption 2. in Lemma 3.3 is readily satisfied. Thus, to prove Lemma 3.4 it remains to construct for each arm a, each $\xi \in [0, 1/2]$ and each $\lambda \in [0, (1+\xi)^{-1})$ a prior $\pi = \pi_{a,\lambda,\xi}$ such that

$$\forall t \in \mathbb{N},\quad \tilde{Z}_a^{\pi_{a,\lambda,\xi}}(t) \geq e^{\lambda X_a(t) - g_\xi(\lambda)},$$

which proves that inequality $(*)$ is satisfied with $M_a^\lambda(t) = Z_a^{\pi_{a,\lambda,\xi}}(t)$.

In the literature, mixture martingales have often been defined for a prior π under which $\tilde{Z}_a^\pi(t)$ can be computed in closed-form, or tightly approximated. For example, in the Gaussian case, a common choice of prior is $\pi = \mathcal{N}(0, y^{-2})$ for which the corresponding mixture martingale is $\tilde{Z}_a^\pi(t) = \frac{y}{\sqrt{y^2+\sigma^2 N_a(t)}}\exp\left(\frac{(S_a(t)-\mu_a N_a(t))^2}{2(y^2+\sigma^2 N_a(t))}\right)$ (see e.g. De La Pena et al. (2004); Abbasi-Yadkori et al. (2011)). Applying the maximal inequality to this martingale directly yields the deviation inequality

$$\mathbb{P}\left(\exists t \in \mathbb{N}^* : \frac{N_a(t)(\hat{\mu}_a(t) - \mu_a)^2}{2\sigma^2} > \sqrt{1 + \frac{y^2}{N_a(t)\sigma^2}}\left(x + \frac{1}{2}\log\left(1 + \frac{\sigma^2 N_a(t)}{y^2}\right)\right)\right) \leq e^{-x}.$$

Compared to what we prove in Theorem 3.1 for $|\mathcal{S}| = 1$, observe that the $\log\log(N_a(t))$ featured in Theorem 3.1 is smaller than the $\log(N_a(t))$ in the right-hand side of the above inequality (at the cost of a less explicit threshold) . Moreover, this $\log\log N_a(t)$ is compatible with the Law of Iterated Logarithm. To obtain this tighter threshold (at least in a regime of large values of $N_a(t)$), and handle general exponential families, we need to resort to a more complex prior distribution. As will be seen shortly, the prior $\pi_{a,\lambda,\xi}$ will be a hierarchical prior, i.e. a continuous average of discrete priors.

The first step of our construction is the following lemma, which states that the process $X_a(t)$ crossing some threshold (slightly larger than) x implies that some mixture martingale with a prior $\pi(x)$ having discrete support exceeds a threshold. The proof of this result is postponed to the end of this section.

Lemma 3.5. *Fix $\xi \in (0, 1/2)$ and $x > 0$. There exists a (discrete) prior $\pi(x) = \pi(x, \xi)$ such that the corresponding mixture martingale satisfies, for all $t \in \mathbb{N}$,*

$$\left\{X_a(t) - (1+\xi)\log\left(C(\xi)\right) \geq x\right\} \subseteq \left\{\tilde{Z}_a^{\pi(x)}(t) \geq e^{\frac{x}{1+\xi}}\right\}.$$

We now exploit Lemma 3.5 to upper bound $e^{\lambda X_a(t) - (1+\xi)\log(C(\xi))}$ by a martingale. To do so, we first note that for every $z > 1$, and every $\lambda > 0$

$$\begin{aligned}\left\{e^{\lambda(X_a(t)-(1+\xi)\log C(\xi))} \geq z\right\} &\subseteq \left\{\tilde{Z}_a^{\pi(\log(z)/\lambda)}(t) \geq e^{\frac{\log(z)}{\lambda(1+\xi)}}\right\} \\ &\subseteq \Big\{\underbrace{\tilde{Z}_a^{\pi(\log(z)/\lambda)}(t)e^{-\frac{\log(z)}{\lambda(1+\xi)}}}_{:=W_a^{z,\lambda}(t)} \geq 1\Big\},\end{aligned}$$

where $W_a^{z,\lambda}(t)$ is a martingale that satisfies $\mathbb{E}[W_a^{z,\lambda}(0)] = e^{-\frac{\log(z)}{\lambda(1+\xi)}}$ and, due to the above inclusion,

$$W_a^{z,\lambda}(t) \geq \mathbb{1}_{\left(e^{\lambda(X_a(t)-(1+\xi)\log C(\xi))} \geq z\right)}. \tag{3.12}$$

We now define another mixture martingale, for $\lambda \in \left]0, \frac{1}{1+\xi}\right[$, this time using a continuous prior:

$$W_a^{\lambda}(t) = 1 + \int_1^{\infty} W_a^{z,\lambda}(t)dz.$$

Using inequality (3.12) yields

$$W_a^{\lambda}(t) \geq e^{\lambda(X_a(t)-(1+\xi)\log C(\xi))}.$$

Moreover, a direct computation shows that $W_a^{\lambda}(0) = \frac{1}{1-\lambda(1+\xi)}$. Finally defining

$$M_a^{\lambda}(t) = (1 - \lambda(1+\xi))W_a^{\lambda}(t),$$

one has that $M_a^{\lambda}(t)$ is a test martingale, i.e. $\mathbb{E}[M_a^{\lambda}(t)] = 1$, that satisfies

$$\begin{aligned}M_a^{\lambda}(t) &\geq \exp\left(\lambda X_a(t) - \lambda(1+\xi)\log(C(\xi)) + \log(1-\lambda(1+\xi))\right) \\ &= \exp\left(\lambda X_a(t) - g_{\xi}(\lambda)\right),\end{aligned}$$

which proves Lemma 3.4. Taking a step back, $M_a^{\lambda}(t)$ is indeed a mixture martingale as it can be written $M_a^{\lambda}(t) = \tilde{Z}_a^{\pi_{a,\lambda,\xi}}(t)$ with

$$\tilde{Z}_a^{\pi_{a,\lambda,\xi}}(t) = (1-\lambda(1+\xi))Z_a^0(t) + \int_1^{\infty}\left(\int Z_a^{\eta}(t)d(\pi(\log(z)/\lambda))(\eta)\right)(1-\lambda(1+\xi))e^{-\frac{\log(z)}{\lambda(1+\xi)}}dz.$$

Proof of Lemma 3.5 Recall that $X_a(t) = N_a(t)d\left(\hat{\mu}_a(t), \mu_a\right) - 3\log(1+\log(N_a(t))$. The first step of the construction consists in relating the deviation of $N_a(t)d^+(\hat{\mu}_a(t), \mu_a)$ and $N_a(t)d^-(\hat{\mu}_a(t), \mu_a)$ to those of $\eta S_a(t) - \phi_{\mu_a}(\eta)N_a(t)$ for a well chosen η, provided that $N_a(t)$ belongs to some "slice" $[(1+\xi)^{i-1}, (1+\xi)^i]$.

More precisely, we prove that for all $i \in \mathbb{N}^*$, there exists $\eta_i^+(x,\xi)$ and $\eta_i^-(x,\xi)$ such that, if $N_a(t) \in [(1+\xi)^{i-1}, (1+\xi)^i]$ it holds that

$$\{N_a(t)d^+(\hat{\mu}_a(t), \mu_a) \geq x\} \subseteq \left\{\eta_i^+ S_a(t) - N_a(t)\phi_{\mu_a}(\eta_i^+) \geq \frac{x}{1+\xi}\right\} \tag{3.13}$$

$$\{N_a(t)d^-(\hat{\mu}_a(t), \mu_a) \geq x\} \subseteq \left\{\eta_i^- S_a(t) - N_a(t)\phi_{\mu_a}(\eta_i^-) \geq \frac{x}{1+\xi}\right\}. \tag{3.14}$$

To prove these two inclusions, we introduce the notation θ for the natural parameter associated to μ_a, defined as $\theta = \dot{b}^{-1}(\mu_a)$ and define $\eta_i^+ < 0$ and $\eta_i^- > 0$ by

$$\mathrm{KL}(\theta + \eta_i^+, \theta) = \mathrm{KL}(\theta + \eta_i^-, \theta) = \frac{x}{(1+\xi)^i},$$

where $\mathrm{KL}(\theta, \theta')$ denotes the Kullback-Leibler divergence between the distributions of natural parameter θ and θ'. Using properties of the KL-divergence, one can write

$$\begin{aligned}
\mathrm{KL}(\theta + \eta_i^+, \theta) &= \eta_i^+ \mu_i^+ - \phi_{\mu_a}(\eta_i^+) \quad \text{with} \quad \mu_i^+ := \dot{b}^{-1}(\theta + \eta_i^+) < \mu_a, \\
\mathrm{KL}(\theta + \eta_i^-, \theta) &= \eta_i^- \mu_i^- - \phi_{\mu_a}(\eta_i^-) \quad \text{with} \quad \mu_i^- := \dot{b}^{-1}(\theta + \eta_i^-) > \mu_a.
\end{aligned}$$

For $N_a(t) \in [(1+\xi)^{i-1}, (1+\xi)^i]$, one can prove (3.13) by writing

$$\begin{aligned}
\{N_a(t) d^+(\hat{\mu}_a(t), \mu_a) \geq x\} &\subseteq \left\{ d^+(\hat{\mu}_a(t), \mu_a) \geq \frac{x}{(1+\xi)^i} \right\} \\
&\subseteq \{\hat{\mu}_a(t) \leq \mu_i^+\} \\
&\subseteq \{\eta_i^+ \hat{\mu}_a(t) - \phi_{\mu_a}(\eta_i^+) \geq \mathrm{KL}(\theta + \eta_i^+, \theta)\} \\
&\subseteq \left\{ (1+\xi)^{i-1} \left(\eta_i^+ \hat{\mu}_a(t) - \phi_{\mu_a}(\eta_i^+)\right) \geq \frac{x}{1+\xi} \right\} \\
&\subseteq \left\{ N_a(t) \left(\eta_i^+ \hat{\mu}_a(t) - \phi_{\mu_a}(\eta_i^+)\right) \geq \frac{x}{1+\xi} \right\},
\end{aligned}$$

where the third inclusion uses that $\eta_i^+ < 0$. Similarly, using this time that $\eta_i^- > 0$ one can prove (3.14):

$$\begin{aligned}
\{N_a(t) d^-(\hat{\mu}_a(t), \mu_a) \geq x\} &\subseteq \{\hat{\mu}_a(t) \geq \mu_i^-\} \\
&\subseteq \{\eta_i^- \hat{\mu}_a(t) - \phi_{\mu_a}(\eta_i^-) \geq \mathrm{KL}(\theta + \eta_i^-, \theta)\} \\
&\subseteq \left\{ N_a(t) \left(\eta_i^- \hat{\mu}_a(t) - \phi_{\mu_a}(\eta_i^-)\right) \geq \frac{x}{1+\xi} \right\}.
\end{aligned}$$

The next step is to relate the deviation of $X_a(t)$ to those of a martingale *for every* $t \in \mathbb{N}$ and not only for $N_a(t)$ is some slice: this will be achieved by a mixture martingale with a well-chosen discrete prior. Given x, we define the following probability distribution. Letting

$$\begin{aligned}
\gamma_i &= \tfrac{1}{2} \tfrac{1}{i^2 \zeta(2)} \qquad x_i = x + \log\left(\tfrac{1}{\gamma_i}\right) \\
\eta_i^+ &= \eta_i^+(x_i, \xi) \qquad \eta_i^- = \eta_i^-(x_i, \xi),
\end{aligned}$$

where $\eta_i^\pm(x, \xi)$ are defined above, we define the discrete prior

$$\pi = \sum_{i=1}^{\infty} \gamma_i \delta_{\eta_i^+} + \sum_{i=1}^{\infty} \gamma_i \delta_{\eta_i^-}$$

and the corresponding mixture martingale

$$\tilde{Z}_a^\pi(t) = \sum_{i=1}^{\infty} \gamma_i Z_a^{\eta_i^+}(t) + \sum_{i=1}^{\infty} \gamma_i Z_a^{\eta_i^-}(t),$$

where we recall that $Z_a^\eta(t) = \exp(\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t))$ for all $\eta \in \mathbb{R}$.

Therefore, we get

$$\begin{aligned}
&\{X_a(t) - (1+\xi)\log C(\xi) \geq x\} \\
&\quad \subseteq \left\{ \left[N_a(t) d(\hat{\mu}_a(t), \mu_a) - 3\log(1 + \log(N_a(t))) \right]^+ \geq x + (1+\xi)\log C(\xi) \right\} \\
&\quad = \left\{ N_a(t) d(\hat{\mu}_a(t), \mu_a) - 3\log(1 + \log(N_a(t))) \geq x + (1+\xi)\log C(\xi) \right\},
\end{aligned}$$

where we use that $x + (1+\xi)\log C(\xi) > 0$ as $\xi < 1/2$. Now, as $2(1+\xi) < 3$, one has

$$\begin{aligned}
&\{X_a(t) - (1+\xi)\log C(\xi) \geq x\} \\
&\quad \subseteq \left\{ N_a(t) d\left(\hat{\mu}_a(t), \mu_a\right) - 2(1+\xi)\log\left(1 + \log(N_a(t))\right) \geq x + (1+\xi)\log\left(\frac{2\zeta(2)}{\log(1+\xi)^2}\right) \right\} \\
&\quad \subseteq \left\{ N_a(t) d\left(\hat{\mu}_a(t), \mu_a\right) \geq x + (1+\xi)\log\left(\frac{2\zeta(2)(1+\log(N_a(t))^2}{\log(1+\xi)^2}\right) \right\} \\
&\quad \subseteq \left\{ N_a(t) d\left(\hat{\mu}_a(t), \mu_a\right) \geq x + (1+\xi)\log\left(\frac{2\zeta(2)(\log(1+\xi)+\log(N_a(t))^2}{\log(1+\xi)^2}\right) \right\},
\end{aligned}$$

where the last inequality uses $\log(1+\xi) \leq \log(3/2) \leq 1$.

We now define $i(t) \geq 1$ to be the integer such that $N_a(t) \in [(1+\xi)^{i-1}, (1+\xi)^i]$ and observe that $\frac{\log N_a(t)}{\log(1+\xi)} \geq i(t) - 1$. Using (3.13) and (3.14) yields

$$\begin{aligned}
&\{X_a(t) - (1+\xi)\log C(\xi) \geq x\} \\
&\quad \subseteq \left\{ N_a(t) d\left(\hat{\mu}_a(t), \mu_a\right) \geq x + (1+\xi)\log\left(\frac{1}{\gamma_{i(t)}}\right) \right\} \\
&\quad \subseteq \left\{ \max_{\eta \in \left\{\eta^+_{i(t)}, \eta^-_{i(t)}\right\}} \left[\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t)\right] \geq \frac{1}{1+\xi}\left[x + (1+\xi)\log\left(\frac{1}{\gamma_{i(t)}}\right)\right] \right\} \\
&\quad \subseteq \left\{ \max_{\eta \in \left\{\eta^+_{i(t)}, \eta^-_{i(t)}\right\}} \gamma_{i(t)} \exp\left(\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t)\right) \geq e^{\frac{x}{1+\xi}} \right\} \\
&\quad \subseteq \left\{ \max_{i \in \mathbb{N}} \max_{\eta \in \left\{\eta^+_i, \eta^-_i\right\}} \gamma_i \exp\left(\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t)\right) \geq e^{\frac{x}{1+\xi}} \right\} \\
&\quad \subseteq \left\{ \tilde{Z}^{\pi}_a(t) \geq e^{\frac{x}{1+\xi}} \right\}.
\end{aligned}$$

□

3.4 From Sequential to Active Testing

We recall that the goal of active identification in a bandit model is to find a sampling rule, a stopping rule and a recommendation rule such that the resulting strategy is δ-correct and also has a *small sample complexity* τ.

From the previous section, we already have a good candidate for the stopping and recommendation rule $(\tau_\delta, \hat{\imath}_{\tau_\delta})$: the Parallel GLRT. However, to minimize the sample complexity, the sampling rule now plays a crucial rule: intuitively it should be designed so that the stopping rule (3.4) is met as early as possible. Optimizing the sampling rule alongside $(\tau_\delta, \hat{\imath}_{\tau_\delta})$, we move from sequential testing to *active testing*. Active testing was pioneered by Chernoff (1959) who considered a setting in which there are two mutually exclusive hypotheses and A experiments that can be chosen to gather information about these hypotheses. Interestingly, the stopping rule proposed in this work for the case of two overlapping hypotheses coincides with the Parallel GLRT rule.

In the next chapters, we focus on the construction of good sampling rules for various active identification problems. As we shall see, *lower bounds* on the sample complexity can guide the design of asymptotically optimal algorithms, just like for regret minimization in structured bandits, that was discussed in Chapter 1.

Chapter 4

Towards Optimal and Efficient Best Arm Identification

In this chapter, we discuss the *sample complexity* of a particular active identification problem: Best Arm Identification. The main result presented in this chapter was obtained in collaboration with Aurélien Garivier in 2016. It provides the first *asymptotically optimal* algorithm for fixed-confidence best arm identification in an exponential bandit model.

4.1 Lower Bounds for Best Arm Identification and Beyond

In this chapter, we consider an exponential bandit model, in which arms belong to a one-parameter exponential family with divergence function $d(x,y)$: the distribution of arm a is ν_{μ_a} with $\mu_a \in \mathcal{I}$. We further assume that there is a unique optimal arm, that is $\boldsymbol{\mu} \in \mathcal{R}$ where

$$\mathcal{R} = \left\{ \boldsymbol{\mu} \in \mathcal{I}^A : \exists a \in [A] : \mu_a > \max_{b \neq a} \mu_b \right\} ,$$

and recall the best arm identification problem in the fixed-confidence setting. Letting $a_\star(\boldsymbol{\mu})$ be the arm with the largest mean in the bandit model parameterized by $\boldsymbol{\mu} \in \mathcal{R}$, the goal is to design a δ-*correct* strategy $(A_t, \tau, \hat{a}_\tau)$, such that for all $\boldsymbol{\mu} \in \mathcal{R}$, $\mathbb{P}(\hat{a}_\tau \neq a_\star(\boldsymbol{\mu})) \leq \delta$. This strategy should further have the *smallest possible sample complexity* $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$. To clarify the presentation, we will materialize the dependency on δ of the sample complexity of a δ-correct strategy, denoted by τ_δ.

Building on the change-of-distribution tools presented in the Introduction and already used in several places in Part I of this document, we provide a tight lower bound on the sample complexity of any δ-correct strategy. The first lower bound of this kind was given by Mannor and Tsitsiklis (2004) for ε-best arm identification in a Bernoulli bandit model: it features the gaps $\mu_\star - \mu_a$ of sub-optimal arms a and the parameter ε, but writes as a sum over a (non-explicit) subset of the arms. For $\varepsilon = 0$, the bound we proved with Aurélien Garivier and Olivier Cappé in Kaufmann et al. (2016) applies to any exponential bandit model and features a sum over all arms. It says that under any δ-correct strategy,

$$\mathbb{E}_\nu[\tau_\delta] \geq \left[\frac{1}{d(\mu_1, \mu_2)} + \sum_{a=2}^{A} \frac{1}{d(\mu_a, \mu_1)} \right] \mathrm{kl}(\delta, 1-\delta),$$

where we assume that $\mu_1 > \mu_2 \geq \cdots \geq \mu_A$ to ease the presentation. This lower bound features an individual complexity quantity for each arm which is the inverse of a Kullback-Leibler divergence and it therefore looks like a (non asymptotic) counterpart of the Lai and Robbins lower bound for regret minimization. However, we found that this simple and natural lower bound is *not* tight.

Just like the Lai and Robbins lower bound for regret minimization, the lower bounds of Mannor and Tsitsiklis (2004) and Kaufmann et al. (2016) were derived by considering an *explicit* change-of-distribution, that is, by picking some explicit alternative bandit models $\boldsymbol{\lambda}$ in which the optimal arm differs from that in $\boldsymbol{\mu}$, in order to obtain constraints on the number of selections of some arms. In the lower bound presented below, that we derived in Garivier and Kaufmann (2016), we instead directly aim for the tightest possible lower bound, without being explicit on the corresponding alternative model $\boldsymbol{\lambda}$ in

$$\mathrm{Alt}(\boldsymbol{\mu}) := \left\{ \boldsymbol{\mu} \in \mathcal{I}^A : a_\star(\boldsymbol{\lambda}) \neq a_\star(\boldsymbol{\mu}) \right\} .$$

Theorem 4.1. *Let $\delta \in (0,1)$. For any δ-PAC strategy and any bandit model $\boldsymbol{\mu} \in \mathcal{R}$,*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T^\star(\boldsymbol{\mu}) \, \mathrm{kl}(\delta, 1-\delta),$$

where

$$T^\star(\boldsymbol{\mu})^{-1} := \sup_{\boldsymbol{w} \in \Sigma_A} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^A w_a d(\mu_a, \lambda_a) \right) . \tag{4.1}$$

with $\Sigma_A = \{\boldsymbol{w} \in [0,1]^A : \sum_{a=1}^A w_a = 1\}$. Note that $\mathrm{kl}(\delta, 1-\delta) \geq \log\left(\frac{1}{3\delta}\right)$.

Proof. Let $\delta \in (0,1)$, $\boldsymbol{\mu} \in \mathcal{S}$, and consider a δ-PAC strategy. We assume that τ_δ is almost surely finite, otherwise $\mathbb{E}[\tau_\delta] = +\infty$ and the lower bound is trivial. From Lemma 0.1 in the Introduction, for any event $\mathcal{E} \in \mathcal{F}_{\tau_\delta}$ and for all $\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})$,

$$\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{I_{\tau_\delta}}, \mathbb{P}_{\boldsymbol{\lambda}}^{I_{\tau_\delta}} \right) \geq \mathrm{kl}\left(\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}), \mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}) \right),$$

where we recall that $I_t = (U_0, A_1, X_1, U_1, \ldots, A_t, X_t, U_t)$ is the information available after t rounds.

On the one hand, for exponential family bandits, it follows from Wald's inequality that

$$\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^{I_{\tau_\delta}}, \mathbb{P}_{\boldsymbol{\lambda}}^{I_{\tau_\delta}} \right) = \sum_{a=1}^A \mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau_\delta)] d(\mu_a, \lambda_a).$$

On the other hand, choosing the event $\mathcal{E} = (\hat{a}_{\tau_\delta} = a_\star(\boldsymbol{\lambda}))$ it holds that $\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}) \leq \delta$ while $\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}) \geq 1-\delta$ and exploiting monotonicity properties of the binary relative entropy yields

$$\mathrm{kl}\left(\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}), \mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}) \right) \geq \mathrm{kl}(\delta, 1-\delta).$$

This leads to

$$\forall \boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu}), \quad \sum_{a=1}^A \mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau_\delta)] d(\mu_a, \lambda_a) \geq \mathrm{kl}(\delta, 1-\delta). \tag{4.2}$$

Instead of choosing for each arm a a specific instance of $\boldsymbol{\lambda}$ that yields a lower bound on $\mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau_\delta)]$, we combine here the inequalities given by all alternatives $\boldsymbol{\lambda}$:

$$\begin{aligned}
\inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^A \mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau_\delta)] d(\mu_a, \lambda_a) &\geq \mathrm{kl}(\delta, 1-\delta) \\
\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \times \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^A \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]} d(\mu_a, \lambda_a) &\geq \mathrm{kl}(\delta, 1-\delta) \\
\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \times \left(\sup_{\boldsymbol{w} \in \Sigma_A} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^A w_a d(\mu_a, \lambda_a) \right) &\geq \mathrm{kl}(\delta, 1-\delta).
\end{aligned}$$

In the last inequality, the strategy-dependent proportions of arm draws are replaced by their supremum (using that $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] = \sum_{a=1}^A \mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau_\delta)]$) so as to obtain a bound valid for any δ-PAC algorithm.

□

It can be observed from this proof that for a strategy to match the lower bound, the last inequality needs to be an equality. Introducing the (set of) vector(s) of optimal proportions

$$\boldsymbol{w}^\star(\boldsymbol{\mu}) := \operatorname*{argmax}_{w \in \Sigma_A} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{A} w_a d(\mu_a, \lambda_a), \tag{4.3}$$

this means that the vector $\left(\frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}\right)_{a=1}^{A}$ should belong to $\boldsymbol{w}^\star(\boldsymbol{\mu})$. Note that, at this stage we have no clue whether the argmax in (4.3) is unique, but we will prove that it is the case in the next section. We will further propose an algorithm to efficiently compute $\boldsymbol{w}^\star(\boldsymbol{\mu})$ as well as the non-explicit characteristic time $T^\star(\boldsymbol{\mu})$ in (4.1).

Lower Bounds for Active Identification The proof technique to establish Theorem 4.1 can be used to derive a (non-explicit) lower bound for any active identification problem with *non-overlapping* regions $\mathcal{R}_1, \dots, \mathcal{R}_I$. By appropriately redefining the set $\mathrm{Alt}(\boldsymbol{\mu}) = \mathcal{R} \backslash \mathcal{R}_{i_\star(\boldsymbol{\mu})}$ if $i_\star(\boldsymbol{\mu})$ is the (only) region to which $\boldsymbol{\mu}$ belongs, we obtain exactly the same lower bound as in Theorem 4.1. Note that for arbitrary active identification problems, computing $T^\star(\boldsymbol{\mu})$ and $w^\star(\boldsymbol{\mu})$ may be harder than what we will present for BAI shortly. We will see another example in Chapter 5.

Lower Bounds for Non-Overlapping regions Deriving lower bounds (and matching algorithms) for active identification problems where the regions $\mathcal{R}_1, \dots, \mathcal{R}_I$ may be overlapping can be much harder, as explained by Degenne and Koolen (2019). In Garivier and Kaufmann (2019) we study the particular case of ε-best arm identification. Interestingly, when $\boldsymbol{\mu}$ belongs to the intersection of two or more regions, one can only provide *asymptotic* lower bounds on $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$ (when δ goes to zero), whereas Theorem 4.1 is non asymptotic as it holds for any $\delta \in (0, 1)$.

4.2 The Track-and-Stop Algorithm

For the best arm identification problem, we now explain how the characteristic time $T^\star(\boldsymbol{\mu})$ and the optimal allocation $\boldsymbol{w}^\star(\boldsymbol{\mu})$ can be efficiently computed for any $\boldsymbol{\mu} \in \mathcal{R}$. Having access to this oracle permits to design a lower-bound inspired algorithm called Track-And-Stop. We prove that the sample complexity of this algorithm is matching the lower bound of Theorem 4.1, at least for a small δ.

4.2.1 Computing the Optimal Allocation

To ease the presentation we present the results for $\boldsymbol{\mu}$ such that $\mu_1 > \mu_2 \geq \cdots \geq \mu_A$. To compute

$$\begin{aligned} T^\star(\boldsymbol{\mu})^{-1} &= \sup_{\boldsymbol{w} \in \Sigma_A} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{A} w_a d(\mu_a, \lambda_a) \right) \\ \text{and} \quad \boldsymbol{w}^\star(\boldsymbol{\mu}) &= \operatorname*{argmax}_{\boldsymbol{w} \in \Sigma_A} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{A} w_a d(\mu_a, \lambda_a) \right) \end{aligned}$$

we start by making the minimization over $\mathrm{Alt}(\boldsymbol{\mu})$ slightly more explicit.

Lemma 4.2. *For every $\boldsymbol{w} \in \Sigma_A$,*

$$\inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{A} w_a d(\mu_a, \lambda_a) \right) = \min_{a \neq 1} \left[w_1 d\left(\mu_1, \frac{w_1 \mu_1 + w_a \mu_a}{w_1 + w_a}\right) + w_a d\left(\mu_a, \frac{w_1 \mu_1 + w_a \mu_a}{w_1 + w_a}\right) \right].$$

Proof. Using the fact that $\mathrm{Alt}(\boldsymbol{\mu}) = \bigcup_{a\neq 1}\{\boldsymbol{\lambda}\in\mathcal{S} : \lambda_a > \lambda_1\}$, one has

$$\begin{aligned} T^\star(\boldsymbol{\mu})^{-1} &= \sup_{\boldsymbol{w}\in\Sigma_A}\min_{a\neq 1}\inf_{\boldsymbol{\lambda}\in\mathcal{S}:\lambda_a>\lambda_1}\sum_{a'=1}^{A} w_{a'}d(\mu_{a'},\lambda_{a'}) \\ &= \sup_{\boldsymbol{w}\in\Sigma_A}\min_{a\neq 1}\inf_{\boldsymbol{\lambda}:\lambda_a\geq\lambda_1}\left[w_1 d(\mu_1,\lambda_1) + w_a d(\mu_a,\lambda_a)\right] . \end{aligned}$$

Minimizing

$$f(\lambda_1,\lambda_a) = w_1 d(\mu_1,\lambda_1) + w_a d(\mu_a,\lambda_a)$$

under the constraint $\lambda_a \geq \lambda_1$ is a convex optimization problem that can be solved analytically. The minimum is obtained for

$$\lambda_1 = \lambda_a = \frac{w_1}{w_1+w_a}\mu_1 + \frac{w_a}{w_1+w_a}\mu_a,$$

which concludes the proof.

□

Using Lemma 4.2 and the fact that at the optimum $w_1^\star \neq 0$ (otherwise the value of the objective is zero), one can write

$$\begin{aligned} T^\star(\boldsymbol{\mu})^{-1} &= \sup_{\boldsymbol{w}\in\Sigma_A}\min_{a\neq 1}\left[w_1 d\left(\mu_1,\frac{w_1\mu_1+w_a\mu_a}{w_1+w_a}\right) + w_a d\left(\mu_a,\frac{w_1\mu_1+w_a\mu_a}{w_1+w_a}\right)\right] \\ &= \sup_{\boldsymbol{w}\in\Sigma_A} w_1 \min_{a\neq 1}\left[d\left(\mu_1,\frac{\mu_1+\frac{w_a}{w_1}\mu_a}{1+\frac{w_a}{w_1}}\right) + \frac{w_a}{w_1}d\left(\mu_a,\frac{\mu_1+\frac{w_a}{w_1}\mu_a}{1+\frac{w_a}{w_1}}\right)\right] \\ &= \sup_{\boldsymbol{w}\in\Sigma_A} w_1 \min_{a\neq 1} g_a\left(\frac{w_a}{w_1}\right) , \end{aligned}$$

where we introduce for all $a\in\{2,\dots,A\}$ the function

$$g_a(x) = d\left(\mu_1,\frac{\mu_1+x\mu_a}{1+x}\right) + x d\left(\mu_a,\frac{\mu_1+x\mu_a}{1+x}\right) .$$

The function g_a is a strictly increasing one-to-one mapping from $[0,+\infty[$ onto $[0,d(\mu_1,\mu_a)[$. We define $x_a : [0,d(\mu_1,\mu_a)[\to[0,+\infty[$ as its inverse function: $x_a(y) = g_a^{-1}(y)$. Denoting by x_1 the function constantly equal to 1, one obtains the following characterization of $\boldsymbol{w}^\star(\boldsymbol{\mu})$.

Theorem 4.3. *For every $a\in[A]$,*

$$w_a^\star(\boldsymbol{\mu}) = \frac{x_a(y^\star)}{\sum_{a=1}^{A} x_a(y^\star)} , \tag{4.4}$$

where $y^\star$ is the unique solution of the equation $F_{\boldsymbol{\mu}}(y) = 1$, and where

$$F_{\boldsymbol{\mu}} : y \mapsto \sum_{a=2}^{A} \frac{d\left(\mu_1,\frac{\mu_1+x_a(y)\mu_a}{1+x_a(y)}\right)}{d\left(\mu_a,\frac{\mu_1+x_a(y)\mu_a}{1+x_a(y)}\right)} \tag{4.5}$$

is a continuous, increasing function on $[0,d(\mu_1,\mu_2)[$ such that $F_{\boldsymbol{\mu}}(0) = 0$ and $F_{\boldsymbol{\mu}}(y)\to\infty$ when $y\to d(\mu_1,\mu_2)$).

Proof. The function g_a rewrites

$$g_a(x) = d\left(\mu_1, m_a(x)\right) + xd(\mu_a, m_a(x)), \quad \text{with} \quad m_a(x) = \frac{\mu_1 + x\mu_a}{1+x} .$$

Using that $m_a'(x) = (\mu_a - \mu_1)/(1+x)^2$ and $\frac{d}{dy}d(x,y) = (y-x)/\dot{b}(\dot{b}^{-1}(y))$ where b denotes the log-partition function of the exponential family, one can show that g_a is strictly increasing, since $g_a'(x) = d(\mu_a, m_a(x)) > 0$. As $g_a(x)$ tends to $d(\mu_1, \mu_a)$ when x goes to infinity, the inverse function $x_a(y) = g_a^{-1}(y)$ is defined on $[0, d(\mu_1, \mu_a)[$ and satisfies

$$x_a'(y) = \frac{1}{d(\mu_a, m_a(x_a(y)))} > 0 .$$

Let $\boldsymbol{w}^\star$ be an element in

$$\operatorname*{argmax}_{\boldsymbol{w}\in\Sigma_A} w_1 \min_{a\neq 1} g_a\left(\frac{w_a}{w_1}\right) .$$

Introducing $x_a^\star = \frac{w_a^\star}{w_1^\star}$ for all $a \neq 1$, one has

$$w_1^\star = \frac{1}{1+\sum_{a=2}^A x_a^\star} \quad \text{and, for } a \geq 2, \quad w_a^\star = \frac{x_a^\star}{1+\sum_{a=2}^A x_a^\star}$$

and $(x_2^\star, \dots, x_A^\star) \in \mathbb{R}^{A-1}$ belongs to

$$\operatorname*{argmax}_{(x_2,\dots,x_A)\in\mathbb{R}^{A-1}} \frac{\min_{a\neq 1} g_a\left(x_a\right)}{1+x_2+\cdots+x_A}. \tag{4.6}$$

We now prove that all the $g_a(x_a^\star)$ have to be equal. Let

$$\mathcal{B} = \left\{ b \in \{2,\dots,A\} : g_b(x_b^\star) = \min_{a\neq 1} g_a(x_a^\star) \right\}$$

and $\mathcal{A} = \{2,\dots,A\}\backslash\mathcal{B}$. Assume that $\mathcal{A} \neq \varnothing$. For all $a \in \mathcal{A}$ and $b \in \mathcal{B}$, one has $g_a(x_a^\star) > g_b(x_b^\star)$. Using the continuity of the g functions and the fact that they are strictly increasing, there exists $\varepsilon > 0$ such that

$$\forall a \in \mathcal{A}, b \in \mathcal{B}, \quad g_a(x_a^\star - \varepsilon/|\mathcal{A}|)) > g_b\left(x_b^\star + \varepsilon/|\mathcal{B}|\right) > g_b\left(x_b^\star\right) .$$

Introducing $\overline{x}_a = x_a^\star - \varepsilon/|\mathcal{A}|$ for all $a \in \mathcal{A}$ and $\overline{x}_b = x_b^\star + \varepsilon/|\mathcal{B}|$ for all $b \in \mathcal{B}$, there exists $b \in \mathcal{B}$:

$$\frac{\min_{a\neq 1} g_a\left(\overline{x}_a\right)}{1+\overline{x}_2+\dots\overline{x}_A} = \frac{g_b\left(x_b^\star + \varepsilon/|\mathcal{B}|\right)}{1+x_2^\star+\cdots+x_A^\star} > \frac{g_b\left(x_b^\star\right)}{1+x_2^\star+\cdots+x_A^\star} = \frac{\min_{a\neq 1} g_a(x_a^\star)}{1+x_2^\star+\cdots+x_A^\star} ,$$

which contradicts the fact that $\boldsymbol{x}^\star$ belongs to (4.6). Hence $\mathcal{A} = \varnothing$ and there exists $y^\star \in [0, d(\mu_1,\mu_2)[$ such that

$$\forall a \in \{2,\dots,A\}, \; g_a(x_a^\star) = y^\star \quad \Leftrightarrow \quad x_a^\star = x_a(y^\star),$$

with the function x_a introduced above. From (4.6), $y^\star$ belongs to

$$\operatorname*{argmax}_{y\in[0,d(\mu_1,\mu_2)[} G(y) \quad \text{with} \quad G(y) = \frac{y}{1+x_2(y)+\cdots+x_A(y)} .$$

G is differentiable and, using the derivative of the x_a given above, $G'(y) = 0$ is equivalent to

$$\begin{aligned} \sum_{a=2}^{A} \frac{y}{d(\mu_a, m_a(x_a(y)))} &= 1 + x_2(y) + \cdots + x_A(y) \\ \sum_{a=2}^{A} \frac{d(\mu_1, m_a(x_a(y))) + x_a(y) d(\mu_a, m_a(x_a(y)))}{d(\mu_a, m_a(x_a(y)))} &= 1 + x_2(y) + \cdots + x_A(y) \\ \sum_{a=2}^{A} \frac{d(\mu_1, m_a(x_a(y)))}{d(\mu_a, m_a(x_a(y)))} &= 1 \,. \end{aligned} \tag{4.7}$$

For the the second equality, we use that $\forall a, d(\mu_1, m_a(x_a(y))) + x_a(y) d(\mu_a, m_a(x_a(y))) = y$. Thus $y^\star$ is a solution to the equation (4.7). This equation has a unique solution since

$$F_{\boldsymbol{\mu}}(y) = \sum_{a=2}^{A} \frac{d(\mu_1, m_a(x_a(y)))}{d(\mu_a, m_a(x_a(y)))} \tag{4.8}$$

is strictly increasing and satisfies $F_{\boldsymbol{\mu}}(0) = 0$ and $\lim_{y \to d(\mu_1,\mu_2)} F_{\boldsymbol{\mu}}(y) = +\infty$. As G is positive and satisfies $G(0) = 0$, $\lim_{y \to d(\mu_1,\mu_2)} G(y) = 0$, the unique local extremum obtained in $y^\star$ is a maximum.

□

Thus, $\boldsymbol{w}^\star(\boldsymbol{\mu})$ can be easily computed by applying (for example) the bisection method to a function whose evaluations require the resolution of A smooth scalar equations. By using efficient numerical solvers, we obtain a fast algorithm of complexity, roughly speaking, proportional to the number of arms. This characterization of $\boldsymbol{w}^\star(\boldsymbol{\mu})$ also permits to establish the following properties.

Proposition 4.4. *1. For all $\boldsymbol{\mu} \in \mathcal{S}$, for all a, $w_a^\star(\boldsymbol{\mu}) \neq 0$.*

2. $\boldsymbol{w}^\star$ is continuous in every $\boldsymbol{\mu} \in \mathcal{R}$.

3. If $\mu_1 > \mu_2 \geq \cdots \geq \mu_A$, one has $w_2^\star(\boldsymbol{\mu}) \geq \cdots \geq w_A^\star(\boldsymbol{\mu})$.

In general, it is not possible to give closed-form formulas for $T^\star(\boldsymbol{\mu})$ and $\boldsymbol{w}^\star(\boldsymbol{\mu})$. In particular, $T^\star(\boldsymbol{\mu})$ cannot be written as a sum over the arms of individual complexity terms as in previous works (Mannor and Tsitsiklis, 2004; Kaufmann et al., 2016). Still, in the Gaussian case, a tight approximation of this form can be given for $T^\star(\boldsymbol{\mu})$: we show in Appendix A of Garivier and Kaufmann (2016) that

$$\frac{2\sigma^2}{(\mu_1 - \mu_2)^2} + \sum_{a=2}^{A} \frac{2\sigma^2}{(\mu_1 - \mu_a)^2} \leq T^\star(\boldsymbol{\mu}) \leq 2 \left[\frac{2\sigma^2}{(\mu_1 - \mu_2)^2} + \sum_{a=2}^{A} \frac{2\sigma^2}{(\mu_1 - \mu_a)^2} \right] .$$

For two-armed bandits, it is also possible to obtain a more explicit expression: $T^\star(\boldsymbol{\mu}) = 1/d_\star(\mu_1, \mu_2)$ where $d_\star(\mu_1, \mu_2) = d(\mu_1, \bar{\mu})$ with $\bar{\mu}$ is defined by $d(\mu_1, \bar{\mu}) = d(\mu_2, \bar{\mu})$, which is some 'reversed' notion of Chernoff information.

4.2.2 Track-and-Stop and Its Analysis

Track-and-Stop combines the Parallel GLRT stopping rule presented in Chapter 4 with a sampling rule under which the empirical proportions of draws of each arm a, $N_a(t)/t$, converges to the corresponding optimal proportion $w_a^\star(\boldsymbol{\mu})$. To achieve this, the sampling rule is "tracking" the empirical optimal proportion $\boldsymbol{w}^\star(\hat{\boldsymbol{\mu}}(t))$ where $\hat{\boldsymbol{\mu}}(t) = (\hat{\mu}_1(t), \ldots, \hat{\mu}_A(t))$: it makes sure that $N_a(t)/t$ is always close to $w_a^\star(\hat{\boldsymbol{\mu}}(t))$. In order to ensure that $w_a^\star(\hat{\boldsymbol{\mu}}(t))$ is also eventually close to $w_a^\star(\boldsymbol{\mu})$, we need to add a bit of forced exploration so that all arms are sufficiently selected and $\hat{\boldsymbol{\mu}}(t)$ converges to $\boldsymbol{\mu}$.

Sampling rule The "tracking" idea can be implemented in different ways by monitoring the ratio or difference between $N_a(t)$ and $t \times w_a^\star(\hat{\boldsymbol{\mu}}(t))$ or the cumulative proportion $\sum_{s=1}^t w_a^\star(\hat{\boldsymbol{\mu}}(s))$. We propose two different tracking rules in Garivier and Kaufmann (2016), and mention only the so-called Direct Tracking rule below. Note that similar tracking approaches have been used in previous work, see e.g. Antos et al. (2008). The (direct) Tracking rule goes as follows. Introducing $U_t = \{a : N_a(t) < \sqrt{t} - A/2\}$, the sampling rule (A_t) is sequentially defined as

$$A_{t+1} \in \begin{cases} \underset{a \in U_t}{\operatorname{argmin}}\, N_a(t) & \text{if } U_t \neq \varnothing, \quad (\textit{forced exploration}) \\ \underset{a \in [A]}{\operatorname{argmax}}\, t\, w_a^\star(\hat{\boldsymbol{\mu}}(t)) - N_a(t) & \text{otherwise.} \quad (\textit{direct tracking}) \end{cases} \tag{4.9}$$

The following property shows that this sampling rule has the desired property: if $w_a^\star(\hat{\boldsymbol{\mu}}(t))$ is close to $w_a^\star(\boldsymbol{\mu})$ for t large enough, $N_a(t)/t$ will eventually be close to $w_a^\star(\boldsymbol{\mu})$.

Lemma 4.5. *Under the tracking rule, for all $a \in [A]$, $N_a(t) \geq (\sqrt{t} - A/2)_+ - 1$ and for all $\varepsilon > 0$, for all t_0, there exists $t_\varepsilon \geq t_0$ such that*

$$\sup_{t \geq t_0} \max_a |w_a^\star(\hat{\boldsymbol{\mu}}(t)) - w_a^\star(\boldsymbol{\mu})| \leq \varepsilon \quad \Rightarrow \quad \sup_{t \geq t_\varepsilon} \max_a \left| \frac{N_a(t)}{t} - w_a^\star(\boldsymbol{\mu}) \right| \leq 3(A-1)\varepsilon \,.$$

In particular, from Lemma 4.5, the law of large numbers and the continuity of $\boldsymbol{w}^\star$ on $\mathcal{R}$, one has

$$\mathbb{P}_{\boldsymbol{\mu}} \left(\lim_{t \to \infty} \frac{N_a(t)}{t} = w_a^\star(\boldsymbol{\mu}) \right) = 1 \,.$$

Stopping rule and recommendation rule We now particularize the Parallel GLRT presented in Chapter 4 for the best arm identification problem. From (3.4), the stopping rule can be written

$$\begin{aligned} \tau_\delta &= \inf \left\{ t \in \mathbb{N} : \max_{a=1,\dots,A} \inf_{\{\boldsymbol{\lambda} \in \mathcal{R} : a_*(\boldsymbol{\lambda}) \neq a\}} \sum_{b=1}^A N_b(t) d\left(\hat{\mu}_b(t), \lambda_b\right) > \beta(t,\delta) \right\} \\ &= \inf \left\{ t \in \mathbb{N} : \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \sum_{a=1}^A N_a(t) d\left(\hat{\mu}_a(t), \lambda_a\right) > \beta(t,\delta) \right\}. \end{aligned} \tag{4.10}$$

The equality comes from the fact that the maximizer in a in the first expression is obtained for $\hat{a}_\star(t) = \operatorname{argmax}_a \hat{\mu}_a(t)$, indeed for all other values of a the infimum is zero as $\hat{\boldsymbol{\mu}}(t)$ belongs to the set. Upon stopping, the arm $\hat{a}_{\tau_\delta} = \operatorname{argmax}_a \hat{\mu}_a(\tau_\delta)$ is recommended.

The expression in (4.10) explains why the Parallel GLRT rule is a natural candidate for matching the lower bound when coupled with a sampling rule that ensures the convergence of each $N_a(t)/t$ to its corresponding optimal proportion. Indeed for large values of t, under such a sampling rule

$$\inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \sum_{a=1}^A N_a(t) d\left(\hat{\mu}_a(t), \lambda_a\right) \quad \simeq \quad t \times \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^A w_a^\star(\boldsymbol{\mu}) d\left(\mu_a, \lambda_a\right) = \frac{t}{T^\star(\boldsymbol{\mu})}$$

With this approximation, τ_δ cannot be much larger than the first t such that $t \geq T^\star(\boldsymbol{\mu})\beta(t,\delta)$, which is of order $T^\star(\boldsymbol{\mu}) \log(1/\delta)$ for the threshold under which we proved that the Parallel GLRT is δ-correct in Chapter 4. Note that this heuristic reasoning could be generalized to any active identification problem for which the weights $\boldsymbol{w}^\star(\boldsymbol{\mu})$ are uniquely defined and continuous on $\mathcal{R}$. However, the corresponding combination of the tracking rule with the Parallel GLRT may be much harder to implement, as it requires an efficient computation of the weights $\boldsymbol{w}^\star(\boldsymbol{\mu})$ for any $\boldsymbol{\mu}$ as well as an efficient computation of the minimizer in (4.10), which may not always exist beyond best arm identification.

For BAI, the minimization in (4.10) can be solved in closed form (with similar arguments as in the proof of Lemma 4.2), leading to

$$\tau_\delta = \inf\left\{ t \in \mathbb{N} : \min_{a \neq \hat{a}_\star(t)} \left[N_{a_\star}(t) d\left(\hat{\mu}_{\hat{a}_\star(t)}(t), \hat{\mu}_{\hat{a}_\star(t),a}(t)\right) + N_a(t) d\left(\hat{\mu}_a(t), \hat{\mu}_{\hat{a}_\star(t),a}(t)\right)\right] > \beta(t,\delta) \right\}, \quad (4.11)$$

where $\hat{\mu}_{a,b}(t) = \frac{N_a(t)\hat{\mu}_a(t) + N_b(t)\hat{\mu}_b(t)}{N_a(t)+N_b(t)}$ is the weighted average of the empirical means of arms a and b.

Theoretical guarantees The Track-and-Stop (TaS) algorithm combines the Tracking rule (4.9) with the Parallel GLRT rule (4.11) using some threshold $\beta(n,\delta)$. We already explained that this algorithm can be efficiently implemented as each computation of $\boldsymbol{w}^\star(\hat{\boldsymbol{\mu}}(t))$ requires to solve a one-dimensional optimization problem using binary search. We formally state below the sample complexity guarantees obtained for this algorithm, showing that TaS is matching the lower bound of Theorem 4.1, at least in a regime of small values of δ. This is why we say that this algorithm is asymptotically optimal.

Theorem 4.6. *Let $\boldsymbol{\mu}$ be an exponential family bandit model. The Track-and-Stop algorithm in which the Parallel GLRT stopping rule uses the threshold $\beta(t,\delta) = 2\mathcal{T}\left(\log(1/\delta)/2\right) + 6\log(1+\log(t))$ where $\mathcal{T}(x) \sim x$ is defined in (3.7) in Chapter 3 is δ-correct and satisfies*

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

The correctness easily follows from Theorem 3.1 in Chapter 3 while the sample complexity analysis can be found in Appendix D of Garivier and Kaufmann (2016). The convergence of the empirical proportions to the optimal weights permits to easily establish that $\limsup_{\delta\to 0} \tau_\delta / \log(1/\delta) \leq T^\star(\boldsymbol{\mu})$ almost surely (like in the proof of Lemma 4.8 in the next section), but the proof of the expected sample complexity bound of Theorem 4.6 requires to control the convergence speed, and further exploits Lemma 4.5.

Practical performance Theorem 4.6 provides an *asymptotic* upper bound on the sample complexity, hence Track-and-Stop is expected to match the optimal performance when δ is small enough. However, performing numerical simulations reveals that TaS actually has a very good sample complexity for any value of δ: its sample complexity is typically twice smaller than that of previous algorithms.

Among existing algorithms for BAI that come with problem-dependent sample complexity guarantees in the fixed-confidence setting, there are two types of algorithms. First, algorithms using uniform sampling and elimination, like Successive Elimination (Even-Dar et al., 2006) or KL-Racing (Kaufmann and Kalyanakrishnan, 2013), under which an arm is eliminated from the list of candidate best arms when an upper confidence bound on its mean is smaller than a lower confidence bound on the mean of the empirical best arm. Such algorithms stop when there is a single arm remaining in the list. The second category is that of algorithms using more adaptive sampling, that are based on upper and possibly lower confidence bounds (calibrated as a function of the risk level δ). Lil'UCB (Jamieson et al., 2014) chooses the arm with largest UCB and stops when one arm has been drawn more than all others. Other algorithms such that (KL)-LUCB (Kalyanakrishnan et al., 2012; Kaufmann and Kalyanakrishnan, 2013) or UGapE Gabillon et al. (2012) also exploit lower confidence bounds and stop when the lower confidence bound of the empirical best arm is larger than the upper confidence bound of all others.

We believe that the main reason for the good practical performance of TaS is the *stopping rule*: compared to confidence-based stopping rules, the Parallel GLRT (also referred to as the Chernoff stopping rule in the paper) exploits the geometry of the distributions better, which leads to earlier stopping. In Figure 4.1 we display the (KL) confidence intervals on the unknown means (represented as the black diamonds) of Bernoulli arms after TaS reaches its stopping rule: we see that the Parallel GLRT has enough information to identify the best arm before the confidence intervals are separated.

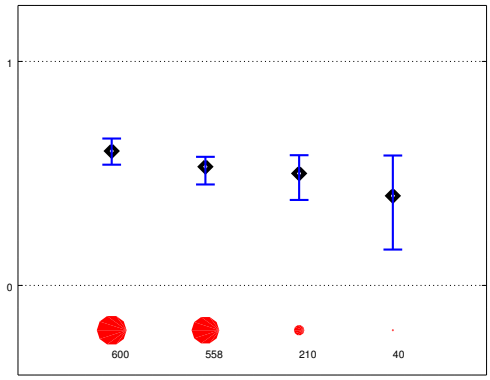


Figure 4.1 – KL confidence intervals for the means (black diamonds) in a 4-armed bandit problem when TaS reaches its stopping rule. The number of selections of each arm is reported on the x axis.

To justify the good performance of TaS, we compare its empirical sample complexity to that of other algorithms for two Bernoulli bandit problems whose means are given by $\boldsymbol{\mu}_1 = [0.5\ 0.45\ 0.43\ 0.4]$ and $\boldsymbol{\mu}_2 = [0.3\ 0.21\ 0.2\ 0.19\ 0.18]$. For these two vectors, the optimal allocations are given by

$$\begin{aligned} \boldsymbol{w}^\star(\boldsymbol{\mu}_1) &= [0.417\ 0.390\ 0.136\ 0.057]\,, \\ \boldsymbol{w}^\star(\boldsymbol{\mu}_2) &= [0.336\ 0.251\ 0.177\ 0.132\ 0.104]\,. \end{aligned}$$

	Track-and-Stop	Chernoff-Racing	KL-LUCB	KL-Racing
$\boldsymbol{\mu}_1$	**4052**	4516	8437	9590
$\boldsymbol{\mu}_2$	**1406**	3078	2716	3334

Table 4.1 – Expected number of draws $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$ for $\delta = 0.1$, estimated over $N = 3000$ runs.

Track-and-Stop was run with the threshold $\beta(t,\delta) = \log\left(\frac{\log(t)+1}{\delta}\right)$, which is an approximation of the threshold in Theorem 4.6 that we recommend to use in practice, as we found that it was sufficient to have an empirical error (much) smaller than δ on all the experiments we performed (not reported here). A similar tweak of the threshold used inside the confidence intervals of confidence-based was suggested before (Kaufmann and Kalyanakrishnan, 2013), and is used in our experiments. We see in Table 4.1 that TaS has indeed a sample complexity twice smaller than that of KL-LUCB and KL-Racing. We also compared to an improved Racing algorithm, called Chernoff-Racing, that uses a GLRT test to assess that one arm is better than an other instead of comparing confidence intervals to perform eliminations. We see that Chernoff-Racing always outperforms KL-Racing, and that on some instances it can be competitive with TaS. However, this approach is not sufficient for asymptotic optimality.

4.3 Beyond Track-and-Stop

We showed that TaS has nice theoretical properties, as it is the first asymptotically optimal algorithm for BAI in exponential family bandit models, and that it has good practical performance, i.e. its empirical sample complexity is much smaller than that of other algorithms. However, we did not spend too much time on its computational cost. While confidence-based algorithm are quite easy to implement, TaS requires a call to the oracle $\boldsymbol{w}^\star$ in each round. This call boils down to solving the equation $F_{\boldsymbol{\mu}}(y) = 1$

with $F_{\boldsymbol{\mu}}$ defined in (4.8) using binary search. But each evaluation of $F_{\boldsymbol{\mu}}$ is costly as it also requires to compute the inverse function x_a for each arm. Hence, the running time of TaS is an order of magnitude larger than that of confidence-based algorithms, which makes its use prohibitive when the number of arms gets large, for example.

Finding alternative approaches to Track-and-Stop that are asymptotically optimal without the need to compute the optimal proportion in every round has been an active line of research over the past years, and I present below two interesting directions, to which I contributed little, but would be very interested to contribute more in the future.

4.3.1 Online Optimization and Optimism

The first work replacing the oracle call in Track and Stop with a gradient ascent step is that of Pierre Ménard (Ménard, 2019). To understand the proposed approach, we first recall that the computation of the optimal weights consists in computing the argmax in $\boldsymbol{w}$ in the optimization problem

$$\sup_{\boldsymbol{w}\in\Sigma_A} F(\boldsymbol{w},\boldsymbol{\mu}), \quad \text{where} \quad F(\boldsymbol{w},\boldsymbol{\mu}) = \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{A} w_a d(\mu_a,\lambda_a).$$

With the appropriate definition of $\mathrm{Alt}(\boldsymbol{\mu})$ this actually permits to compute the optimal weights for any active identification problem. For best arm identification, we proposed an ad hoc algorithm for computing $\boldsymbol{w}^\star$, however we can take a step back and investigate the use of generic optimization techniques.

On the use of online optimization algorithms The simplex Σ_A is convex and $F(\boldsymbol{w},\boldsymbol{\mu})$ is concave in $\boldsymbol{w}$, however it is not necessarily smooth. Still in the Best Arm Identification case, recalling that

$$F(\boldsymbol{w},\boldsymbol{\mu}) = \min_{a\neq a_\star(\boldsymbol{\mu})} \left[w_{a_\star(\boldsymbol{\mu})} d\left(\mu_{a_\star(\boldsymbol{\mu})}, \mu_{a_\star(\boldsymbol{\mu}),a}(\boldsymbol{w})\right) + w_a d\left(\mu_a, \mu_{a_\star(\boldsymbol{\mu}),a}(\boldsymbol{w})\right)\right] \tag{4.12}$$

with the weighted average $\mu_{a,b}(\boldsymbol{w}) = \frac{w_a\mu_a + w_b\mu_b}{w_a+w_b}$, we notice that F has sub-gradients. Indeed, some computations show that for every arm a that is a minimizer in (4.12), the vector

$$\partial_{\boldsymbol{w}} F(\boldsymbol{w},\boldsymbol{\mu}) = d(\mu_{a_\star(\boldsymbol{\mu})}, \mu_{a_\star(\boldsymbol{\mu}),a}(\boldsymbol{w}))e_{a_\star(\boldsymbol{\mu})} + d(\mu_a, \mu_{a_\star(\boldsymbol{\mu}),a}(\boldsymbol{w}))e_a$$

where e_i is the i-th vector of the canonical basic of $\mathbb{R}^A$ is a sub-gradient of F. Hence, one could use Projected Subgradient Descent or Mirror Descent (see, e.g. Bubeck (2015)) to compute $\boldsymbol{w}^\star(\boldsymbol{\mu})$. These online optimization algorithms would however be much slower than the algorithm suggested by Theorem 4.3 to produce accurate oracle calls.

The idea of Ménard (2019) is not to use a gradient-based algorithm to perform the oracle calls in Track-and-Stop but instead to replace the oracle call by one step of a (sub)-gradient based algorithm in each round of the algorithm. The optimization algorithm chosen is lazy mirror descent. Denoting by $\pi_0 = \left(\frac{1}{A},\dots,\frac{1}{A}\right)$ the uniform distribution on $\{1,\dots,A\}$, the proposed algorithm takes the form

$$\begin{aligned}
\widetilde{w}(t+1) &= \operatorname*{argmax}_{\boldsymbol{w}\in\Sigma_A} \left[\eta_{t+1} \boldsymbol{w}^\top \left(\sum_{s=A}^{t} \mathrm{Clip}_s\left(\partial_{\boldsymbol{w}} F\left(\widetilde{w}(s), \hat{\mu}(s)\right)\right)\right) - \mathrm{KL}\left(\boldsymbol{w},\pi_0\right)\right] \\
w'(t+1) &= (1-\gamma_t)\widetilde{w}(t+1) + \gamma_t \pi_0 \\
A_{t+1} &\in \operatorname*{argmax}_{a\in[A]} \sum_{s=1}^{t+1} w'_a(s) - N_a(t),
\end{aligned}$$

where the clipping $\mathrm{Clip}_s(x) = (x_a \wedge M\sqrt{s})_{a\in[A]}$ for an arbitrary constant M is a technical trick to handle possibly un-bounded sub-gradients. Note also that the mixing with the uniform distribution in the

second line can be seen as a counterpart of the forced exploration used in Track-and-Stop. Ménard (2019) proves that this sampling rule used in conjunction with the Parallel GLRT is δ-correct and asymptotically optimal for the choices $\eta_t = 1/\sqrt{t}$ and $\gamma_t = 1/(4\sqrt{t})$. This algorithm avoids the computational cost of the oracle calls, but still uses some kind of forced exploration.

How about Frank-Wolfe? Even if the Frank-Wolfe algorithm is in principle suited for optimizing smooth functions and F is not smooth, one can write the Frank-Wolfe update on the objective $\boldsymbol{w} \mapsto F(\boldsymbol{w}, \boldsymbol{\mu})$:

$$\boldsymbol{w}_{t+1} = (1-\eta_t)\boldsymbol{w}_t + \eta_t \boldsymbol{s}_t \quad \text{where} \quad \boldsymbol{s}_t \in \operatorname*{argmax}_{\boldsymbol{w} \in \Sigma_A} \boldsymbol{w}^\top \partial_{\boldsymbol{w}_t} F(\boldsymbol{w}_t, \boldsymbol{\mu}) \ .$$

With the above expression of the sub-gradient in the BAI case, $\boldsymbol{s}_t \in \{e_{a_\star}, e_{c_t}\}$ where

$$c_t \in \operatorname*{argmin}_{a \neq a_\star} \left[w_{a_\star} d(\mu_{a_\star}, \mu_{a_\star,a}(\boldsymbol{w}_t)) + w_a d(\mu_a, \mu_{a_\star,a}(\boldsymbol{w}_t)) \right]$$

and $\boldsymbol{s}_t = e_{a_\star}$ if and only if $d(\mu_{a_\star}, \mu_{a_\star,c_t}(\boldsymbol{w}_t)) > d(\mu_{c_t}, \mu_{a_\star,c_t}(\boldsymbol{w}_t))$.

This update inspires the following sampling rule: letting $\boldsymbol{w}(t) = \left(\frac{N_a(t)}{t}\right)_{a \in [A]}$ be the vector of empirical fraction of selection of each arm, on can compute

$$\begin{aligned} B_t &= \operatorname*{argmax}_{a \in [A]} \hat{\mu}_a(t) && \textit{(candidate best arm)} \\ C_t &= \operatorname*{argmin}_{a \neq B_t} \left[w_{B_t}(t) d(\hat{\mu}_{B_t}(t), \hat{\mu}_{B_t,a}(\boldsymbol{w}(t))) + w_a(t) d(\hat{\mu}_a(t), \hat{\mu}_{B_t,a}(\boldsymbol{w}(t))) \right] && \textit{(challenger)} \end{aligned}$$

and select

$$A_{t+1} = \begin{cases} B_t & \text{if } d\left(\hat{\mu}_{B_t}, \hat{\mu}_{B_t,C_t}(\boldsymbol{w}(t))\right) > d\left(\hat{\mu}_{C_t}, \hat{\mu}_{B_t,C_t}(\boldsymbol{w}(t))\right) \ , \\ C_t & \text{else} \ . \end{cases}$$

This sampling rule, proposed by Ménard (2019), is reminiscent of the Best Challenger sampling rule that we proposed in Garivier and Kaufmann (2016). However, it is not clear that it ensures convergence to the optimal weights (even with forced exploration), or that the noise-free version w_t converges to $\boldsymbol{w}^\star(\boldsymbol{\mu})$. Indeed, numerical experiments that we performed with Pierre Ménard revealed that in bandit models such that $\mu_1 > \mu_2 = \mu_3$, Frank-Wolfe and the Best Challenger rule fail to converge.

Yet, this does not mean that a Frank-Wolfe algorithm cannot work for other active identification problems: as explained by Ménard (2019), it leads to an optimal algorithm for the thresholding bandit problem (Locatelli et al., 2016), even if the corresponding F is not smooth. Meanwhile, Berthet and Perchet (2017) studied different bandit optimization problems in which there is also an underlying function $F(\boldsymbol{w}, \boldsymbol{\mu})$ to optimize in $\boldsymbol{w}$ based on sampling the arms of the bandit model $\boldsymbol{\mu}$: when F is smooth, they advocate the use of Frank-Wolfe combined with upper confidence bounds on the gradients.

A two-player game interpretation By proposing a new interpretation of the lower bound as the solution of a two-players game, Degenne et al. (2019) open new directions for algorithms. The quantity

$$\sup_{\boldsymbol{w} \in \Sigma_A} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{A} w_a d(\mu_a, \lambda_a)$$

is viewed as the value of a "pure exploration game" between a MAX player who selects actions in $[A]$ (using a mixed strategy) and a MIN player, who plays $\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})$. The payoff of the MAX player selecting a when the MIN player selects $\boldsymbol{\lambda}$ is $d(\mu_a, \lambda_a)$ (or equivalently, her loss is $-d(\mu_a, \lambda_a)$).

The idea of Degenne et al. (2019) consists in playing two no-regret learning algorithms (aimed at minimizing the loss) against each other: an algorithm $\mathcal{A}^{\mathrm{MAX}}$ for the MAX player (or rather one algorithm

$\mathcal{A}_a^{\text{MAX}}$ for each possible best arm a) and an algorithm $\mathcal{A}^{\text{MIN}}$ for the MIN player. Both algorithms may play distributions over $[A]$ or over the alternative of the current best arm. An additional feature of the algorithm is that it uses some optimism, in the sense that the learner used by the MAX player is fed with upper-confidence bounds on the payoffs. This allows to get rid of the forced exploration used in Track-and-Stop and in the algorithm of Ménard (2019) described above.

Several instances of this general principle (corresponding to different learning algorithms for the two players) are studied by Degenne et al. (2019) and are proved to be asymptotically optimal for Bernoulli and Gaussian bandits. For the sake of concreteness, we present only one for best arm identification, in which the MIN player uses Best Response and the learning algorithm $\mathcal{A}_a^{\text{MAX}}$ is AdaHedge (de Rooij et al., 2014), an Exponentiated Gradient algorithm for minimizing a sum of loss functions $\sum_{s=1}^t \ell_s(\boldsymbol{w})$ which can cope with unbounded gradients. In each round t, the empirical means $\hat{\boldsymbol{\mu}}(t)$ and the empirical best arm B_t are computed. Then the MAX player calls $\mathcal{A}_{B_t}^{\text{MAX}}$ which outputs a weight vector $\boldsymbol{w}(t)$. The best response for the MIN player is

$$\boldsymbol{\lambda}(t) = \operatorname*{argmin}_{\boldsymbol{\lambda} \in \text{Alt}(\hat{\boldsymbol{\mu}}(t))} \sum_{a=1}^{A} w_a(t) d(\hat{\mu}_a(t), \lambda_a).$$

To update the learning algorithm, an upper confidence bound on the payoff of the MAX player is computed: $U_a(t) = \max_{\xi \in [\text{LCB}_a(t), \text{UCB}_a(t)]} d(\xi, \lambda_a(t))$ for some underlying confidence interval and $\mathcal{A}_{B_t}^{\text{MAX}}$ is fed with the loss function $\ell(\boldsymbol{w}) = -\sum_{a=1}^A w_a U_a(t)$. To deduce an arm to play from the sequence of weights $\boldsymbol{w}(s)$ used by the algorithm, a Tracking procedure is needed and $A_{t+1} = \operatorname*{argmax}_{a \in [A]} \frac{N_a(t)}{\sum_{s=1}^t w_a(s)}$.

Hence, we presented an asymptotically optimal algorithm (at least for Bernoulli and Gaussian bandits) which avoids the computational complexity of the oracle calls and gets rid of forced exploration with optimism. It would be interesting to investigate whether a Thompson scheme could be used instead, which avoids the (conservative) design of a confidence interval. We now discuss other possible variants of Thompson Sampling for BAI.

4.3.2 Bayesian Approaches to the Rescue

Another idea to propose a simple and asymptotically optimal algorithm is to come up with a variant of Thompson Sampling, which is a Bayesian, anytime algorithm, for best arm identification. Why is a variant needed? Thompson Sampling is asymptotically optimal for regret minimization, hence we know that for all $a \in [A]$, it satisfies $N_a(t)/t \to \delta_{(a=a_\star)}$. Therefore, there is no hope for Thompson to converge to the optimal proportions, that are supported on all arms. This fundamental difference between regret minimization and pure exploration algorithms has been known since the work of Bubeck et al. (2011).

The question of whether Thompson Sampling could be adapted to best arm identification is actually the very reason I discovered the best arm identification literature a long time ago, as it was asked to me by Shivaram Kalyanakrishnan, with whom I collaborated during my PhD. But this collaboration made me discover many other interesting questions, especially regarding lower bounds, and I did not come back to answering the very first one. Fortunately, Daniel Russo answered it in 2016 by proposing Top-Two Thompson Sampling and several other variants (Russo, 2016).

Top-Two Thompson Sampling (TTTS) follows a simple idea: as vanilla Thompson Sampling 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. The pseudo-code of the TTTS sampling rule can be found in Algorithm 10, where Π_t denotes the posterior distribution after t observations, and Π_0 the prior distribution.

The TTTS sampling rule can be used in conjunction with the Parallel GLRT and can be studied in the fixed-confidence setting, as will be seen shortly. However, this is not the setting initially considered

Algorithm 10 Top-Two Thompson Sampling (TTTS)

Input: Parameter $\beta \in (0,1)$
for $t = 1, 2, \ldots$ **do**
  Sample $\boldsymbol{\theta} \sim \Pi_{t-1}$
  $I^{(1)} \leftarrow \underset{a \in [A]}{\text{argmax}} \theta_a$
  Draw an independent Bernoulli random variable $b \sim \mathcal{B}(\beta)$
  if $b = 1$ **then**
    $A_t = I^{(1)}$
  else
    Repeat sample $\boldsymbol{\theta}' \sim \Pi_{t-1}$
    $I^{(2)} \leftarrow \underset{a \in [A]}{\text{argmax}} \theta'_a$
    until $I^{(2)} \neq I^{(1)}$
    $A_t = I^{(2)}$
  end if
  Select arm A_t and observe $X_t \sim \nu^{\mu_{A_t}}$
  Compute the new posterior Π_t
end for

by Russo (2016). Instead, he studies some intrinsic (Bayesian) properties of the sampling rule in terms of posterior convergence. More precisely, under some assumption on the prior distribution Π_0 to be specified shortly, he proves that TTTS satisfies

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}_{\boldsymbol{\theta}\sim\Pi_n} \left(\theta_{a_\star(\boldsymbol{\mu})} < \max_a \theta_a \right) = \Gamma^\star_\beta(\boldsymbol{\mu}), \ \mathbb{P}_{\boldsymbol{\mu}} - a.s., \tag{4.13}$$

where the quantity $\Gamma^\star_\beta(\boldsymbol{\mu})$ can be expressed with our notation as

$$\Gamma^\star_\beta(\boldsymbol{\mu}) = \sup_{\substack{\boldsymbol{w}\in\Sigma_A: \\ w_{a_\star(\boldsymbol{\mu})}=\beta}} \inf_{\boldsymbol{\lambda}\in \text{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{A} w_a d(\mu_a, \lambda_a).$$

Interestingly, $\Gamma^\star_\beta(\boldsymbol{\mu})$ is very close to the inverse of the characteristic time $T^\star(\boldsymbol{\mu})$ defined in (4.1), the only difference being the extra condition that $w_{a^\star(\boldsymbol{\mu})} = \beta$ in the supremum. Intuitively, this condition comes from the fact that by design TTTS spends a fraction β selecting the same arm as Thompson Sampling would, and this arm is often $a_\star(\boldsymbol{\mu})$.

In words, Equation (4.13) says that the posterior probability of the set of bandit models that have an optimal arm different from $a_\star(\boldsymbol{\mu})$ decays in $e^{-n\Gamma^\star_\beta(\boldsymbol{\mu})}$. The assumption made by Russo on the prior distribution are the following: the prior Π_0 on the vector of natural parameter $\boldsymbol{\theta}$ should satisfy:

- Π_0 is supported on $\Theta = (\underline{\theta}, \overline{\theta})^A$, where $\underline{\theta}, \overline{\theta} \in \mathbb{R}$.
- Π_0 has a density π_0 with respect to the Lebesgue measure that satisfies

$$0 < \inf\{\pi_0(\boldsymbol{\lambda}), \boldsymbol{\lambda} \in \Theta\} < \sup\{\pi_0(\boldsymbol{\lambda}), \boldsymbol{\lambda} \in \Theta\} < \infty.$$

Under these assumption, it is further shown that $\Gamma^\star_\beta(\boldsymbol{\mu})$ is the best possible rate of decay for strategies that spend a fraction β of the time on the optimal arm, that is, any strategy such that $N_{a_\star(\boldsymbol{\mu})}(t)/t$ converges to β satisfies

$$\limsup_{n\to\infty} -\frac{1}{n} \log \mathbb{P}_{\boldsymbol{\theta}\sim\Pi_n} \left(\theta_{a_\star(\boldsymbol{\mu})} < \max_a \theta_a \right) \leq \Gamma^\star_\beta(\boldsymbol{\mu}).$$

Note that in order to have the fastest decay rate, one should further select $\beta^\star = \text{argmax}_\beta \Gamma^\star_\beta(\boldsymbol{\mu})$ which is such that $T^\star(\boldsymbol{\mu}) = 1/\Gamma^\star_{\beta^\star}(\boldsymbol{\mu})$. Yet this optimal tuning depends on the arms' means.

From Bayesian to fixed-confidence guarantees Under some conditions on the prior distribution, a sufficient condition to prove (4.13) is to establish the almost sure convergence of the vector $(N_a(t)/t)_{a=1}^A$ to the weight vector $\boldsymbol{w}^\beta(\boldsymbol{\mu})$, where

$$\boldsymbol{w}^\beta(\boldsymbol{\mu}) \in \underset{\substack{\boldsymbol{w}\in\Sigma_A:\\ w_{a_\star(\boldsymbol{\mu})}=\beta}}{\operatorname{argmax}} \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^A w_a d(\mu_a, \lambda_a).$$

Indeed Russo (2016) proves the following result, under the above condition on the prior Π_0.

Lemma 4.7. *If the prior Π_0 has a bounded support in natural parameter space with a positive density, any allocation rule satisfying $N_{a_\star(\boldsymbol{\mu})}(t)/t \to w_a^\beta(\boldsymbol{\mu})$ $\mathbb{P}_{\boldsymbol{\mu}}$-a.s. for all $a \in [A]$ satisfies*

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}_{\boldsymbol{\theta}\sim\Pi_n}\left(\theta_{a_\star(\boldsymbol{\mu})} < \max_a \theta_a\right) = \Gamma_\beta^\star(\boldsymbol{\mu}), \quad \mathbb{P}_{\boldsymbol{\mu}} - a.s..$$

The assumptions on Π_0 are a bit restrictive, as they rule out the typical conjugate priors used for Bernoulli and Gaussian bandits. In the paper Qin et al. (2017), the author later proved that this result also holds for Gaussian distribution with the usual improper prior, and in the paper Shang et al. (2020), we prove that it also holds for Bernoulli bandits with a Beta prior (see Theorem 6 therein). In this work, we also show that under TTTS with parameter β the empirical proportions of draws of each arm do converge to $\boldsymbol{w}_\beta^\star(\boldsymbol{\mu})$ for Gaussian and Bernoulli bandits with these practical prior distributions, hence TTTS is optimal in terms of the decay of posterior probability of the set of wrong models.

A natural question is whether this convergence property is sufficient to provide sample complexity guarantees for the δ-correct algorithm that combines the Parallel GLRT stopping and recommendation rule with the TTTS sampling rule [1]. From the convergence property, one can easily deduce the following bound on τ_δ, which holds almost surely.

Lemma 4.8. *Fix $\boldsymbol{\mu}$ a bandit model. Under an* anytime *sampling rule that satisfies, for all $a \in [A]$ $\lim_{t\to\infty} \frac{N_a(t)}{t} = w_a^\beta(\boldsymbol{\mu})$, almost surely, it holds that*

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\limsup_{\delta\to 0} \frac{\tau_\delta}{\log(1/\delta)} \le \frac{1}{\Gamma_\beta^\star(\boldsymbol{\mu})}\right) = 1.$$

Proof. We assume that $a_\star(\boldsymbol{\mu}) = 1$ to fix the ideas and let $\mathcal{E}$ be the event

$$\mathcal{E} = \left\{\forall a \in [A], \frac{N_a(t)}{t} \underset{t\to\infty}{\to} \left(w_\beta^\star(\boldsymbol{\mu})\right)_a \text{ and } \hat{\boldsymbol{\mu}}(t) \underset{t\to\infty}{\to} \boldsymbol{\mu}\right\}.$$

From the assumption on the sampling rule and the law of large numbers (as all arms are drawn a linear amount of time since the weights are non-zero), $\mathcal{E}$ is of probability 1.

On $\mathcal{E}$, there exists t_0 such that for all $t \ge t_0$, $\hat{\mu}_1(t) > \max_{a\neq 1} \hat{\mu}_a(t)$, hence $\mathrm{Alt}(\hat{\boldsymbol{\mu}}(t)) = \mathrm{Alt}(\boldsymbol{\mu})$ and the Parallel GLRT stops when $Z(t)$ exceed $\beta(t,\delta)$ where

$$Z(t) = \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^A N_a(t) d(\hat{\mu}_a(t), \lambda_a) = t \times \min_{a\neq 1} G_a\left(\hat{\boldsymbol{w}}(t), \hat{\boldsymbol{\mu}}(t)\right),$$

with $\hat{w}_a(t) = \frac{N_a(t)}{t}$ and

$$G_a(\boldsymbol{w}, \boldsymbol{\lambda}) = w_1 d\left(\lambda_1, \frac{w_1\lambda_1 + w_a\lambda_a}{w_1 + w_a}\right) + w_a d\left(\lambda_a, \frac{w_1\lambda_1 + w_a\lambda_a}{w_1 + w_a}\right).$$

1. Recall that the Parallel GLRT rule is δ-correct under any sampling rule, in particular with TTTS.

For all $a \geq 2$, the mapping $(\boldsymbol{w}, \boldsymbol{\lambda}) \to G_a(\boldsymbol{w}, \boldsymbol{\lambda})$ is continuous at $(\boldsymbol{w}^\beta(\boldsymbol{\mu}), \boldsymbol{\mu})$. Therefore, for all $\varepsilon > 0$ there exists $t_1 \geq t_0$ such that for all $t \geq t_1$ and all $a \in [A]$,

$$G_a\left(\hat{\boldsymbol{w}}(t), \hat{\boldsymbol{\mu}}(t)\right) \geq \frac{G_a\left(\boldsymbol{w}^\beta(\boldsymbol{\mu}), \boldsymbol{\mu}\right)}{1+\varepsilon}.$$

Hence, for $t \geq t_1$,

$$Z(t) \geq \frac{t}{1+\varepsilon} \min_{a \neq 1} G_a\left(\boldsymbol{w}^\beta(\boldsymbol{\mu}), \boldsymbol{\mu}\right) = \frac{t}{1+\varepsilon} \min_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{A} w_a^\beta(\boldsymbol{\mu}) d(\mu_a, \lambda_a) = \frac{t}{1+\varepsilon} \Gamma_\beta^\star(\boldsymbol{\mu}).$$

Consequently,

$$\begin{aligned} \tau_\delta &= \inf\{t \in \mathbb{N} : Z(t) \geq \beta(t, \delta)\} \\ &\leq t_1 \vee \inf\left\{t \in \mathbb{N} : t \geq \frac{1+\varepsilon}{\Gamma_\beta^\star(\boldsymbol{\mu})} \beta(t, \delta)\right\}. \end{aligned}$$

The threshold $\beta(t, \delta)$ satisfies $\beta(t, \delta) \leq \log\left(\frac{t}{\delta}\right) + o_{\delta \to 0}\left(\log\left(\frac{1}{\delta}\right)\right)$ and simple algebra (e.g. Lemma 18 in Garivier and Kaufmann (2016)) shows that

$$\tau_\delta \leq t_1 \vee \left[\frac{(1+\varepsilon)}{\Gamma_\beta^\star(\boldsymbol{\mu})} \log\left(\frac{1}{\delta}\right) + o_{\delta \to 0}\left(\log\left(\frac{1}{\delta}\right)\right)\right].$$

Thus τ_δ is finite on $\mathcal{E}$ for every $\delta \in (0, 1)$, and

$$\limsup_{\delta \to 0} \frac{\tau_\delta}{\log(1/\delta)} \leq \frac{(1+\varepsilon)}{\Gamma_\beta^\star(\boldsymbol{\mu})}.$$

Letting ε go to zero concludes the proof.

□

However, the almost sure convergence of the empirical proportions is not sufficient to prove a bound on the *expected* sample complexity. To prove such a result for Track-and-Stop, a more precise result on the convergence speed was needed (Lemma 4.5). In Shang et al. (2020), we establish an expected sample complexity bound for TTTS by following the same technique as the one introduced by Qin et al. (2017) to analyse the sample complexity of another Bayesian algorithm named TTEI (for Top-Two Expected Improvement). They show that a sufficient condition to establish

$$\limsup_{\delta \to \infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leq \frac{1}{\Gamma_\beta^\star(\boldsymbol{\mu})} \tag{4.14}$$

for any sampling rule used in conjunction with the Parallel GLRT is to show that this sampling rule satisfies, for all $\varepsilon \in (0, 1)$, $\mathbb{E}\left[T_\beta^\varepsilon\right] < \infty$ where T_β^ε is the random variable

$$T_\beta^\varepsilon := \inf\left\{N \in \mathbb{N} : \max_{a \in [A]} \left|\frac{N_a(n)}{n} - w_a^\beta\right| \leq \varepsilon; \forall n \geq N\right\}.$$

In Theorem 3 of Shang et al. (2020), we prove that TTTS satisfies $\mathbb{E}\left[T_\beta^\varepsilon\right] < \infty$ for Gaussian bandits. Establishing this property for TTTS is much more intricate than for TTEI due to the randomized nature of TTTS.

Hence, the BAI algorithm that uses TTTS as a sampling rule and the Parallel GLRT stopping and recommendation rule satisfies (4.14). The following lower bound (proved using the exact same technique as for Theorem 4.1) shows that this is optimal, among algorithms allocating a β fraction of the samples to the optimal arm. We call an algorithm matching this lower bound a *β-optimal* algorithm.

Theorem 4.9. *Let $\delta \in (0,1)$. For any δ-PAC strategy, for all $\boldsymbol{\mu}$ such that the sampling rule satisfies $\frac{N_{a_\star(\boldsymbol{\mu})}(t)}{t} \to \beta$ $\mathbb{P}_{\boldsymbol{\mu}}$ - a.s., it holds that*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq \frac{1}{\Gamma^\star_\beta(\boldsymbol{\mu})} \mathrm{kl}(\delta, 1-\delta).$$

Is TTTS a good alternative to Track-And-Stop? We established that a Bayesian BAI algorithm can have a *near-optimal* sample complexity without performing a costly computation of the optimal proportion in every-round. However, to be asymptotically optimal for BAI, TTTS would need to be run with the oracle parameter $\beta^\star$ where $\beta^\star = \boldsymbol{w}^\star_{a^\star(\boldsymbol{\mu})}(\boldsymbol{\mu})$, which is of course unknown.

In practice, Russo (2016) suggests the use of $\beta = 1/2$ which is proved to satisfy $\Gamma^\star_{1/2}(\boldsymbol{\mu}) \geq \Gamma^\star_{\beta^\star}(\boldsymbol{\mu})/2$ for all $\boldsymbol{\mu}$. But he also shows that if the algorithm uses in each step t a parameter β_t that satisfies $\beta_t \to \beta^\star$ almost surely, then the Bayesian optimality properties are preserved. However, this is less clear for the fixed-confidence guarantees, and an efficient adaptive tuning (that does not need a complex oracle call in every round) has still to be found.

Under TTTS, the cost of an oracle call to $\boldsymbol{w}^\star$ is replaced by the cost of sampling the posterior. While sampling from the typical conjugate posterior used in exponential families is easy, note that the *re-sampling* step (l. 9-11 in Algorithm 10) can be quite long. Indeed, when the posterior on all arms become concentrated, a lot of samples are needed before the optimal arm appears sub-optimal. In Shang et al. (2020), we propose an efficient approximation of this costly re-sampling step, which leads to the T3C algorithm (for Top Two Transportation Cost). Introducing the transportation cost

$$W_n(a,b) := \begin{cases} 0 & \text{if } \hat{\mu}_b(n) \geq \hat{\mu}_a(n), \\ N_a(n) d\left(\hat{\mu}_a(n), \hat{\mu}_{a,b}(n)\right) + N_b(n) d\left(\hat{\mu}_b(n), \hat{\mu}_{a,b}(n)\right) & \text{otherwise,} \end{cases}$$

it can be observed that under TTTS, if $I_n^{(1)}$ and $I_n^{(2)}$ are the two candidate arms that can be sampled in round $n+1$,

$$\mathbb{P}\left(I_n^{(2)} = b | I_n^{(1)} = a\right) = \frac{a_{n,b}}{\sum_{k \neq a} a_{n,k}} \tag{4.15}$$

where $a_{n,k}$ is the posterior probability that arm k is optimal: $a_{n,k} = \Pi_n\left(\theta_k > \max_{i \neq k} \theta_i\right)$. When a coincides with the empirical best mean (this will often be the case for $I_n^{(1)}$ when n is large due to posterior convergence) one can write

$$a_{n,b} \simeq \Pi_n\left(\theta_b \geq \theta_a\right) \simeq \exp\left(-W_n(a,b)\right),$$

where the last step is justified for Gaussian and Bernoulli distribution by upper bound that we provide in Shang et al. (2020). In T3C we propose to replace sampling from the distribution (4.15) by an approximation of its mode, which is easy to compute. That is, we define

$$I_n^{(2)} = \operatorname*{argmin}_{b \neq I_n^{(1)}} W_n\left(I_n^{(1)}, b\right)$$

where $I_n^{(1)} = \mathrm{argmax}_a \theta_a$ with $\boldsymbol{\theta} \sim \Pi_n$ and the T3C sampling rule selects

$$A_{n+1} = \begin{cases} I_n^{(1)} & \text{with probability } \beta\,, \\ I_n^{(2)} & \text{with probability } 1-\beta\,. \end{cases}$$

Experiments performed in Shang et al. (2020) reveal that T3C has comparable performance to TTTS but with smaller computational cost.

Hence, the next step is to combine T3C with an adaptive tuning of β in order to propose a computationally simple asymptotically optimal algorithm.

Chapter 5

Applications to Monte-Carlo Tree Search

In this chapter, we focus on two more sophisticated examples of Active Identification problems that are motivated by the search for a better theoretical understanding of Monte-Carlo Tree Search. The presented works are the outcome of a fruitful collaboration with Wouter Koolen and Aurélien Garivier.

5.1 A Simple Model for Planning in Games

The word "planning" has different meanings in the context of reinforcement learning. It often means computing the optimal policy in a Markov Decision Process (MDP) with known reward function and transition probabilities, i.e. find out what is the best next action to take in each possible state. But when the MDP is large, this task is already very hard as the usual Dynamic Programming solution is intractable. A simpler objective is therefore to find the best action to perform in the *current* state, based on possible calls to a generative model, which simulates transitions in the MDP often in the form of trajectories starting in the current state. This is the planning task we have in mind in this chapter: finding the best action to take in a given state using as few calls to the generative model as possible.

Such a planning problem has also been studied in the context of two player games, where the goal is to find the best possible move to take for a given player in a given state of the game, by exploring several trajectories (sequences of successive moves by the two players) starting from this state. Besides simple (solved) games in which the tree of all possible trajectories can be fully stored, Monte-Carlo Tree Search methods (see, e.g., Browne et al., 2012, for a survey) may be used to smartly explore a subset of this very large game tree in order to find a good move. The architecture of a typical MCTS algorithm in shown in Figure 5.1.

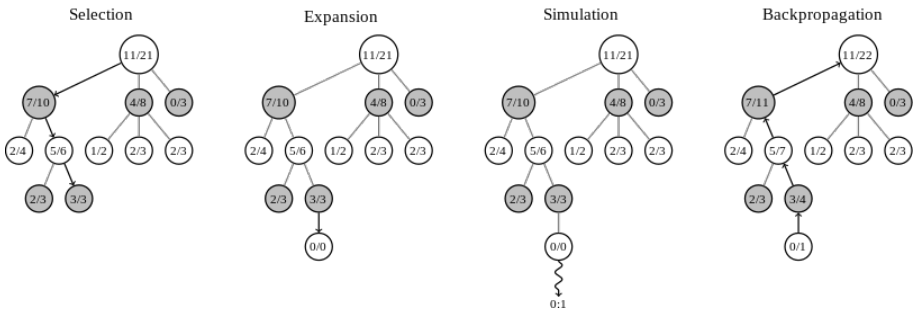


Figure 5.1 – The 4 steps of common MCTS algorithms. White nodes present the states in which the root player takes action (MAX nodes), grey states the states in which the adversary plays (MIN nodes)

A MCTS algorithm has two crucial components: a *selection strategy*, which (adaptively) chooses a path down the search tree, and a *leaf evaluation* method. The first MCTS algorithms often relied on a random evaluation (called simulation in Figure 5.1) in the form of a *playout* which consists in finishing the game from the current position using a simple heuristic (typically, playing at random) and reporting 1 if the game was won by the root player, 0 otherwise. Outcomes of theses playouts are then back-propagated to the state visited before reaching the leaf, and may be used by an adaptive selection strategy. In more recent MCTS algorithms that lead to a breakthrough in Go, namely AlphaGo (Silver et al., 2016) and AlphaZero (Silver et al., 2018), playouts are combined or even replaced by a leaf evaluation function that is learned by a neural network, mapping the board to a value. Still, we propose below an idealized model for the vanilla playout based Monte-Carlo Tree Search, trying to answer the following question: how many (random) leaf evaluations are needed to be able to find a good move at the root of the game tree?

As will be seen shortly, in the proposed model this question can be framed as an active identification problem in a multi-armed bandit model. Since the seminal work of Kocsis and Szepesvári (2006) who proposed the UCT algorithm (for Upper Confidence bounds applied to Trees), selection rules based on bandit tools have been successfully used for MCTS, in particular for the design of the first computer Go programs (Coulom, 2006). UCT goes down the search tree by selecting the most promising children to explore using a UCB algorithm, that is, exploration is targeted toward the parts of the search tree that lead to successful playouts. However, the goal of an MCTS algorithm is not to win a large number of simulated games (i.e. maximize some notion of reward), but rather to identify the best first action to take when exploration in over. In light of the fundamental difference between algorithms for reward maximization and best arm identification that has been discussed in previous chapters, it is natural to wonder whether alternative approaches ignoring these rewards could have a better performance.

Despite the practical success of UCT, this algorithm has only been analyzed in terms of convergence towards a good recommendation when the number of sampled trajectories is large, and no sample complexity guarantees are available for this algorithm. With Aurélien Garivier and Wouter Koolen, we worked on new MCTS algorithms based on BAI tools and gave upper bounds on their sample complexity (Garivier et al., 2016a; Kaufmann and Koolen, 2017).

A simple model for MCTS in games We present an idealized model for Monte-Carlo Tree Search in a two-player game represented by a game tree $\mathcal{G}$. This tree models the possible action sequences by a collection of MAX nodes, that correspond to states in the game in which player A should take action, MIN nodes, for states in the game in which player B should take action, and leaves in which random evaluations of the positions can be performed. The root of this tree is a MAX node, and the goal is to find the best first action to take at the root for player A.

Letting $\mathcal{L}$ be the set of leaves of this tree, for each $\ell \in \mathcal{L}$ we introduce a stochastic oracle $\mathcal{O}_\ell$ that represents the playout performed when this leaf is reached by an MCTS algorithm. In this model, we do not try to optimize the evaluation or playout strategy, but we rather assume that the oracle $\mathcal{O}_\ell$ produces i.i.d. samples from an unknown distribution whose mean μ_ℓ is the value of the position ℓ. To ease the presentation, we focus on binary oracles (indicating the win or loss of a playout), in which the oracle $\mathcal{O}_\ell$ is a Bernoulli distribution with unknown mean μ_ℓ, which is the probability of player A winning the game in the corresponding state.

We denote by $\boldsymbol{\mu} = (\mu_\ell)_{\ell\in\mathcal{L}}$ the collection of leaf values. For each node s in the tree, we denote by $\mathcal{C}(s)$ the set of its children and by $\mathcal{P}(s)$ its parent. The root is denoted by s_0. The *value* (for player A) of any node s is recursively defined by $V_{\boldsymbol{\mu}}(\ell) = \mu_\ell$ if $\ell \in \mathcal{L}$ and

$$V_{\boldsymbol{\mu}}(s) = \begin{cases} \max_{c\in\mathcal{C}(s)} V_{\boldsymbol{\mu}}(c) & \text{if s is a MAX node,} \\ \min_{c\in\mathcal{C}(s)} V_{\boldsymbol{\mu}}(c) & \text{if s is a MIN node.} \end{cases}$$

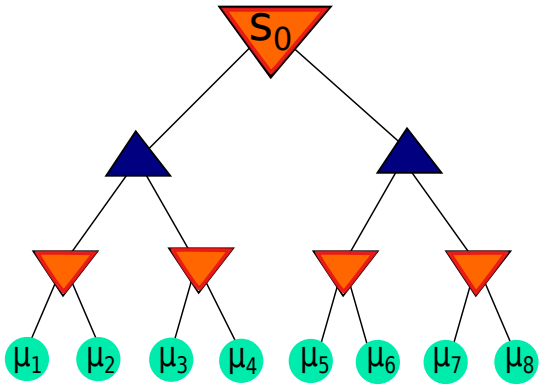


Figure 5.2 – A depth three maxmin tree with expected payoff in the leaves

Assuming that player B is strategic, an optimal move for player A is the action leading to a depth-one node with highest value,

$$s^\star \in \operatorname*{argmax}_{s \in \mathcal{C}(s_0)} V_{\boldsymbol{\mu}}(s) \ ,$$

so that $V_{\boldsymbol{\mu}}(s_0) = V_{\boldsymbol{\mu}}(s^\star)$. In order to identify a near-optimal move, an MCTS algorithm sequentially selects paths in the game tree and calls the corresponding leaf oracle. At round t, a leaf $L_t \in \mathcal{L}$ is chosen by this adaptive *sampling rule*, after which a sample $X_t \sim \mathcal{O}_{L_t}$ is collected. The strategy also requires a *stopping rule* τ, after which leaves are no longer evaluated, and a *recommendation rule* that outputs upon stopping a guess $\hat{s}_\tau \in \mathcal{C}(s_0)$ for the best move of player A.

Given a risk level $\delta \in (0, 1]$ and some accuracy parameter $\varepsilon \geq 0$ our goal is to have a recommendation $\hat{s}_\tau \in \mathcal{C}(s_0)$ whose value is within ε of the value of the best move, with probability larger than $1 - \delta$:

$$\mathbb{P}\left(V_{\boldsymbol{\mu}}(s^\star) - V_{\boldsymbol{\mu}}(\hat{s}_\tau) \leq \varepsilon\right) \geq 1 - \delta.$$

An algorithm satisfying this property is called (ε, δ)-correct. The main challenge is to design (ε, δ)-correct algorithms that minimize the *sample complexity*, that is, the number of leaf evaluations τ needed to make the recommendation.

This problem fits the active identification framework introduced in Chapter 3, as in a bandit model parameterized by the leaf values $\boldsymbol{\mu} = (\mu_\ell)_{\ell \in \mathcal{L}}$ the goal it to identify a region $\mathcal{R}_s = \{\boldsymbol{\mu} : V_{\boldsymbol{\mu}}(s) \geq V_{\boldsymbol{\mu}}(s^\star) - \varepsilon\}$ for $s \in \mathcal{C}(s_0)$ to which $\boldsymbol{\mu}$ belongs by adaptively sampling the arms of means $\boldsymbol{\mu}$. Note however that the regions are expressed with the value function $V_{\boldsymbol{\mu}}$ which depends in a complex way of the means $\boldsymbol{\mu}$. This problem can be viewed as a particular instance of the structured best arm identification framework of Huang et al. (2017), which was also introduced as a possible model for MCTS.

5.2 Monte-Carlo Tree Search by Best Arm Identification

We now present the UGapE-MCTS algorithm, which fits the more general BAI-MCTS framework that we introduced in Kaufmann and Koolen (2017), and later discuss optimal MCTS algorithms.

5.2.1 UGapE-MCTS

Recall that our goal is to identify the action at the root that leads to the state at depth one that has largest value. If one had access to a stochastic oracle producing i.i.d. samples with means $V(s)$ for each $s \in \mathcal{C}(s_0)$ this would be a standard best arm identification problem. However, the difficulty in the MCTS

problem comes from the fact that we can only sample the leaves to refine our estimates about the values of the target depth-one nodes.

Still, UGapE-MCTS builds on a best arm identification algorithm called UGapE (Gabillon et al., 2012) (designed for finding the m best arms but used here for $m = 1$) for choosing the first step towards a leaf. The specificity of UGapE is that its stopping, sampling and recommendation rules are fully determined by confidence intervals on the means of the arms. Hence, used in conjunction with *confidence intervals on the values at depth one*, UGapE permits to select a promising depth-one node to explore. This is the first step of the two-staged UGapE-MCTS, which then select a *representative leaf* to sample from the selected depth-one node.

Before giving a more precise description of UGapE-MCTS, we elaborate on its two central elements: the construction of the confidence intervals and the notion of a representative leaf.

Confidence intervals and representative nodes For each leaf $\ell \in \mathcal{L}$, using the past (i.i.d.) observations from this leaf we may build a confidence interval

$$\mathcal{I}_\ell(t) = [\mathrm{LCB}_\ell(t), \mathrm{UCB}_\ell(t)],$$

where $\mathrm{UCB}_\ell(t)$ (resp. $\mathrm{LCB}_\ell(t)$) is an Upper Confidence Bound (resp. a Lower Confidence Bound) on the value $V(\ell) = \mu_\ell$. The specific confidence interval we shall use will be discussed later.

These confidence intervals are then propagated upwards in the tree using the following construction. For each internal node s, we recursively define $\mathcal{I}_s(t) = [\mathrm{LCB}_s(t), \mathrm{UCB}_s(t)]$ with

$$\mathrm{LCB}_s(t) = \begin{cases} \max_{c\in\mathcal{C}(s)} \mathrm{LCB}_c(t) & \text{for a MAX node } s, \\ \min_{c\in\mathcal{C}(s)} \mathrm{LCB}_c(t) & \text{for a MIN node } s, \end{cases}$$

$$\mathrm{UCB}_s(t) = \begin{cases} \max_{c\in\mathcal{C}(s)} \mathrm{UCB}_c(t) & \text{for a MAX node } s, \\ \min_{c\in\mathcal{C}(s)} \mathrm{UCB}_c(t) & \text{for a MIN node } s. \end{cases}$$

Note that these intervals are the tightest possible on the parent under the sole assumption that the confidence intervals of the children are all valid. A similar construction was used in the OMS algorithm of Borsoniu et al. (2014) in a different context. It is easy to convince oneself (or prove by induction) that the accuracy of the confidence intervals is preserved under this construction, as stated below.

Proposition 5.1. *Let $t \in \mathbb{N}$. One has* $\bigcap_{\ell\in\mathcal{L}} (\mu_\ell \in \mathcal{I}_\ell(t)) \;\Rightarrow\; \bigcap_{s\in\mathcal{G}} (V_{\boldsymbol{\mu}}(s) \in \mathcal{I}_s(t))$.

We further define the *representative child* $c_s(t)$ of an internal node s as

$$c_s(t) \in \begin{cases} \mathrm{argmax}_{c\in\mathcal{C}(s)} \mathrm{UCB}_c(t) & \text{if } s \text{ is a MAX node}, \\ \mathrm{argmin}_{c\in\mathcal{C}(s)} \mathrm{LCB}_c(t) & \text{if } s \text{ is a MIN node}, \end{cases}$$

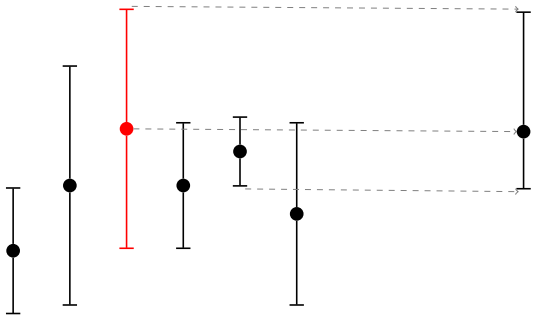
and the *representative leaf* $\ell_s(t)$ of a node $s \in \mathcal{G}$, which is the leaf obtained when going down the tree by always selecting the representative child:

$$\ell_s(t) = s \;\text{ if } s \in \mathcal{L}, \qquad \ell_s(t) = \ell_{c_s(t)}(t) \;\text{ otherwise.}$$

The confidence intervals in the tree represent the statistically plausible values in each node, hence the representative child can be interpreted as an "optimistic move" in a MAX node and a "pessimistic move" in a MIN node (assuming we play against the best possible adversary). This is reminiscent of the behavior of the UCT algorithm (Kocsis and Szepesvári, 2006), which uses however different types of confidence intervals. In UCT, the confidence interval used in a given node is built as if the outcomes of the playouts performed in the whole sub-tree rooted at this node were i.i.d. whereas in UGapE-MCTS the confidence

intervals are built in the leaves only (in which the observations are indeed i.i.d.) and propagated in the tree.

The construction of confidence intervals and the representative child are illustrated in Figure 5.3 while Figure 5.4 illustrates the confidence intervals in the whole tree as well as the path down to the representative leaf.



(a) Children (b) Parent

Figure 5.3 – Construction of confidence interval and representative child (in red) for a MAX node.

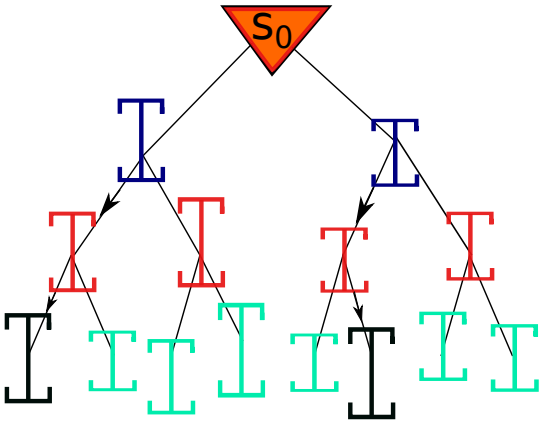


Figure 5.4 – Confidence intervals in the whole tree built recursively from the leaves. Arrows show the path down to the representative leaf from each depth-one node.

UGapE-MCTS Following what is done in the UGapE algorithm of Gabillon et al. (2012), we introduce for each depth-one node the index

$$B_s(t) = \max_{s' \in \mathcal{C}(s_0) \setminus \{s\}} \mathrm{UCB}_{s'}(t) - \mathrm{LCB}_s(t).$$

$B_s(t)$ is meant to be an upper confidence bound on the simple regret $V_{\boldsymbol{\mu}}(s_\star) - V_{\boldsymbol{\mu}}(s)$. Next, we define the following two promising depth-one nodes, a candidate best arm $\underline{b}_t$ and a challenger $\underline{c}_t$:

$$\underline{b}_t = \operatorname*{argmin}_{a \in \mathcal{C}(s_0)} B_a(t) \quad \text{and} \quad \underline{c}_t = \operatorname*{argmax}_{b \in \mathcal{C}(s_0) \setminus \{\underline{b}_t\}} \mathrm{UCB}_b(t).$$

At round $t+1$, UGapE-MCTS first computes and selects as the first node to explore the one among b_t and c_t whose confidence interval is the largest (that is, the most uncertain node):

$$R_{t+1} = \operatorname*{argmax}_{i \in \{\underline{b}_t, \underline{c}_t\}} \left[\mathrm{UCB}_i(t) - \mathrm{LCB}_i(t)\right].$$

Then the algorithm goes down the tree and samples the representative leaf of R_{t+1}: $L_{t+1} = \ell_{R_{t+1}}(t)$. UGapE-MCTS stops when the confidence intervals of the two promising arms overlap by less than ε:

$$\tau = \inf\left\{t \in \mathbb{N} : \mathrm{UCB}_{\underline{c}_t}(t) - \mathrm{LCB}_{\underline{b}_t}(t) < \varepsilon\right\},$$

and it recommends $\hat{s}_\tau = \underline{b}_\tau$.

Analysis of UGapE-MCTS The confidence interval for the mean of each leaf μ_ℓ depends on $N_\ell(t)$, the number of visits of this leaf in the first t rounds, and $\hat{\mu}_\ell(t)$ the empirical mean of the playouts from this leaf. We analyze UGapE-MCTS based on the following confidence intervals:

$$\mathrm{LCB}_\ell(t) = \max\{q : N_\ell(t)\mathrm{kl}(\hat{\mu}_\ell(t), q) \le \beta(N_\ell(t), \delta)\} \tag{5.1}$$

$$\text{and } \mathrm{UCB}_\ell(t) = \min\{q : N_\ell(t)\mathrm{kl}(\hat{\mu}_\ell(t), q) \le \beta(N_\ell(t), \delta)\}\ , \tag{5.2}$$

where kl is the binary relative entropy and $\beta(s,\delta)$ is some *exploration function*. An interesting practical feature of these confidence intervals is that they only depend on the local number of draws $N_\ell(t)$, whereas most of the confidence-based BAI algorithms use exploration functions that depend on the number of rounds t. Hence the only confidence intervals that need to be updated at round t are those of the ancestors of the selected leaf, which can be done recursively.

Using Theorem 3.1 in Chapter 3, the threshold can be tuned to ensure that the event

$$\mathcal{E} = \bigcap_{\ell \in \mathcal{L}} (\mu_\ell \in [\mathrm{LCB}_\ell(t), \mathrm{UCB}_\ell(t)])$$

holds with probability $1-\delta$. On this event, it can be observed that by definition of the stopping rule, the algorithm outputs an action which is ε-optimal, which yields the following result.

Lemma 5.2. *Letting $\mathcal{T}(x)$ be the threshold function defined in (3.7) in Chapter 3. Choosing*

$$\beta(s,\delta) = \mathcal{T}\left(\log\left(\frac{|\mathcal{L}|}{\delta}\right)\right) + 3\log(\log s + 1) \tag{5.3}$$

UGapE-MCTS *satisfies* $\mathbb{P}(V(s^\star) - V(\hat{s}_\tau) \le \varepsilon) \ge 1-\delta$.

To introduce our sample complexity guarantees, we first introduce some notation. Recall that $s^\star$ is the optimal action at the root, identified with the depth-one node satisfying $V(s^\star) = V(s_0)$, and define the second-best depth-one node as $s_2^\star = \mathrm{argmax}_{s \in \mathcal{C}(s_0)\setminus\{s^\star\}} V(s)$. Recall $\mathcal{P}(s)$ denotes the parent of a node s different from the root. Introducing furthermore the set $\mathrm{Anc}(s)$ of all the ancestors of a node s, we define the complexity term by

$$H_\varepsilon^\star(\boldsymbol{\mu}) := \sum_{\ell \in \mathcal{L}} \frac{1}{\Delta_\ell^2 \vee \Delta_*^2 \vee \varepsilon^2}, \quad \text{where} \quad \begin{array}{rcl} \Delta_* & := & V(s^\star) - V(s_2^\star) \\ \Delta_\ell & := & \max_{s \in \mathrm{Anc}(\ell)\setminus\{s_0\}} |V(s) - V(\mathcal{P}(s))| \end{array} \tag{5.4}$$

The intuition behind these squared terms in the denominator is the following. We will sample a leaf ℓ until we either prune it (by determining that it or one of its ancestors is a bad move), prune everyone else (this happens for leaves below the optimal arm) or reach the required precision ε.

Theorem 5.3. *Let $\delta \in (0,1)$. UGapE-MCTS using the exploration function (5.3) is such that, with probability larger than $1-\delta$, $(V(s^\star) - V(\hat{s}_\tau) < \varepsilon)$ and, letting $\overline{\Delta}_{\ell,\varepsilon} = \Delta_\ell \vee \Delta_* \vee \varepsilon$,*

$$\tau \le \sum_{\ell \in \mathcal{L}} \frac{8}{\overline{\Delta}_{\ell,\varepsilon}^2}\left(\mathcal{T}\left(\log\left(\frac{|\mathcal{L}|}{\delta}\right)\right) + 4\log\log\frac{1}{\overline{\Delta}_{\ell,\varepsilon}^2}\right) + 32 H_\varepsilon^\star(\boldsymbol{\mu})\log\log\left(8e\mathcal{T}\left(\log\frac{|\mathcal{L}|}{\delta}\right)\right) + 1\ .$$

Recalling that $\mathcal{T}(x) \sim x + \log\log(x)$, with high probability the sample complexity of UGapE-MCTS is of order $\mathcal{O}\left(H_\varepsilon^\star(\boldsymbol{\mu}) \log\left(\frac{1}{\delta}\right)\right)$. This sample complexity improves over the one of the FindTopWinner algorithm previously proposed by Teraoka et al. (2014), which is based on eliminations and stops exploring a sub-tree rooted in a particular node if a gap between this node and its parents is detected. Numerical simulations reported in Kaufmann and Koolen (2017) also reveal that UGapE-MCTS greatly outperforms FindTopWinner in practice.

The independent work of Huang et al. (2017) proposed the LUCBMinMax algorithm, which is very similar to UGapE-MCTS. The two algorithms only differ in the way the best guess $\underline{b}_t$ is picked. The analysis is very similar to ours, but features some refined complexity measure, in which Δ_ℓ (which is the maximal distance between *consecutive* ancestors of the leaf, see (5.4)) is replaced by the maximal distance between *any* ancestors of that leaf. We note that an improvement along these lines could be obtained in our analysis, by updating Lemma 5.5, as explained in the proof below.

Sketch of proof of Theorem 5.3. Letting $\mathcal{E}_t = \bigcap_{\ell \in \mathcal{L}} \left(\mu_\ell \in \mathcal{I}_\ell(t)\right)$ and $\mathcal{E} = \bigcap_{t\in\mathbb{N}} \mathcal{E}_t$, we upper bound τ assuming the event $\mathcal{E}$ holds. To do so, we shall relate the number of selections of each leaf $N_\ell(\tau)$ to the associated gap $\overline{\Delta}_{\ell,\varepsilon} = \Delta_\ell \vee \Delta_* \vee \varepsilon$.

A first observation is the following important consequence of the definition of the representative leaf, which says that along a path $s_0, s_1, \ldots, s_D$ to a leaf $\ell = s_D$, the confidence intervals are nested.

Lemma 5.4. *Let $t \in \mathbb{N}$ and $s_0, s_1, \ldots, s_D$ be a path from the root down to a leaf $\ell = s_D$.*

$$\left(\ell_{s_1}(t) = s_D\right) \Rightarrow \left(\forall k = 2, \ldots, D, \quad \mathcal{I}_{s_{k-1}}(t) \subseteq \mathcal{I}_{s_k}(t)\right) .$$

As a consequence, on the event $\mathcal{E}$, if a path $s_0, s_1, \ldots, s_D = \ell$ is selected by UGapE-MCTS at round $t+1$, the interval $\mathcal{I}_{L_{t+1}}(t)$ – whose width is upper bounded by $\sqrt{2\beta(N_\ell(t),\delta)/N_\ell(t)}$ by Pinsker's inequality – contains all the values $V(s_1), \ldots, V(s_D)$. Hence the following result.

Lemma 5.5. *Let $t \in \mathbb{N}$ and $s_0, s_1, \ldots, s_D$ be a path from the root down to a leaf $\ell = s_D$. If $\mathcal{E}_t$ holds and ℓ is selected at round $t+1$ then*

$$\max_{k=2\ldots D} |V(s_k) - V(s_{k-1})| \le \sqrt{\frac{2\beta(N_\ell(t),\delta)}{N_\ell(t)}}.$$

We note that this lemma could be tightened to obtain a sample complexity in the spirit of the one given by Huang et al. (2017): as all the values $V(s_k)$ belong to $\mathcal{I}_\ell(t)$ for $k = 1, \ldots, D$, the width of this confidence interval is also larger than the maximal distance between any two $V(s_k)$, and not only consecutive values.

The next step of the analysis is a consequence of some properties of the UGapE mechanism used at depth one, which relates the width of $\mathcal{I}_{R_{t+1}}(t)$ (and thus that of $\mathcal{I}_{L_{t+1}}(t)$) to the gap in depth one.

Lemma 5.6. *If $\mathcal{E}_t$ holds, $t < \tau$ and $R_{t+1} = s_1$, then*

$$\left(V(s_0) - V(s_1)\right) \vee \Delta_* \vee \varepsilon \le 2\left(\mathrm{UCB}_{s_1}(t) - \mathrm{LCB}_{s_1}(t)\right).$$

Using that

$$\mathrm{UCB}_{R_{t+1}}(t) - \mathrm{LCB}_{R_{t+1}}(t) \le \mathrm{UCB}_{L_{t+1}}(t) - \mathrm{LCB}_{L_{t+1}}(t) \le \sqrt{\frac{2\beta(N_\ell(t),\delta)}{N_\ell(t)}}$$

by Lemma 5.4 and Pinsker's inequality, it follows from Lemma 5.5 and Lemma 5.6 that if the path $s_0, s_1, \ldots, s_D = \ell$ is selected in round $t+1$ when $\mathcal{E}_t$ holds and $t < \tau$,

$$\left(V(s_0) - V(s_1)\right) \vee \Delta_* \vee \varepsilon \vee \max_{k=2\ldots D} |V(s_k) - V(s_{k-1})| \le 2\sqrt{\frac{2\beta(N_\ell(t),\delta)}{N_\ell(t)}}.$$

From the definition of Δ_ℓ, this is equivalent to the following statement.

Lemma 5.7. *Let* $t \in \mathbb{N}$. $\mathcal{E}_t \cap (\tau > t) \cap (L_{t+1} = \ell) \quad \Rightarrow \quad N_\ell(t) \leq \frac{8\beta(N_\ell(t),\delta)}{\Delta_\ell^2 \vee \Delta_*^2 \vee \varepsilon^2}$.

The last tool for the proof is the following lemma (which is a variant of Lemma 6 in Kaufmann and Koolen (2017)), which permits to invert the above inequality and find the largest possible value of $N_\ell(t)$ for which it holds, for the specific exploration rate in (5.3).

Lemma 5.8. *Let* $\beta(s) = C + 3\log(1+\log(s))$ *and define* $S = \sup\{s \geq 1 : a\beta(s) \geq s\}$. *Then*

$$S \leq aC + 4a\log(1+\log(aC)).$$

On the event $\mathcal{E}$, letting τ_ℓ be the last instant before τ at which the leaf ℓ has been played before stopping, one has $N_\ell(\tau-1) = N_\ell(\tau_\ell)$ that satisfies by Lemma 5.7

$$N_\ell(\tau_\ell) \leq \frac{8\beta(N_\ell(\tau_\ell),\delta)}{\Delta_\ell^2 \vee \Delta_*^2 \vee \varepsilon^2}.$$

Applying Lemma 5.8 with $a = a_\ell = \frac{8}{\Delta_\ell^2 \vee \Delta_*^2 \vee \varepsilon^2}$ and $C = \mathcal{T}\left(\log\left(\frac{|\mathcal{L}|}{\delta}\right)\right)$ leads to

$$N_\ell(\tau-1) \leq \frac{8}{\Delta_\ell^2 \vee \Delta_*^2 \vee \varepsilon^2}\left[\mathcal{T}\left(\log\left(\frac{|\mathcal{L}|}{\delta}\right)\right) + 4\log\left(1+\log\left(\frac{8}{\Delta_\ell^2 \vee \Delta_*^2 \vee \varepsilon^2}\mathcal{T}\left(\log\left(\frac{|\mathcal{L}|}{\delta}\right)\right)\right)\right)\right].$$

Recalling that $\overline{\Delta}_{\ell,\varepsilon} = \Delta_\ell \vee \Delta_* \vee \varepsilon$ and summing over arms, we find

$$\begin{aligned}
\tau &= 1 + \sum_\ell N_\ell(\tau-1) \\
&\leq 1 + \sum_\ell \frac{8}{\overline{\Delta}_{\ell,\varepsilon}^2}\left[\mathcal{T}\left(\log\left(\frac{|\mathcal{L}|}{\delta}\right)\right) + 4\log\left(1+\log\left(\frac{8}{\overline{\Delta}_{\ell,\varepsilon}^2}\mathcal{T}\left(\log\left(\frac{|\mathcal{L}|}{\delta}\right)\right)\right)\right)\right] \\
&= 1 + \sum_\ell \frac{8}{\overline{\Delta}_{\ell,\varepsilon}^2}\left(\mathcal{T}\left(\log\left(\frac{|\mathcal{L}|}{\delta}\right)\right) + 4\log\log\frac{1}{\overline{\Delta}_{\ell,\varepsilon}^2}\right) + 32H_\varepsilon^\star(\boldsymbol{\mu})\log\log\left(8e\mathcal{T}\left(\log\frac{|\mathcal{L}|}{\delta}\right)\right).
\end{aligned}$$

To conclude the proof, we use that, as already noted above, $V(s^\star) - V(\hat{s}_\tau) < \varepsilon$ on the event $\mathcal{E}$ and that the threshold $\beta(s,\delta)$ was calibrated in Lemma 5.2 so that $\mathcal{E}$ holds with probability larger than $1-\delta$.

□

5.2.2 Towards Optimal Strategies

By leveraging a sub-optimal best arm identification algorithm, UGapE, to propose an algorithm for the MCTS problem, one may be quite far from designing a MCTS algorithm with a *minimal* sample complexity. This is corroborated by the fact that the complexity terms $H_\varepsilon^\star(\boldsymbol{\mu})$ features a sum over *all* leaves, while one could expect only a few leaves to be useful, as in the best case of $\alpha-\beta$ pruning.

As the MCTS is a particular active identification problem, to measure the room for improvement, one can write the lower bound of Theorem 4.1, at least in the special case $\varepsilon = 0$.

Theorem 5.9. *Assume* $\varepsilon = 0$ *and define* $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} \in [0,1]^{|\mathcal{L}|} : s^\star(\boldsymbol{\lambda}) \neq s^\star(\boldsymbol{\mu})\}$. *Any* δ*-correct algorithm satisfies*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq T^\star(\boldsymbol{\mu})\log\left(\frac{1}{3\delta}\right), \quad \textit{where} \quad T^\star(\boldsymbol{\mu})^{-1} := \sup_{\boldsymbol{w}\in\Sigma_{|\mathcal{L}|}} \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu})} \sum_{\ell\in\mathcal{L}} w_\ell \mathrm{kl}(\mu_\ell,\lambda_\ell) \tag{5.5}$$

with $\Sigma_k = \{\boldsymbol{w} \in [0,1]^k : \sum_{i=1}^k w_i = 1\}$ *and* $\mathrm{kl}(x,y) = x\log(x/y) + (1-x)\log((1-x)/(1-y))$.

To compare $T^\star(\boldsymbol{\mu})$ and $H_0^\star(\boldsymbol{\mu})$, a more explicit expression of the former quantity is needed. For depth-two trees with K (resp. M) actions for player A (resp. B), $T^\star(\boldsymbol{\mu})$ can be made more explicit, revealing an interesting *sparsity pattern*.

Lemma 5.10. *Assume $\varepsilon = 0$ and consider a tree of depth two with $\boldsymbol{\mu} = (\mu_{i,j})_{1\le i\le K, 1\le j\le M}$ such that $\forall (i,j), \mu_{1,1} > \mu_{i,1},\ \mu_{i,1} < \mu_{i,j}$. The supremum in the definition of $T^\star(\boldsymbol{\mu})^{-1}$ can be restricted to*

$$\tilde{\Sigma}_{K,M} := \{\boldsymbol{w} \in \Sigma_{K\times M} : w_{i,j} = 0 \text{ if } i \ge 2 \text{ and } j \ge 2\}$$

and

$$T^\star(\boldsymbol{\mu})^{-1} = \max_{\boldsymbol{w}\in\tilde{\Sigma}_{K,M}} \min_{\substack{i=2,\dots,K \\ a=1,\dots,M}} \left[w_{1,a} d\left(\mu_{1,a}, \frac{w_{1,a}\mu_{1,a} + w_{i,1}\mu_{i,1}}{w_{1,a} + w_{i,1}}\right) + w_{i,1}\mathrm{kl}\left(\mu_{i,1}, \frac{w_{1,a}\mu_{1,a} + w_{i,1}\mu_{i,1}}{w_{1,a} + w_{i,1}}\right)\right].$$

For depth-two trees, the fact that the optimal weights $\boldsymbol{w}^\star(\boldsymbol{\mu})$ are supported on only $K + M - 1$ of the $K \times M$ leaves suggests that matching algorithms should draw many of the leaves *much less than* $O(\log(1/\delta))$ *times*. This shows that the complexity quantity $H_0^\star(\boldsymbol{\mu})$, which scales in $\frac{KM}{\Delta_\star^2}$ for depth-two trees, is *not* optimal in the case (assuming the lower bound can be reached).

In order to design an asymptotically optimal algorithm, one could try to follow a similar approach as that of Track-and-Stop for best arm identification, described in Chapter 4: use a sampling rule tracking the optimal weights (with some forced exploration) together with the Parallel GLRT as a stopping rule. We first tried this approach for depth-two trees with $K = 2$ actions for the root player, for which we could obtain an efficient algorithm for computing the optimal weights in the lines of the one proposed in Theorem 4.3 for best arm identification. However, the resulting sample complexity was not significantly smaller than that of UGapE-MCTS, unlike what happens for best arm identification where the sample complexity of Track-and-Stop is twice smaller than that of UGapE.

Beyond $K = 2$ and beyond depth-two trees, computing the optimal weights is more involved. Using software for disciplined optimization (CVX), Wouter Koolen wrote a program for computing the optimal weights for depth-three trees. We found out that, as in the depth-two cases, a lot of these optimal weights are indeed zero. However we couldn't provide a characterization of their support as we do in Lemma 5.10 for depth-two trees.

Hence, finding an asymptotically optimal algorithm for MCTS that is both numerically efficient and has a small empirical sample complexity for moderate values of δ remains an open question.

5.3 From Thompson Sampling to Murphy Sampling

A crucial component of the UGapE-MCTS algorithm is the construction of confidence intervals on the values $V(s)$ of all nodes in the tree. UGapE-MCTS relies on a simple construction that builds an upper confidence bound on the minimum (resp. a lower confidence bound on the maximum) of several values by taking the minimum of the upper confidence bounds on these values (resp. the maximum of the lower confidence bounds). Motivated by possible improvements to UGapE-MCTS, we worked with Wouter Koolen and Aurélien Garivier on improved confidence intervals on the minimum and maximum, and on a related active identification problem: that of finding whether the minimum of a set of means is smaller or larger than a threshold. This work was published in Kaufmann et al. (2018) and we now describe its key features.

Given a finite collection of probability distributions in a one-parameter exponential family parameterized by their means $\boldsymbol{\mu} = (\mu_1, \dots, \mu_A) \in \mathcal{I}^A$, we are interested in learning about $\mu_{\min} = \min_a \mu_a$ from

adaptive samples $X_t \sim \mu_{A_t}$, where A_t indicates the distribution sampled at time t. More precisely, given a threshold $\gamma \in \mathcal{I}$, our goal is to decide whether $\mu_{\min} < \gamma$ or $\mu_{\min} > \gamma$. We introduce the regions

$$\mathcal{R}_< = \left\{\boldsymbol{\mu} \in \mathcal{I}^A : \mu_{\min} < \gamma\right\} \text{ and } \mathcal{R}_> = \left\{\boldsymbol{\mu} \in \mathcal{I}^A : \mu_{\min} > \gamma\right\}, \text{ and their union } \mathcal{R} = \mathcal{R}_< \cup \mathcal{R}_>$$

We want to propose a sequential and adaptive testing procedure, that consists of a *sampling rule* A_t, a *stopping rule* τ and a *decision rule* $\hat{m} \in \{<, >\}$. The algorithm samples $X_t \sim \mu_{A_t}$ while $t \leq \tau$, and then outputs a decision $\hat{m}$. Given a risk parameter $\delta \in (0, 1]$, we aim for a δ-*correct* algorithm, that satisfies $\mathbb{P}_{\boldsymbol{\mu}}(\boldsymbol{\mu} \in \mathcal{R}_{\hat{m}}) \geq 1 - \delta$ for all $\boldsymbol{\mu} \in \mathcal{R}$. Our goal is to build δ-correct algorithms that use a small number of samples τ_δ in order to reach a decision. In particular, we want the *sample complexity* $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ to be small.

This active identification problem can have different types of applications. For example in e-learning, we may want to certify that a given student has sufficient understanding of a range of subjects, asking as few questions as possible about the different subjects. Then in anomaly detection, the more anomalies are present, the faster we may want to flag the presence of an anomaly. Yet, our motivation for studying this problem came from the modeling of a simple Monte-Carlo Tree Search problem, in which the value of a MIN node should be compared to the known values of its neighbors, as illustrated in Figure 5.5. Note that we study minimums to fix the ideas, but one could develop similar methods to adaptively compare the maximum of a bunch of means to a threshold.

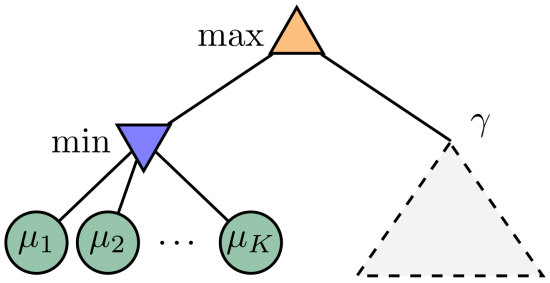


Figure 5.5 – Game tree search problem of "depth $1 + 1/2$". We consider the scenario where it has been established that the right subtree (grey) of the root has value γ. Learning the optimal action at the root (orange) is equivalent to determining how the minimum (blue) of the leaf means (green) compares to γ.

Computing the lower bound For the active identification problem studied in this section, it is actually possible to provide an *explicit* expression of the sample complexity lower bound —and the associated optimal proportions— which can be derived with the methodology presented in Chapter 3.

Lemma 5.11. *Any δ-correct strategy satisfies $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T^\star(\boldsymbol{\mu})\mathrm{kl}(\delta, 1-\delta)$ with*

$$T^\star(\boldsymbol{\mu}) = \begin{cases} \frac{1}{\mathrm{kl}(\mu_{\min}, \gamma)} & \text{if } \mu_{\min} < \gamma, \\ \sum_{a=1}^{A} \frac{1}{\mathrm{kl}(\mu_a, \gamma)} & \text{if } \mu_{\min} > \gamma. \end{cases}$$

The vector of optimal proportions $\boldsymbol{w}^\star(\boldsymbol{\mu})$ [1] *is such that*

$$\forall a \in [A],\ w_a^\star(\boldsymbol{\mu}) = \begin{cases} \frac{\mathbb{1}(\mu_a = \mu_{\min})}{\sum_{j=1}^{A} \mathbb{1}(\mu_j = \mu_{\min})} & \text{if } \mu_{\min} < \gamma, \\ \frac{\frac{1}{\mathrm{kl}(\mu_a, \gamma)}}{\sum_{j=1}^{A} \frac{1}{\mathrm{kl}(\mu_j, \gamma)}} & \text{if } \mu_{\min} > \gamma. \end{cases}$$

1. If the arm with smallest mean is not unique, any probability distribution whose support is included in the set of arms with mean $\mu_{\min}$ is an optimal allocation, not just the example given

Proof. We recall from Chapter 4 that

$$(T^\star(\boldsymbol{\mu}))^{-1} = \sup_{w \in \Sigma_A} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{A} w_a \mathrm{kl}(\mu_a, \lambda_a)$$

where for this active identification problem $\mathrm{Alt}(\boldsymbol{\mu}) = \mathcal{R}_<$ (resp. $\mathcal{R}_>$) if $\boldsymbol{\mu} \in \mathcal{R}_>$ (resp. $\mathcal{R}_<$).

If $\mu_{\min} < \gamma$, then we find

$$\frac{1}{T^\star(\boldsymbol{\mu})} = \max_{\boldsymbol{w} \in \Sigma_A} \sum_{a:\mu_a < \gamma} w_a \mathrm{kl}(\mu_a, \gamma) = \max_{a:\mu_a < \gamma} \mathrm{kl}(\mu_a, \gamma) = \mathrm{kl}(\mu_{\min}, \gamma)$$

and a maximizer is obtained for any distribution supported on arms a whose mean is equal to $\mu_{\min}$. On the other hand, if $\mu_{\min} > \gamma$, we find

$$\frac{1}{T^\star(\boldsymbol{\mu})} = \max_{\boldsymbol{w} \in \Sigma_A} \min_a \; w_a \, \mathrm{kl}(\mu_a, \gamma) = \frac{1}{\sum_a \frac{1}{\mathrm{kl}(\mu_a, \gamma)}}$$

and the maximum is attained for $w_a^\star(\boldsymbol{\mu}) = \frac{\frac{1}{\mathrm{kl}(\mu_a, \gamma)}}{\sum_j \frac{1}{\mathrm{kl}(\mu_j, \gamma)}}$.

□

As explained in Chapter 3, the oracle weights $\boldsymbol{w}^\star(\boldsymbol{\mu})$ correspond to the fraction of samples that should be allocated to each arm under a strategy matching the lower bound. The interesting feature here is that the lower bound indicates that an oracle algorithm should have very different behavior depending on whether $\boldsymbol{\mu}$ belongs to $\mathcal{R}_<$ or to $\mathcal{R}_>$. On $\mathcal{R}_<$ it should sample the arm with lowest mean (or all lowest means, if there are several) exclusively, while on $\mathcal{R}_>$ it should sample *all* arms with certain specific proportions.

In light of the lower bound in Lemma 5.11, we now investigate the design of optimal learning algorithms (sampling rule A_t and stopping rule τ). To ease the presentation, we assume that there is a unique arm with smallest mean, which we denote by $a_{\min}$.

5.3.1 Sampling Rule: Murphy Sampling

The key contribution of this work is a new sampling rule under which the empirical fraction of selections of each arm, $N_a(t)/t$, converges to the corresponding optimal proportion $w_a^\star(\boldsymbol{\mu})$, for any $\boldsymbol{\mu} \in \mathcal{R}$. Unlike previously proposed asymptotically optimal algorithms for active identification, this sampling rule does not rely on the Tracking idea, hence doesn't need any forced exploration. Before presenting Murphy Sampling, we explain that it is easy to propose sampling rules that converge to $\boldsymbol{w}^\star(\boldsymbol{\mu})$ for either $\boldsymbol{\mu} \in \mathcal{R}_<$ or $\boldsymbol{\mu} \in \mathcal{R}_>$, but not for both.

Thompson Sampling For $\boldsymbol{\mu} \in \mathcal{R}_<$, the empirical proportion of draws of the arm $a_{\min}$ with smallest mean should converge to 1. The literature on regret minimization provides candidate algorithms that have this type of behavior, for example Thompson Sampling. Given independent prior distributions on the mean of each arm, we recall that this Bayesian algorithm selects an arm at random according to its posterior probability of being optimal (in our case, the arm with smallest mean). Letting π_a^t refer to the posterior distribution of μ_a after t samples, this can be implemented as

TS: Sample $\forall a \in [A], \theta_a(t) \sim \pi_a^{t-1}$, then play $A_t = \arg\min_{a \in [A]} \; \theta_a(t)$.

If Thompson Sampling is run without stopping, one can show (see Theorem 12 in Kaufmann et al. (2018)) that $N_{a_{\min}}(t)/t$ converges almost surely to 1, for every $\boldsymbol{\mu} \in \mathcal{R}$. Hence, TS converges to the optimal proportions for $\boldsymbol{\mu} \in \mathcal{R}_<$. However, TS cannot be optimal for $\boldsymbol{\mu} \in \mathcal{R}_>$, as the empirical proportion of draws doesn't converge to $\boldsymbol{w}^\star(\boldsymbol{\mu}) \neq \mathbb{1}_{a_{\min}}$.

LCB For $\boldsymbol{\mu} \in \mathcal{R}_>$, one can show that a different algorithm, called LCB, can converge to the optimal proportions. It requires to define confidence regions, i.e. upper and lower confidence bounds

$$\mathrm{UCB}_a(t) = \max\{q : N_a(t) d^+(\hat{\mu}_a(t), q) \geq C_<(\delta, N_a(t))\}, \tag{5.6}$$

$$\mathrm{LCB}_a(t) = \min\{q : N_a(t) d^-(\hat{\mu}_a(t), q) \geq C_>(\delta, N_a(t))\}, \tag{5.7}$$

for two threshold functions $C_<(\delta, r)$ and $C_>(\delta, r)$. The LCB strategy selects at each round the arm with smallest Lower Confidence Bound:

$$\boxed{\texttt{LCB: Play } A_t = \mathrm{argmin}_a \, \mathrm{LCB}_a(t) \, .}$$

In Appendix E of Kaufmann et al. (2018) we prove that LCB is optimal for $\boldsymbol{\mu} \in \mathcal{R}_>$ however we show that on instances of $\mathcal{R}_<$ it draws all arms $a \neq a_{\min}$ too much and cannot converge to the optimal proportions.

Murphy Sampling We denote by $\Pi_n = \mathbb{P}(\cdot|\mathcal{F}_n)$ the posterior distribution of the mean parameters after n rounds. We introduce a new (randomised) sampling rule called *Murphy Sampling* after Murphy's Law, as it performs some conditioning on the "worst event" ($\boldsymbol{\mu} \in \mathcal{R}_<$):

$$\boxed{\text{MS: Sample } \boldsymbol{\theta}_t \sim \Pi_{t-1}(\cdot|\mathcal{R}_<), \text{ then play } A_t = a_{\min}(\boldsymbol{\theta}_t) \, .} \tag{5.8}$$

As we will argue below, the subtle difference of sampling from $\Pi_{n-1}(\cdot|\mathcal{R}_<)$ instead of Π_{n-1} (as regular Thompson Sampling does) ensures the required split personality behavior. Note that MS always conditions on $\mathcal{R}_<$ (and never on $\mathcal{R}_>$) regardless of the position of $\boldsymbol{\mu}$ w.r.t. γ. This is different from the symmetric Top Two Thompson Sampling (Russo, 2016), which essentially conditions on $a^\star(\theta) \neq a^\star(\boldsymbol{\mu})$ a fixed fraction $1 - \beta$ of the time, where β is a parameter that needs to be tuned with knowledge of $\boldsymbol{\mu}$. MS on the other hand needs no parameters. Also note that MS is an anytime sampling algorithm, being independent of the risk parameter δ.

MS is technically an instance of Thompson Sampling with a joint prior Π supported only on $\mathcal{R}_<$. This viewpoint is conceptually interesting, as we will apply MS identically to $\mathcal{R}_<$ *and* $\mathcal{R}_>$. To implement MS, we use that independent conjugate per-arm priors induce likewise posteriors, admitting efficient (unconditioned) posterior sampling. Rejection sampling then achieves the required conditioning. Its computational cost is limited: the acceptance probability cannot be much smaller than the risk δ provided to the algorithm. Indeed, the fact that the stopping rule (see Section 5.3.2) has not yet fired, combined with the posterior concentration (Proposition 5.13) and the convergence of the sampling efforts to track the sampling proportions (Theorem 5.12) reveals that the MS rejection sampling step accepts with probability at least of order $\delta/(\log t)^3$. So for reasonable values of δ, this can be small and require a few thousands of draws (not a big deal for today's computers), but it cannot be *prohibitively* small.

We now prove the convergence of Murphy Sampling under the same assumptions made by Russo (2016): the parameter space $\Theta \ni \boldsymbol{\mu}$ (or the support of the prior) is the interior of a bounded subset of $\mathbb{R}^A$. This ensures that $\sup_{\mu,\theta\in\Theta} \mathrm{kl}(\mu, \theta) < \infty$ and $\sup_{\mu,\theta\in\Theta} \|\mu - \theta\| < \infty$. We further assume that the prior Π has a density π with bounded ratio $\sup_{\mu,\theta\in\Theta} \frac{\pi(\theta)}{\pi(\mu)} < \infty$.

Theorem 5.12. *Under the above assumptions, Murphy Sampling ensures that for any* $\boldsymbol{\mu} \in \mathcal{R}$

$$\forall a \in [A], \; \frac{N_a(t)}{t} \to w_a^\star(\boldsymbol{\mu}) \;\; a.s. \, .$$

Proof. Following Russo (2016), we denote the sampling probabilities in round n by

$$\psi_a(n) = \mathbb{P}\left(A_n = a | \mathcal{F}_{n-1}\right) = \Pi_{n-1}\left(a = \arg\min_j \theta_j \middle| \mathcal{R}_<\right),$$

and abbreviate $\Psi_a(n) = \sum_{t=1}^n \psi_a(t)$ and $\bar{\psi}_a(n) = \frac{\Psi_a(n)}{n}$.

The first step of the analysis is to show that for all $\boldsymbol{\mu} \in \mathcal{R}$, under Murphy Sampling all arms are drawn infinitely often, that is $N_a(t) \to +\infty, a.s.$, which is proved in Proposition 11 of Kaufmann et al. (2018). An interesting corollary (which follows from a martingale convergence theorem as proved by Russo (2016)) is the fact that $\lim_{n\to\infty} \frac{\Psi_a(n)}{N_a(n)} = 1$ $a.s.$. Hence, the convergence of the empirical proportions to $\boldsymbol{w}^\star(\boldsymbol{\mu})$ is equivalent to the convergence of $\bar{\boldsymbol{\psi}}(t) = (\bar{\psi}_a(t))_{a\in[A]}$ to $\boldsymbol{w}^\star(\boldsymbol{\mu})$.

Then, consider $\boldsymbol{\mu} \in \mathcal{R}_<$. In this case the conditioning in MS is asymptotically immaterial as $\Pi_n(\mathcal{R}_<)$ converges to 1, and the algorithm behaves like regular Thompson Sampling (targeted toward minimizing rewards instead of maximizing them). As Thompson sampling has sublinear regret (Agrawal and Goyal, 2012), we should have $\mathbb{E}[N_{a_{\min}}(t)]/t \to 1$. The crux of the proof is to show that the convergence occurs almost surely. For this, we prove that $\bar{\psi}_{a_{\min}}(t) \to 1$, which follows from the fact that for all $\zeta \in (\mu_{\min}, \min_{a\neq a_{\min}} \mu_a)$,

$$\psi_{a_{\min}}(t) \geq \Pi_t\left(\arg\min_j \theta_j = a_{\min}, \min_j \theta_j \leq \gamma\right) \geq \underbrace{\Pi_t\left(\theta_{a_{\min}} \leq \zeta\right)}_{\to 1} \prod_{a\neq a_{\min}} \underbrace{\Pi_t\left(\theta_a \geq \zeta\right)}_{\to 1} \to 1.$$

The fact that $\Pi_t\left(\theta_{a_{\min}} \leq \zeta\right)$ and $\Pi_t\left(\theta_a \geq \zeta\right)$ converge to 1 for $a \neq a_{\min}$ follows from posterior convergence and the fact that all arms are drawn infinitely often.

Next, consider $\boldsymbol{\mu} \in \mathcal{R}_>$. Let us abbreviate $\boldsymbol{w}^\star = \boldsymbol{w}^\star(\boldsymbol{\mu})$. Our strategy is similar to that Russo (2016) for proving the convergence of TTTS. It relies on two important results stated below.

Lemma 5.13 ((Russo, 2016, Proposition 4))**.** *For any open subset $\tilde{\Theta} \subseteq \Theta$, the posterior concentrates at rate $\Pi_n(\tilde{\Theta}) \doteq \exp\left(-n \min_{\boldsymbol{\lambda}\in\tilde{\Theta}} \sum_a \bar{\psi}_a(n) \mathrm{kl}(\mu_a, \lambda_a)\right)$ a.s. where $a_n \doteq b_n$ means $\frac{1}{n}\log\frac{a_n}{b_n} \to 0$.*

Lemma 5.14 ((Russo, 2016, Simplified version of Lemma 11))**.** *Consider any sampling rule $(A_t)_t$. If for any arm $a \in [A]$ and all $c > 0$*

$$\sum_n \psi_a(n) \mathbb{1}\left(\bar{\psi}_a(n) \geq w_a^\star + c\right) < \infty\ ,$$

then $\bar{\boldsymbol{\psi}}(n) \to \boldsymbol{w}^\star$.

First, recall from Lemma 5.11 that

$$T^\star(\boldsymbol{\mu})^{-1} = \max_{\boldsymbol{w}} \min_{\boldsymbol{\lambda}:\min_a \lambda_a<\gamma} \sum_a w_a d\left(\mu_a, \lambda_a\right) = \max_{\boldsymbol{w}} \min_a w_a \mathrm{kl}\left(\mu_a, \gamma\right) = w_a^\star \mathrm{kl}(\mu_a, \gamma) \quad \forall a. \tag{5.9}$$

Furthermore, by Lemma 5.13, for any $a \in [A]$

$$\Pi_n(\theta_a < \gamma) \doteq \exp\left(-n \min_{\boldsymbol{\lambda}:\lambda_a<\gamma} \sum_b \bar{\psi}_b(n) \mathrm{kl}\left(\mu_b, \lambda_b\right)\right) = \exp\left(-n\bar{\psi}_a(n) d\left(\mu_a, \gamma\right)\right).$$

In particular, there is a sequence ε_n decreasing to zero such that

$$\forall n: \quad \Pi_n(\theta_a < \gamma) \in \exp\left(-n\left(\bar{\psi}_a(n) d\left(\mu_a, \gamma\right) \pm \varepsilon_n\right)\right).$$

To establish the precondition of Lemma 5.14 above, fix $a \in [A]$ and $c > 0$ and consider any round n where $\bar{\psi}_a(n) \geq w_a^\star + c$. Then

$$\psi_a(n) = \frac{\Pi_{n-1}\left(a = \arg\min_j \theta_j, \min_j \theta_j < \gamma\right)}{\Pi_{n-1}\left(\min_j \theta_j < \gamma\right)} \leq \frac{\Pi_{n-1}\left(\theta_a < \gamma\right)}{\max_a \Pi_{n-1}\left(\theta_a < \gamma\right)}$$
$$\leq \frac{e^{-n(\bar{\psi}_a(n)\mathrm{kl}(\mu_a,\gamma)-\varepsilon_n)}}{\max_a e^{-n(\bar{\psi}_a(n)\mathrm{kl}(\mu_a,\gamma)+\varepsilon_n)}} = e^{-n\left(\bar{\psi}_a(n)\mathrm{kl}(\mu_a,\gamma)-\min_a \bar{\psi}_a(n)\mathrm{kl}(\mu_a,\gamma)-2\varepsilon_n\right)}.$$

By (5.9) $\min_a \bar{\psi}_a(n)\mathrm{kl}(\mu_a,\gamma) \leq \max_{\boldsymbol{w}} \min_a w_a \mathrm{kl}(\mu_a,\gamma) = w_a^\star \mathrm{kl}(\mu_a,\gamma)$. Also $\bar{\psi}_a(n) \geq w_a^\star + c$, so

$$\psi_a(n) \leq e^{-n((w_a^\star+c)\mathrm{kl}(\mu_a,\gamma)-w_a^\star\mathrm{kl}(\mu_a,\gamma)-2\varepsilon_n)} = e^{-n(c\mathrm{kl}(\mu_a,\gamma)-2\varepsilon_n)}.$$

Now as $\varepsilon_n \to 0$, this establishes eventual exponential decay, hence ensuring that

$$\sum_n \psi_a(n)\mathbb{1}\left(\bar{\psi}_a(n) \geq w_a^\star + c\right) < \infty,$$

as required. The conclusion follows from Lemma 5.14.

□

5.3.2 Stopping Rules

Before resorting to (variants) of the Parallel GLRT test, we will prove that when combined with a simple and intuitive stopping rule based on confidence intervals, the sample complexity of Murphy Sampling is close to $T^\star(\boldsymbol{\mu})\log(1/\delta)$.

The "Box" stopping rule The first stopping rule that comes to mind consists in comparing each arm to the threshold *separately* and stopping when either one arm looks significantly below the threshold or all arms look significantly above. With the Kullback-Leibler upper and lower confidence bounds defined in (5.6) and (5.7), we let $\tau^{\mathrm{Box}} = \tau_< \wedge \tau_>$ where

$$\tau_< = \inf\left\{t \in \mathbb{N} : \min_{a\in[A]} \mathrm{UCB}_a(t) < \gamma\right\}$$
$$\tau_> = \inf\left\{t \in \mathbb{N} : \min_{a\in[A]} \mathrm{LCB}_a(t) > \gamma\right\}$$

This intuitive stopping rule is illustrated in Figure 5.6. When the algorithm stops when $\tau = \tau_<$ the recommendation $\hat{m}$ is $<$, whereas the recommendation is $>$ if $\tau = \tau_>$.

We recall the threshold function $\mathcal{T}(x)$ introduced in (3.7) in Chapter 3, which satisfies $\mathcal{T}(x) \simeq x + c\log\left(x\right)$ for some constant c:

$$\mathcal{T}(x) = 2\tilde{h}\left(\frac{h^{-1}(1+x) + \log(2\zeta(2))}{2}\right) \tag{5.10}$$

where for $u \geq 1$, $h(u) = u - \log u$ and for any $x \geq 0$

$$\tilde{h}(x) = \begin{cases} e^{1/h^{-1}(x)}h^{-1}(x) & \text{if } x \geq h^{-1}(1/\log(3/2)), \\ (3/2)(x - \log\log(3/2)) & \text{otherwise.} \end{cases}$$

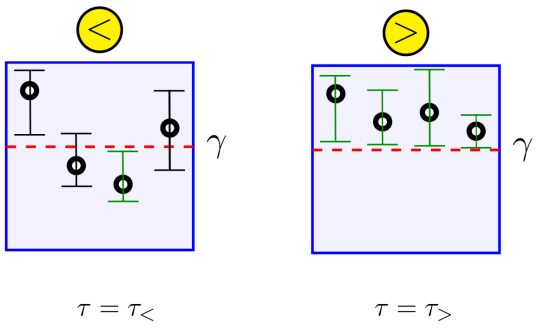


Figure 5.6 – A Box stopping rule

Theorem 3.1 in Chapter 3 permits to prove that for a calibration of the confidence intervals with

$$\begin{aligned} C_<(n,\delta) &= \mathcal{T}(\log(1/A\delta)) + 3\log(1+\log(n)) \\ C_>(n,\delta) &= \mathcal{T}(\log(1/\delta)) + 3\log(1+\log(n)) \end{aligned}$$

the corresponding Box stopping rule is δ-correct. From Lemma 5.15 below we deduce that Murphy Sampling combined with this stopping rule satisfies $\lim_{\delta\to 0} \frac{\tau_\delta}{\log(1/\delta)} \le T^\star(\boldsymbol{\mu})$, almost surely. The proof of this result is very similar to that of Lemma 4.8 in Chapter 4.

Lemma 5.15. *Fix $\boldsymbol{\mu} \in \mathcal{R}$. Fix an anytime sampling strategy (A_t) ensuring $\frac{N_a(t)}{t} \to w_a^\star(\boldsymbol{\mu})$ for all $a \in [A]$. Let τ_δ be a stopping rule such that $\tau_\delta \le \tau_\delta^{Box}$, for a Box stopping rule whose threshold $C_\lessgtr$ in the confidence intervals (5.6) and (5.7) satisfy the following: they are non-decreasing in r and there exists a function f such that,*

$$\forall r \ge r_0,\ C_\lessgtr(\delta, r) \le f(\delta) + \log r, \quad \textit{where} \quad f(\delta) = \log(1/\delta) + o(\log(1/\delta)).$$

Then $\limsup_{\delta\to 0} \frac{\tau_\delta}{\log\frac{1}{\delta}} \le T^\star(\boldsymbol{\mu})$ almost surely.

We remark that in order to prove that Murphy Sampling is asymptotically optimal, one should rather provide an upper bound on the expectation $\mathbb{E}[\tau_\delta]$, which is technically more involved that obtaining the result of Lemma 5.15. A possibility to do that would be to use the tools presented in Chapter 4 for the fixed-confidence analysis of Top-Two Thompson Sampling and prove that the expected number of time steps before the empirical proportions are ε-close to the optimal proportions is bounded.

More sophisticated stopping rules It can be checked that the stopping rule $\tau_<$ is already that of a sequential Generalized Likelihood Ratio Test for rejecting $H_0 : (\boldsymbol{\mu} \in \mathcal{R}_>)$. However, a GLRT for rejecting $H_0 : (\boldsymbol{\mu} \in \mathcal{R}_<)$ is summing evidence across arms whose empirical means are smaller than γ in the sense that it stops when $\sum_{a:\hat\mu_a(t)\le\gamma} N_a(t) d^+(\hat\mu_a(t),\gamma)$ is larger than some threshold. Still based on Theorem 3.1, a δ-correct Parallel GLRT is $\tau^{\mathrm{GLRT}} = \tau_<^{\mathrm{GLRT}} \wedge \tau_>$ where

$$\tau_<^{\mathrm{GLRT}} = \inf\left\{ t \in \mathbb{N}^\star : \sum_{a:\hat\mu_a(t)\le\gamma} [N_a(t)\mathrm{kl}(\hat\mu_a(t),\gamma) - 3\log(1+\log(N_a(t)))]^+ \ge A\mathcal{T}\left(\frac{\log(1/\delta)}{A}\right)\right\}$$

Instead of summing evidence across arms, one can go further and *aggregate* evidence by merging samples from a well-chosen subset of arms. This is justified by a new time-uniform deviation inequality, featuring, for every subset $\mathcal{S} \subseteq [A]$, the quantities

$$N_\mathcal{S}(t) = \sum_{a\in\mathcal{S}} N_a(t) \qquad \text{and} \qquad \hat\mu_\mathcal{S}(t) = \frac{\sum_{a\in\mathcal{S}} N_a(t)\hat\mu_a(t)}{N_\mathcal{S}(t)}\,.$$

The proof of Theorem 5.16 bears strong similarities with that of Theorem 3.1 presented in Chapter 3.

Theorem 5.16. *Let $\mathcal{T}$ be the threshold function defined in (5.10). For every subset $\mathcal{S}$ of arms and $x \geq 0$,*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : N_{\mathcal{S}}(t) d^{+}\left(\hat{\mu}_{\mathcal{S}}(t), \min_{a \in \mathcal{S}} \mu_a\right) \geq 3\log(1 + \log(N_{\mathcal{S}}(t))) + \mathcal{T}(x)\right) \leq e^{-x}, \tag{5.11}$$

$$\mathbb{P}\left(\exists t \in \mathbb{N} : N_{\mathcal{S}}(t) d^{-}\left(\hat{\mu}_{\mathcal{S}}(t), \max_{a \in \mathcal{S}} \mu_a\right) \geq 3\log(1 + \log(N_{\mathcal{S}}(t))) + \mathcal{T}(x)\right) \leq e^{-x}. \tag{5.12}$$

Fix a subset prior $\pi : \wp(\{1, \dots, A\}) \to \mathbb{R}^+$ such that $\sum_{\mathcal{S} \subseteq \{1,\dots,A\}} \pi(\mathcal{S}) = 1$. We define the stopping rule $\tau^{\pi} := \tau_{>} \wedge \tau^{\pi}_{<}$, where

$$\begin{aligned}
\tau_{>} &= \inf\left\{t \in \mathbb{N}^{\star} : \forall a \in \{1, \dots, A\} N_a(t) d^{-}\left(\hat{\mu}_a(t), \gamma\right) \geq 3\log(1 + \log(N_a(t))) + \mathcal{T}\left(\log(1/\delta)\right)\right\}, \\
\tau^{\pi}_{<} &= \inf\left\{t \in \mathbb{N}^{\star} : \exists \mathcal{S} : N_{\mathcal{S}}(t) d^{+}\left(\hat{\mu}_{\mathcal{S}}(t), \gamma\right) \geq 3\log(1 + \log(N_{\mathcal{S}}(t))) + \mathcal{T}\left(\log(1/(\delta\pi(\mathcal{S})))\right)\right\}.
\end{aligned}$$

For the practical computation of $\tau^{\pi}_{<}$, note that the search over subsets can be reduced to nested subsets including arms sorted by increasing empirical mean and smaller than γ. The following result is a consequence of Theorem 5.16.

Lemma 5.17. *Any algorithm using the stopping rule τ^{π} and selecting $\hat{m} = >$ iff $\tau^{\pi} = \tau_{>}$, is δ-correct.*

Practical performance As explained in the beginning of this section, the Box stopping rule based on confidence intervals is sufficient to obtain an asymptotically optimal algorithm. Yet we report results of experiments that reveal that a well chosen aggregate stopping rule may indeed stop earlier. Quantifying the potential benefits of such a stopping rule is left for future work.

We discuss the results of numerical experiments performed on Gaussian bandits with variance 1, using the threshold $\gamma = 0$. Thompson and Murphy sampling are run using a flat (improper) prior on $\mathbb{R}$, which leads to a conjugate Gaussian posterior. The experiments demonstrate the flexibility of our MS sampling rule, which attains optimal performance on instances from both $\mathcal{R}_{<}$ and $\mathcal{R}_{>}$. Moreover, they show the advantage of using a stopping rule aggregating samples from subsets of arms when $\boldsymbol{\mu} \in \mathcal{R}_{<}$. This aggregating stopping rule, that we refer to as τ^{Agg} is an instance of the τ^{π} stopping rule presented above for $\pi(\mathcal{S}) = A^{-1}\binom{A}{|\mathcal{S}|}^{-1}$. We investigate the combined use of three sampling rules, MS, LCB and Thompson Sampling with three stopping rules, τ^{Agg}, τ^{Box} and τ^{GLRT}.

We first study an instance $\boldsymbol{\mu} \in \mathcal{R}_{<}$ with $A = 10$ arms that are linearly spaced between -1 and 1. We run the different algorithms (excluding the TS sampling rule, that essentially coincides with MS on $\mathcal{R}_{<}$) for different values of δ and report the estimated sample complexity in Figure 5.7 (left). For each sampling rule, it appears that $\mathbb{E}[\tau^{\text{Agg}}] \leq \mathbb{E}[\tau^{\text{Box}}] \leq \mathbb{E}[\tau^{\text{GLRT}}]$. Moreover, for each stopping rule, MS is outperforming LCB, with a sample complexity of order $T^{\star}(\boldsymbol{\mu})\log(1/\delta) + C$. Then we study an instance $\boldsymbol{\mu} \in \mathcal{R}_{>}$ with $A = 5$ arms that are linearly spaced between 0.5 and 1, with τ^{Agg} as the sampling rule (which matters little as the algorithm mostly stops because of $\tau_{>}$ on $\mathcal{R}_{>}$). Results are reported in Figure 5.7 (right), in which we see that MS is performing very similarly to LCB (that is also proved optimal on $\mathcal{R}_{>}$), while vanilla TS fails dramatically. On those experiments, the empirical error was always zero, which shows that our theoretical thresholds are still quite conservative.

We report in Figure 5.8 further results illustrating the convergence of the sampling proportions $N_a(\tau)/\tau$ under the two instances of $\mathcal{R}_{<}$ and $\mathcal{R}_{>}$ described above, for the smallest value of δ used in each experiment and under the stopping rule τ^{Agg}. Under $\mathcal{R}_{<}$ we see that MS has indeed spent a larger fraction of the time on the arm with smallest mean (arm 1 in these experiments), even if it does not yet reach the fraction 1 prescribed by the lower bound. Under $\mathcal{R}_{>}$, we see that the empirical fractions of draws of both MS and LCB converge to $\boldsymbol{w}^{\star}(\boldsymbol{\mu})$ whereas the TS sampling rule departs significantly from those optimal weights, by drawing mostly arm $a_{\min} = 1$.

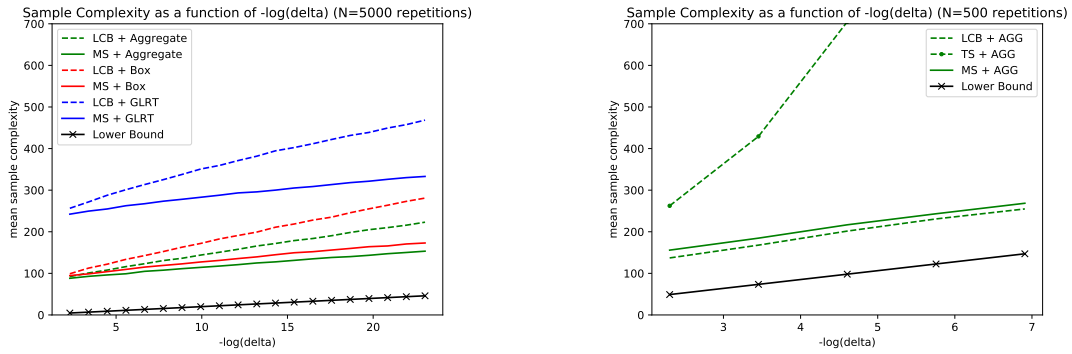


Figure 5.7 – $\mathbb{E}[\tau_\delta]$ as a function of $\log(1/\delta)$ for several algorithms on an instance $\mu \in \mathcal{R}_<$ (left) and $\mu \in \mathcal{R}_>$ (right), estimated using $N = 5000$ (resp. 500) repetitions.

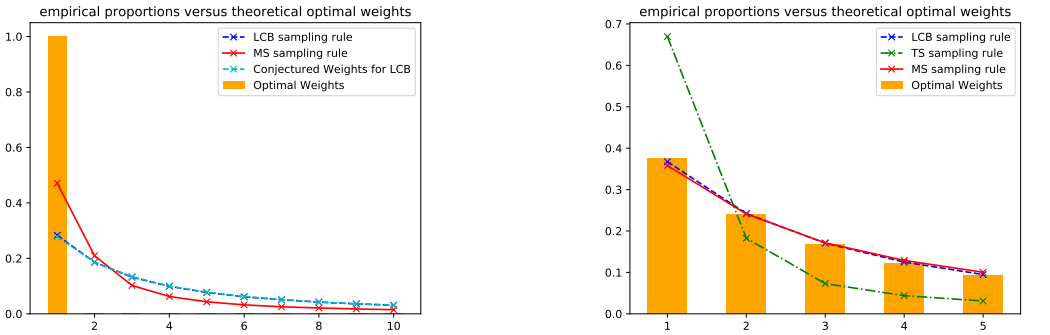


Figure 5.8 – Empirical proportions of samples versus $w^\star(\mu)$ for one instance in $\mathcal{R}_<$ (left) and one instance in $\mathcal{R}_>$ (right), in the same experimental setup as that of Figure 5.7.

To summarize, we presented in this chapter two active identification problems motivated by the design of MCTS algorithms with sample complexity guarantees. For the first full MCTS problem, designing an asymptotically optimal algorithm is hard as there is in general no efficient way to compute the optimal allocation $w^\star(\mu)$. However, we managed to propose an efficient algorithm based on confidence intervals, albeit doomed to be sub-optimal.

For the second "1 + 1/2" MCTS problem, the optimal allocation is easy to compute, and is very different under each of the two possible hypotheses. A variant of Track-and-Stop presented in Chapter 4 for best arm identification could certainly be applied. However, in this particular case we managed to propose a variant of Thompson Sampling that also ensures the convergence of the sampling proportions towards $w^\star(\mu)$, under both hypotheses. Unlike Top Two Thompson Sampling that was presented in Chapter 4 as a possible alternative to Track-and-Stop, Murphy Sampling does not need any problem-dependent tuning (the parameter β in TTTS) to be asymptotically optimal.

In future work, I will continue to investigate variants of Thompson Sampling for different active identification problems, hoping to propose a unified algorithm using possibly some conditioning in the spirit of Murphy Sampling.

Perspective

We presented several bandit algorithms in different contexts: rewards maximization in complex, for example structured bandit models, and active identification problems in the fixed-confidence setting. In both cases, our focus was on proposing *asymptotically optimal* algorithms, i.e. algorithms whose regret or sample complexity is matching a distribution-dependent lower bound, for every possible bandit instance $\boldsymbol{\mu}$, at least in an asymptotic regime in the horizon T or in the risk parameter δ. Beyond this asymptotic notion of optimality, that can be viewed as a sanity-check for the design of algorithms, we also cared about their finite-time performance, often assessed through numerical simulation.

Now we can take a step back and examine the similarities between the different approaches that have been discussed. For both regret minimization in structured bandits and active identification, we have presented lower-bound inspired algorithms that leverage the computation of the oracle allocation under a given configuration of the means. As the computation of this allocation is costly, it is important to come up with algorithms that can avoid these oracle calls, possibly based on Thompson Sampling, an approach also discussed in both parts of this document. We elaborate on these two aspects below.

Beyond lower-bound inspired algorithms Lower bound-inspired algorithms for structured bandits were presented in Chapter 1, while the Track-and-Stop approach for best arm identification (and other active identification problems) was described in Chapter 4. Several approaches to overcome the computational cost of Track-and-Stop for pure exploration have already been proposed, including a nice game view on the lower bound given by Degenne et al. (2019). The generic approach proposed relies on the use of an online learning algorithm for each of the two players, and understanding what is the best combination of these algorithms depending on the problem at hand would be interesting.

In particular, we have been investigating with Pierre Ménard the use of a Saddle-Point Frank-Wolfe algorithm, as least for the best arm identification problem. In that case the optimization problem to solve in $\boldsymbol{w}$ can be written as, letting $F_{a_\star,a}(\boldsymbol{w},\boldsymbol{\mu}) = w_{a_\star} d\left(\mu_{a_\star}, \mu_{a_\star,a}(\boldsymbol{w})\right) + w_a d\left(\mu_a, \mu_{a_\star,a}(\boldsymbol{w})\right)$,

$$\sup_{\boldsymbol{w}\in\Sigma_A} \min_{a\neq a_\star(\boldsymbol{\mu})} F_{a_\star,s}(\boldsymbol{w},\boldsymbol{\mu}) = \sup_{\boldsymbol{w}\in\Sigma_A} \inf_{\boldsymbol{q}\in\Sigma_{A-1}} \sum_{a\neq a_\star(\boldsymbol{\mu})} q_a F_{a_\star,s}(\boldsymbol{w},\boldsymbol{\mu}) \,.$$

The function $G(\boldsymbol{w},\boldsymbol{q}) = \sum_{a\neq a_\star(\boldsymbol{\mu})} q_a F_{a_\star,s}(\boldsymbol{w},\boldsymbol{\mu})$ is concave in $\boldsymbol{w}$, convex in $\boldsymbol{q}$ and the set $\Sigma_A \times \Sigma_{A-1}$ is convex and compact. Hence we could apply saddle-point solvers such as Mirror Prox (Bubeck, 2015) or Saddle-Point Frank-Wolfe (SP-FW). However, the function G does not satisfy the assumptions under which Gidel et al. (2017) manage to analyze this algorithm. Despite this lack of theoretical guarantees, in our numerical experiments, SP-FW seems to converge to the optimal proportions $\boldsymbol{w}^\star(\boldsymbol{\mu})$, and we would like to investigate a noisy version of this algorithm as a possible sampling rule for BAI that may converge faster to the optimal weights.

The game view on lower bounds was recently also used for reward maximization in structured bandits by (Degenne et al., 2020), leading to algorithms with a better performance than OSSB (Combes et al., 2017) at a smaller computational cost. Still, several open research directions are pointed out by the

authors, including the question whether it is possible to exploit the structure in a small horizon regime, and always improve over UCB approaches.

Towards a universal Thompson Sampling algorithm We mentioned several variants of Thompson Sampling in the previous chapters. First, for regret minimization in structured bandits, it is known that vanilla Thompson Sampling cannot always be asymptotically optimal as it fails for linear bandits (Lattimore and Szepesvári, 2017). Still, for unimodal bandits knowing the structure of the (sparse) optimal allocation allows to design a simple variant of TS that is asymptotically optimal, as explained in Chapter 1. It would be interesting to understand for which other structures variants of TS can be designed by using some minimal information about the optimal allocation.

For active identification problems, vanilla Thompson Sampling fails for an even simpler reason: under this algorithm the empirical allocation converges towards a Dirac in the optimal arm, whereas the optimal allocation can be supported on more than one arm. Still, we studied two variants of TS that have near-optimal performance. Murphy Sampling, presented in Chapter 5 for deciding whether one of the means in the bandit model is smaller than a threshold, is particularly interesting as this parameter-free algorithm ensures convergence towards the optimal proportions in all possible bandit models, even if the optimal allocation can have very different forms under different bandit instances. We also proposed in Chapter 4 a new fixed-confidence analysis of Top-Two Thompson Sampling (Russo, 2016) which is β-optimal for best arm identification and requires an oracle tuning of the parameter β to be optimal.

Both algorithms use some kind of conditioning which triggers some posterior re-sampling. In future work, I want to investigate whether it is possible to propose a generic re-sampling version of Thompson Sampling for active identification and possibly also for regret minimization in structured bandits.

Regarding the mathematical tools that have been useful in this study, we highlighted the importance of information-theoretic lower bounds that can shape the design of optimal algorithms, and that of concentration inequalities. In particular, we presented the proof of a concentration inequality that measures the deviations for multiple arms simultaneously (Theorem 3.1). Such an inequality can be useful in the analysis of algorithms for structured bandits, for which constructing individual confidence intervals on each arm can be sub-optimal. But we mostly highlighted in Chapter 3 its use as a means to prove the correctness of a Parallel Generalized Likelihood Ratio test for any active identification problem in an exponential bandit model, possibly involving overlapping regions.

Generalized Likelihood Ratio is a very old statistical tool (Wilks, 1938) to handle composite hypotheses, and a GLR test is often calibrated using some asymptotic properties, such as the Wilks phenomenon. We thus presented examples of sequential, active GLR tests whose calibration (choice of threshold) is not based on asymptotic arguments. In the preprint Garivier and Kaufmann (2019) we advocate the use of a Parallel GLRT for any (sequential, adaptive) statistical testing problem with composite and possibly overlapping hypotheses, for which other deviation inequalities may be found. We are also currently investigating the non-asymptotic calibration of change-point detection tests based on GLRT, for which an asymptotic calibration was given by Lai and Xing (2010). This work requires some adaptation of the techniques to prove Theorem 3.1.

I now highlight below a few other research directions that I want to pursue.

Beyond fixed-confidence best arm identification In part II of this document, we studied active identification in the fixed-confidence setting, in which one should find a correct answer with probability larger than $1-\delta$ using as few samples from the arms as possible. As explained in Section 3.1.2, several other mathematical frameworks exist, including the fixed-budget setting in which given a budget n one should minimize the error probability of a recommendation made after n arm selections.

I am quite intrigued by the possible difference in complexity between the fixed-confidence and fixed-budget setting. If the complexity were similar, one could expect a minimal error probability of order $\exp(-n/T_\star(\boldsymbol{\mu}))$ for a budget n, as $T_\star(\boldsymbol{\mu})\log(1/\delta)$ is the minimal number of samples needed to have an error of at most δ in the fixed-confidence setting. However, it is unclear whether this conjecture is true. First, it seems that lower bounds are much harder to prove in the fixed-budget setting and to the best of my knowledge, the tightest existing lower bound does not feature information-theoretic quantities (Carpentier and Locatelli, 2016). One can note that the assumption made on the algorithm in the fixed-budget setting (its probability of error should go to zero for all instances $\boldsymbol{\mu}$) is much weaker than having the probability of error uniformly bounded by δ, which makes the use of our lower bound methodology more intricate. Second, there are elements suggesting that the complexity of the two problems may be different. It is possible to provide tight upper and lower bounds on the error probability of a strategy using a fixed allocation $\boldsymbol{w}$ and recommending the empirical best arm (using e.g. a similar technique as that of Glynn and Juneja (2004)) which reveals that the allocation that yields the smallest error probability is

$$\operatorname*{argmax}_{\boldsymbol{w}\in\Sigma_A} \min_{a\neq a_\star(\boldsymbol{\mu})} \min_x \left[w_{a_\star} d(x,\mu_{a_\star}) + w_a d(x,\mu_a)\right].$$

The difference with the optimal allocation in the fixed-confidence setting is that the arguments of the KL-divergence are *reversed*. However, it is unclear whether an adaptive algorithm, that is agnostic to $\boldsymbol{\mu}$ and thus to the optimal static allocation, can have a probability of error as small as that of the best static allocation. Indeed if a Tracking procedure were to be applied, the dominant term in the error probability would be the estimation error. Hence, I am very eager to understand what is the smallest error that an adaptive algorithm can achieve in the fixed-budget setting.

Among all the settings studied for best arm identification or more generally active identification, the one that I find the most interesting is the anytime setting of Jun and Nowak (2016). In this framework, there is no risk δ, no budget n but one needs the algorithm to make a recommendation, whenever it is stopped, that is as accurate as possible. In particular, anytime algorithms (that neither depend on n or δ) are needed, while state of the art fixed-budget algorithms require the knowledge of n to tune the size of some elimination phases (Karnin et al., 2013). Among the fixed-confidence algorithms that we proposed, we noted that Track-and-Stop and variants of Thompson Sampling are naturally anytime. In future work, I would like to obtain a non-asymptotic bound on the error probability for variants of TTTS (hopefully adaptive to β) in each round t, to justify the practical use of such algorithms in applications in which there is no natural risk δ or budget n.

Reinforcement learning I have also been interested in more general reinforcement learning problems. While in reward maximization in a bandit model the agent repeatedly faces the same actions, in reinforcement learning it is assumed that the state of the agent evolves after each action, following a Markov Decision Process (MDP). I have been working on reinforcement learning since the beginning of the PhD of Omar Darwiche Domingues, in October 2018, which is part of the collaborative DELTA project on lifelong reinforcement learning, lead by Anders Jonsson at UPF (Barcolona).

Monte-Carlo Tree Search algorithms that are discussed in Chapter 5 were originally proposed both for planning in games and in Markov Decision Processes (i.e. for finding the best action to take in a given state of the MDP based on trajectories sampled using a generative model). However, the simple model we proposed for MCTS and the BAI-MCTS algorithm are specific to deterministic games with random terminal rewards, whereas in MDPs intermediate rewards may be collected and transitions are stochastic. In the recent work (Jonsson et al., 2020) we proposed the MDP-GapE algorithm for planning in Markov Decision Processes. The algorithm is inspired by UGapE-MCTS for games but its analysis is more involved and notably requires new time uniform confidence intervals for transition probabilities, that were derived by Pierre Ménard. The algorithm is suited for planning in MDPs with a finite branching

factor, in which there is a (known) maximal number B of possible next states. We note that state-of-the-art planning algorithms for $B = \infty$ (Grill et al., 2016, 2019) are mostly theoretical, as their actual runtime is prohibitively high. Designing practical planning algorithms for $B = \infty$ is therefore an interesting direction of future work.

In this document, we noted in several places the fundamental difference between optimal sampling strategies designed for minimizing regret and optimal sampling strategies for identifying the best arm, at least through our problem-dependent lens. In reinforcement learning, one can also define two similar objectives: on the one hand learning to interact with the environment in order to maximize the total reward received (a notion of regret can also be defined) and on the other hand learning a good *policy* (a mapping from state to actions). In the paper Jin et al. (2018), the authors explain that if a regret minimizing algorithm is run for enough episodes, picking a policy uniformly at random among the ones used by the algorithm produces a policy that is ε optimal with high probability. However, it is not clear that this conversion (which has a very bad scaling in the error probability δ) is the best thing one can do if the initial goal is to find a good policy, regardless of the received rewards. In future work I want to get a better understanding of the relationship between regret minimization and best policy identification in reinforcement learning.

Alternative exploration methods How to achieve good exploration or a good exploration/exploitation trade-off in a bandit model is a well-studied topic also because it inspires some exploration methods for reinforcement learning. In this document, we encountered two families of exploration methods: the use of confidence intervals and that of posterior sampling. While they can lead to asymptotically optimal algorithms, they are also very much tailored to the specific distributions of the reward. For example kl-UCB requires the right divergence function $d(\cdot, \cdot)$ to build the upper confidence bound whereas TS requires the right prior distribution to match the Lai and Robbins lower bound. A single algorithm that is simultaneously optimal for different classes of distributions would therefore be more powerful.

To come up with such an algorithm, it is necessary to exploit the empirical distribution of the rewards, regardless of any parametric assumption. This has been done with the IMED (Honda and Takemura, 2015) and kl-UCB-Switch (Garivier et al., 2018) algorithms but it applies only to bounded distributions, with known bounds. Under the same assumption, Non Parametric Thompson Sampling (Riou and Honda, 2020) is also asymptotically optimal for any bounded distribution, matching the lower bound of Burnetas and Katehakis (1996). This index policy computes for each arm a weighted sum of the observed rewards and the upper bound of the support, in which the weights are drawn uniformly at random in the simplex. This can be seen as some kind of bootstrap estimator of the mean. Another natural bootstrap idea is to use a sample obtained using the non-parametric bootstrap as an index for each arm. That is, given a history of n rewards from an arm, n rewards are sampled with replacement from this history and their average is computed. Yet this approach is known to suffer from linear regret, as shown by (Kveton et al., 2019b) who also propose a fix. This fix consists in adding a certain amount of fake samples in each history before bootstrapping (see also Kveton et al. (2019a)). The Giro and PHE algorithms are proved to have logarithmic regret for bounded distributions, but their asymptotic optimality, e.g. for Bernoulli rewards, remains unclear.

In an on-going work with Odalric-Ambrym Maillard and our new PhD student Dorian Baudry, we are currently investigating alternative re-sampling approaches based on pairwise comparisons and subsampling, in the spirit of the BESA algorithm of Baransi et al. (2014), with an emphasis on proving the asymptotic optimality of the proposed approach for different distributions, see the recent publication Baudry et al. (2020).

On the applications side I am very interested in understanding why bandit algorithms, whose motivating story is often clinical trials —at least in the introduction of theoretical papers— have seldom been used in this field, and what (variants of) bandit algorithms could be useful. The gold standard in clinical trials seems to be randomized clinical trials, in which each treatment is given to the same number of patients, often decided in advance. Both the statistical power and the number of saved lives have the potential to be improved by using Adaptive Clinical Trials, whose use has been recently promoted (Food and Drugs Administration (FDA), 2018). In particular, Bayesian Adaptive Designs have been studied a lot (Berry et al., 2010).

A particular field in which a Bayesian Adaptive Design is concretely used is that of dose-finding clinical trials in oncology, that is the context of our work Aziz et al. (2018) presented in Chapter 1. The algorithm is called the Continual Reassessment Method (CRM) and can be seen from a bandit eye as a greedy algorithm for toxicity probabilities following a Bayesian logistic regression model. We showed through numerical experiments that variants of Thompson Sampling can also be competitive in this setting, but we could not provide a theoretical analysis beyond the simpler model studied in Chapter 1. In future work, I plan to start new projects on the use of bandit tools for clinical trials, beyond early-stage dose-finding designs.

On a related topic, I also started to work on possible applications of sequential decision making to drug repurposing with the co-supervision of the PhD thesis of Clémence Reda with Andrée Delahaye-Duriez from INSERM (Paris) since September 2019.

Index of Notation

MAB	Multi-Armed Bandit
$[n]$	set of integers $\{1, \dots, n\}$
$\overline{\mathcal{E}}$	the complement of the event $\mathcal{E}$
A	number of arms in a bandit model
ν_a	distribution associated to arm a
μ_a	mean of arm a
t	index of a decision round
T	horizon of the bandit game
$(X_{a,t})_{s\in\mathbb{N}^*}$	i.i.d. rewards stream associated to arm a
$(Y_{a,s})_{s\in\mathbb{N}^*}$	i.i.d. sequence of successive observations from arm a
A_t	arm selected in round t by a bandit algorithm
X_t	observation in round t (called R_t for reward in part I)
$\mathcal{F}_t$	σ-algebra generated by the observation available after t rounds
$a_\star$	arm with largest mean
$\mu_\star$	largest mean in the bandit model
$\mathcal{R}_{\boldsymbol{\nu}}(\mathcal{A}, T)$	regret of a bandit algorithm $\mathcal{A}$ in T rounds on the bandit model $\boldsymbol{\nu}$
$N_a(t)$	number of selections of arm a in the first t rounds
$b(\theta)$	log-partition function of an exponential family with canonical parameter θ
$f_\theta(x)$	density in an exponential family as a function of the natural parameter: $f_\theta(x) = \exp(\theta x - b(\theta))$
$\mathcal{I}$	set of possible means in an exponential family
ν^μ	distribution with mean μ in a one-parameter exponential family
$\mathrm{KL}(\nu, \nu')$	Kullback-Leibler divergence between the distributions ν and ν'
$d(\mu, \mu')$	Kullback-Leibler divergence between ν^μ and $\nu^{\mu'}$ in an exponential family
$\mathrm{kl}(\mu, \mu')$	binary relative entropy, i.e. $d(\mu, \mu')$ for Bernoulli distributions
$\boldsymbol{\nu}$	bandit model $\boldsymbol{\nu} = (\nu_1, \dots, \nu_A)$
$\boldsymbol{\mu}$	vector of means that parameterized the (exponential family) bandit model $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_A)$

$d^+(x,y)$ $d^+(x,y) = d(x,y)\mathbb{1}_{(x \le y)}$

$d^-(x,y)$ $d^-(x,y) = d(x,y)\mathbb{1}_{(x \ge y)}$

$\mu_{a,s}$ empirical mean based on the first s observations from arm a

$\hat{\mu}_a(t)$ empirical mean of arm a after t rounds: $\hat{\mu}_a(t) = \hat{\mu}_{a,N_a(t)}$

$\mathrm{UCB}_a(t)$ upper confidence bound on μ_a

Π_t posterior distribution on $\boldsymbol{\mu}$

$\pi_a(t)$ posterior distribution on μ_a

-- – – – – – – – – – – – – – – – – – – –

$a_\star(\boldsymbol{\mu})$ optimal arm in the bandit model parameterized by $\boldsymbol{\mu}$

$\mathcal{S}$ set of possible means in a structured bandit model

$\mathrm{Alt}_{\mathcal{S}}(\boldsymbol{\mu})$ set of bandit models in $\mathcal{S}$ that do not share an optimal arm with $\boldsymbol{\mu}$

$C_{\mathcal{S}}(\boldsymbol{\mu})$ constant in the optimal regret rate in a structured bandit

$\boldsymbol{c}_{\mathcal{S}}(\boldsymbol{\mu})$ optimal allocation for the sub-optimal arms

$\mathcal{N}_G(a)$ neighborhood of arm a in a graphical unimodal bandit with graph G

$\mathcal{N}_G^+(a)$ extended neighborhood of arm a: $\mathcal{N}_G^+(a) = \mathcal{N}_G(a) \cup \{a\}$

-- – – – – – – – – – – – – – – – – – – –

MP-MAB Multi-Player Multi-Armed Bandit

M number of players in a multi-player bandit model

μ_a^m mean of arm a for player m

X_t^m sensing information for player m in round t

C_t^m collision information for player m in round t: 1 if player m experiences a collision at time t

R_t^m reward of player m in round t: $R_t^m = X_t^m(1 - C_t^m)$

$\mathrm{TopM}(\boldsymbol{\mu})$ set of M arms with largest means

$N_a^m(t)$ number of selections of arm a by player m in the first t rounds

$C_a(t)$ number of collisions on arm a in the first t rounds

$\mathrm{UCB}_a^m(t)$ upper confidence bound on arm a for player m after t rounds

$\widehat{M^m}(t)$ candidate top M arms for player M after t rounds

$\mathrm{TopM}(\boldsymbol{\mu})$ set of M arms with largest mean in an homogeneous MP-MAB

$U(\pi)$ Utility of a matching $\pi : [M] \to [A]$ from players to arms

$\Delta(\pi)$ Gap of a matching π : $\Delta(\pi) = U^\star - U(\pi)$

Δ Smallest positive gap

$\hat{p}_T$ Number of epochs in the M-ETC-Elim algorithm run with a budget T

-- – – – – – – – – – – – – – – – – – – –

I	number of regions in an active identification problem
$\mathcal{R}$	set of possible values of $\boldsymbol{\mu}$
$\mathcal{R}_i$	region i to which $\boldsymbol{\mu}$ could belong
BAI	Best Arm Identification
τ	stopping rule
$\hat{\imath}_\tau$	recommendation rule (denoted by $\hat{a}_\tau$ for best arm identification, $\hat{s}_\tau$ for MCTS...)
δ	risk parameter
$\beta(t,\delta)$	threshold used in the parallel GLRT
$\mathcal{T}(x)$	function used to express the threshold in Theorem 3.1
$\ell(X_1,\dots,X_t;\boldsymbol{\mu})$	likelihood of the observation under the model parameterized by $\boldsymbol{\mu}$
$\phi_\mu(x)$	log moment generating function of the distribution with mean $\mu \in \mathcal{I}$
$S_a(t)$	sum of observations made from arm a in the first t rounds
$\tilde{Z}_a^\pi(t)$	mixture martingale with prior π: $\tilde{Z}_a^\pi(t) = \int \exp(\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t)) d\pi(\eta)$
--	- - - - - - - - - - - - - - - - - - - -
Σ_A	simplex of dimension $A-1$: $\Sigma_A = \{\boldsymbol{w} \in [0,1]^A : \sum_{a=1}^A w_a = 1\}$
$\mathrm{Alt}(\boldsymbol{\mu})$	set of bandit models that a different best arm than that in $\boldsymbol{\mu}$
$T^\star(\boldsymbol{\mu})$	characteristic number of samples for $\boldsymbol{\mu}$
$\boldsymbol{w}^\star(\boldsymbol{\mu})$	vector of optimal sampling proportions under $\boldsymbol{\mu}$
$\hat{\boldsymbol{\mu}}(t)$	vector of empirical means of the arms after t rounds
TaS	the Track-and-Stop algorithm
$\mu_{a,b}(\boldsymbol{w})$	weighted average of μ_a and μ_b: $\mu_{a,b}(\boldsymbol{w}) = \frac{w_a\mu_a + w_b\mu_b}{w_a + w_b}$
β	fraction of samples assigned to the optimal arm under Top-Two Thompson Sampling (TTTS)
--	- - - - - - - - - - - - - - - - - - - -
$\mathcal{G}$	a maxmin game tree
$\mathcal{L}$	the set of leaves of $\mathcal{G}$
$V_{\boldsymbol{\mu}}(s)$	value of a node $s \in \mathcal{G}$ when the leaves values are given by $\boldsymbol{\mu}$
$\mathcal{C}(s_0)$	depth-one nodes in $\mathcal{G}$ (children of the root s_0)
$s^\star$	depth-one node with largest value
$\mathcal{I}_s(t)$	confidence interval on the value of node s after t rounds $\mathcal{I}_s(t) = [\mathrm{LCB}_s(t), \mathrm{UCB}_s(t)]$
$\ell_s(t)$	representative leaf of node s after t rounds
$\mu_{\min}$	minimum of the arms means: $\mu_{\min} = \min_a \ \mu_a$
γ	value of the threshold
$\mathcal{R}_<$	set of means whose minimum is smaller than γ
$\mathcal{R}_>$	set of means whose minimum is larger than γ

Bibliography

Abbasi-Yadkori, Y., D.Pál, and C.Szepesvári (2011). Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*.

Agrawal, R. (1995). Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078.

Agrawal, R., Teneketzis, D., and Anantharam, V. (1989). Asymptotically Efficient Adaptive Allocation Schemes for Controlled i.i.d. Processes: Finite Parameter Space. *IEEE Transactions on Automatic Control*, 34(3):258–267.

Agrawal, S. and Goyal, N. (2012). Analysis of Thompson Sampling for the multi-armed bandit problem. In *Proceedings of the 25th Conference On Learning Theory*.

Agrawal, S. and Goyal, N. (2013a). Further Optimal Regret Bounds for Thompson Sampling. In *Proceedings of the 16th Conference on Artificial Intelligence and Statistics*.

Agrawal, S. and Goyal, N. (2013b). Thompson Sampling for Contextual Bandits with Linear Payoffs. In *International Conference on Machine Learning (ICML)*.

Anandkumar, A., Michael, N., and Tang, A. K. (2010). Opportunistic Spectrum Access with multiple users: Learning under competition. In *IEEE INFOCOM*.

Anantharam, V., Varaya, P., and Walrand, J. (1987). Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: I.i.d. rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976.

Antos, A., Grover, V., and Szepesvári, C. (2008). Active learning in multi-armed bandits. In *Algorithmic Learning Theory*.

Audibert, J., Bubeck, S., and Lugosi, G. (2014). Regret in online combinatorial optimization. *Math. Oper. Res.*, 39(1):31–45.

Audibert, J.-Y., Bubeck, S., and Munos, R. (2010). Best Arm Identification in Multi-armed Bandits. In *Proceedings of the 23rd Conference on Learning Theory*.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. (2002b). The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 32(1):48–77.

Avner, O. and Mannor, S. (2015). Learning to Coordinate Without Communication in Multi-User Multi-Armed Bandit Problems. *arXiv preprint arXiv:1504.08167*.

Aziz, M., Kaufmann, E., and Riviere, M. (2018). On multi-armed bandit designs for dose-finding clinical trials. *arXiv:1903.07082*.

Balsmubramani, A. (2015). Sharp finite-time iterated-logarithm martingale concentration. *arXiv:1405.2639*.

Baransi, A., Maillard, O., and Mannor, S. (2014). Sub-sampling for multi-armed bandits. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML / PKDD*.

Baudry, D., Kaufmann, E., and Maillard, O.-A. (2020). Sub-sampling for Efficient Non-Parametric Bandit Exploration. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bechhofer, R. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics*, 25:16–39.

Bechhofer, R., Kiefer, J., and Sobel, M. (1968). *Sequential identification and ranking procedures*. The University of Chicago Press.

Berry, S., Carlin, B., Lee, J., and Muller, P. (2010). *Bayesian adaptive methods for clinical trials*. CRC Press.

Berthet, Q. and Perchet, V. (2017). Fast rates for bandit optimization with upper-confidence frank-wolfe. In *Advances in Neural Information Processing Systems (NIPS)*.

Besson, L. and Kaufmann, E. (2018). Multi-player bandits revisited. In *International Conference on Algorithmic Learning Theory (ALT)*.

Bistritz, I. and Leshem, A. (2018). Distributed multi-player bandits - a game of thrones approach. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bistritz, I. and Leshem, A. (2019). Game of thrones: Fully distributed learning for multi-player bandits. *arXiv.org:1810.11162v3*.

Bonnefoi, R., Besson, L., Moy, C., Kaufmann, E., and Palicot, J. (2017). Multi-Armed Bandit Learning in IoT Networks: Learning helps even in non-stationary settings. In *12th EAI Conference on Cognitive Radio Oriented Wireless Network and Communication*, CROWNCOM Proceedings.

Borsoniu, L., Munos, R., and Páll, E. (2014). An analysis of optimistic, best-first search for minimax sequential decision making. In *ADPRL14*.

Boursier, E., Kaufmann, E., Mehrabian, A., and Perchet, V. (2020). A practical algorithm for multiplayer bandits when arm means vary among players. In *The 23rd International Conference on Artificial Intelligence and Statistics, (AISTATS)*.

Boursier, E. and Perchet, V. (2019). SIC-MMAB: synchronisation involves communication in multiplayer multi-armed bandits. In *Advances in Neural Information Processing Systems (NeuRIPS)*.

Browne, C., Powley, E., Whitehouse, D., Lucas, S., Cowling, P., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., and Colton, S. (2012). A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games,*, 4(1):1–49.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717.

Bubeck, S., Munos, R., and Stoltz, G. (2011). Pure Exploration in Finitely Armed and Continuous Armed Bandits. *Theoretical Computer Science 412, 1832-1852*, 412:1832–1852.

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

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

Carpentier, A. and Locatelli, A. (2016). Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Proceedings of the 29th Conference on Learning Theory (COLT)*.

Centenaro, M., Vangelista, L., Zanella, A., and Zorzi, M. (2016). Long-range communications in unlicensed bands: the rising stars in the IoT and smart city scenarios. *IEEE Wireless Communications*, 23(5):60–67.

Chapelle, O. and Li, L. (2011). An empirical evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems*.

Chen, L., Gupta, A., Li, J., Qiao, M., and Wang, R. (2017). Nearly optimal sampling algorithms for combinatorial pure exploration. In *Proceedings of the 30th Conference on Learning Theory (COLT)*.

Chen, W., Wang, Y., and Yuan, Y. (2013). Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*.

Chernoff, H. (1959). Sequential design of Experiments. *The Annals of Mathematical Statistics*, 30(3):755–770.

Combes, R., Magureanu, S., and Proutière, A. (2017). Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Combes, R. and Proutière, A. (2014b). Unimodal Bandits without Smoothness. Technical report.

Combes, R., Talebi, S., Proutière, A., and Lelarge, M. (2015). Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Coulom, R. (2006). Efficient selectivity and backup operators in monte-carlo tree search. In *Computers and Games, 5th International Conference, CG 2006, Turin, Italy, May 29-31, 2006. Revised Papers*, pages 72–83.

Cowan, W., Honda, J., and Katehakis, M. N. (2017). Normal bandits of unknown means and variances. *Journal of Machine Learning Research*, 18:154:1–154:28.

De La Pena, V., Klass, M., and Lai, T. (2004). Self-Normalized Processes: Exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 32(3A):1902–1933.

de Rooij, S., van Erven, T., Grünwald, P. D., and Koolen, W. M. (2014). Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15(1):1281–1316.

Degenne, R. and Koolen, W. M. (2019). Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Degenne, R., Koolen, W. M., and Ménard, P. (2019). Non-asymptotic pure exploration by solving games. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Degenne, R., Shao, H., and Koolen, W. M. (2020). Structure adaptive algorithms for stochastic bandits. In *International Conference on Machine Learning (ICML)*.

Dwork, C., Mcsherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *In Proceedings of the 3rd Theory of Cryptography Conference*, pages 265–284. Springer.

Even-Dar, E., Mannor, S., and Mansour, Y. (2006). Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *Journal of Machine Learning Research*, 7:1079–1105.

Food and Drugs Administration (FDA) (2018). Adaptive design clinical trials for drugs and biologics.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. (2012). Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence. In *Advances in Neural Information Processing Systems*.

Gai, Y., Krishnamachari, B., and Jain, R. (2012). Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Trans. Netw.*, 20(5):1466–1478.

Gajane, P., Urvoy, T., and Kaufmann, E. (2018). Corrupt bandits for preserving local privacy. In *International Conference on Algorithmic Learning Theory (ALT)*.

Garivier, A. and Cappé, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Conference on Learning Theory*.

Garivier, A., Hadiji, H., Ménard, P., and Stoltz, G. (2018). Kl-ucb-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints. *arXiv:1805.05071*.

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference On Learning Theory*.

Garivier, A. and Kaufmann, E. (2019). Non-asymptotic sequential tests for overlapping hypotheses and application to near optimal arm identification in bandit models. *arXiv:1905.03495*.

Garivier, A., Kaufmann, E., and Koolen, W. (2016a). Maximin action identification: A new bandit framework for games. In *Proceedings of the 29th Conference On Learning Theory*.

Garivier, A., Kaufmann, E., and Lattimore, T. (2016b). On explore-then-commit strategies. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Garivier, A., Ménard, P., and Rossi, L. (2019a). Thresholding bandit for dose-ranging: The impact of monotonicity. In *International Conference on Machine Learning, Artificial Intelligence and Applications*.

Garivier, A., Ménard, P., and Stoltz, G. (2019b). Explore first, exploit next: The true shape of regret in bandit problems. *Math. Oper. Res.*, 44(2):377–399.

Gidel, G., Jebara, T., and Lacoste-Julien, S. (2017). Frank-wolfe algorithms for saddle point problems. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Glynn, P. and Juneja, S. (2004). A large deviations perspective on ordinal optimization. In *Proceedings of the 2004 Winter Simulation Conference (IEEE)*.

Graves, T. and Lai, T. (1997). Asymptotically Efficient adaptive choice of control laws in controlled markov chains. *SIAM Journal on Control and Optimization*, 35(3):715–743.

Grill, J.-B., Domingues, O. D., Ménard, P., Munos, R., and Valko, M. (2019). Planning in entropy-regularized markov decision processes and games. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Grill, J.-B., Valko, M., and Munos, R. (2016). Blazing the trails before beating the path: Sample-efficient monte-carlo planning. In *Neural Information Processing Systems (NeurIPS)*.

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

Huang, R., Ajallooeian, M. M., Szepesvári, C., and Müller, M. (2017). Structured best arm identification with fixed confidence. In *International Conference on Algorithmic Learning Theory (ALT)*.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014). lil'UCB: an Optimal Exploration Algorithm for Multi-Armed Bandits. In *Proceedings of the 27th Conference on Learning Theory*.

Jedor, M., Perchet, V., and Louëdec, J. (2019). Categorized bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jonsson, A., Kaufmann, E., Ménard, P., Domingues, O. D., Leurent, E., and Valko, M. (2020). Planning in markov decision processes with gap-dependent sample complexity. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jouini, W., Ernst, D., Moy, C., and Palicot, J. (2009). Multi-armed bandit based policies for cognitive radio's decision making issues. In *International Conference Signals, Circuits and Systems (IEEE)*.

Jun, K.-W. and Nowak, R. (2016). Anytime exploration for multi-armed bandits using confidence information. In *International Conference on Machine Learning (ICML)*.

Juneja, S. and Krishnasamy, S. (2019). Sample complexity of partition identification using multi-armed bandits. In *Conference on Learning Theory (COLT)*.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012). PAC subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning (ICML)*.

Karnin, Z., Koren, T., and Somekh, O. (2013). Almost optimal Exploration in multi-armed bandits. In *International Conference on Machine Learning (ICML)*.

Katariya, S., Kveton, B., Szepesvári, C., Vernade, C., and Wen, Z. (2017a). Bernoulli rank-1 bandits for click feedback. In *International Joint Conferences on Artificial Intelligence (IJCAI)*.

Katariya, S., Kveton, B., Szepesvári, C., Vernade, C., and Wen, Z. (2017b). Stochastic rank-1 bandits. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*.

Kaufmann, E., Bonald, T., and Lelarge, M. (2017). A Spectral Algorithm with Additive Clustering for the Recovery of Overlapping Communities in Networks. *Journal of Theoretical Computer Science*.

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the Complexity of Best Arm Identification in Multi-Armed Bandit Models. *Journal of Machine Learning Research*, 17(1):1–42.

Kaufmann, E. and Garivier, A. (2017). Learning the distribution with largest mean: two bandit frameworks. *ESAIM: Proceedings and Surveys*, 60:114–131.

Kaufmann, E. and Kalyanakrishnan, S. (2013). Information complexity in bandit subset selection. In *Proceeding of the 26th Conference On Learning Theory*.

Kaufmann, E. and Koolen, W. (2018). Mixture martingales revisited with applications to sequential tests and confidence intervals. *arXiv:1811.11419*.

Kaufmann, E., Koolen, W., and Garivier, A. (2018). Sequential test for the lowest mean: From Thompson to Murphy Sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kaufmann, E. and Koolen, W. M. (2017). Monte-Carlo tree search by best arm identification. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson Sampling : an Asymptotically Optimal Finite-Time Analysis. In *Proceedings of the 23rd conference on Algorithmic Learning Theory*.

Kocsis, L. and Szepesvári, C. (2006). Bandit based monte-carlo planning. In *Proceedings of the 17th European Conference on Machine Learning*, ECML'06, pages 282–293, Berlin, Heidelberg. Springer-Verlag.

Komiyama, J., Honda, J., and Nakagawa, H. (2015). Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*.

Komiyama, J., Honda, J., and Nakagawa, H. (2016). Copeland dueling bandit problem: Regret lower bound, optimal algorithm, and computationally efficient algorithm. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*.

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

Kveton, B., Szepesvári, C., Ghavamzadeh, M., and Boutilier, C. (2019a). Perturbed-history exploration in stochastic multi-armed bandits. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*.

Kveton, B., Szepesvári, C., Vaswani, S., Wen, Z., Lattimore, T., and Ghavamzadeh, M. (2019b). Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvári, C. (2015). Tight regret bounds for stochastic combinatorial semi-bandits. In *AISTATS*.

Kwon, J., Perchet, V., and Vernade, C. (2017). Sparse stochastic bandits. In *Proceedings of the 30th Conference on Learning Theory (COLT)*.

Lai, T. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *Annals of Statistics*, 15(3):1091–1114.

Lai, T. (1988). Boundary Crossing problems for samples means. *Annals of Probability*, 16(1):375–396.

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

Lai, T. L. and Xing, H. (2010). Sequential change-point detection when the pre-and post-change parameters are unknown. *Sequential analysis*, 29(2):162–175.

Lattimore, T. and Munos, R. (2014). Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems (NIPS)*.

Lattimore, T. and Szepesvári, C. (2017). The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *AISTATS*.

Lattimore, T. and Szepesvari, C. (2019). *Bandit Algorithms*. Cambridge University Press.

Lattimore, T. and Szepesvári, C. (2019). Cleaning up the neighborhood: A full classification for adversarial partial monitoring. In *Algorithmic Learning Theory (ALT)*.

Liu, K. and Zhao, Q. (2010). Distributed learning in Multi-Armed Bandit with multiple players. *IEEE Transaction on Signal Processing*, 58(11):5667–5681.

Locatelli, A., Gutzeit, M., and Carpentier, A. (2016). An optimal algorithm for the thresholding bandit problem. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1690–1698.

Luedtke, A., Kaufmann, E., and Chambaz, A. (2019). Asymptotically optimal algorithms for budgeted multiple play bandits. *Machine Learning*, 108(11):1919–1949.

Lugosi, G. and Mehrabian, A. (2018). Multiplayer bandits without observing collision information. *arXiv:1808.08416*.

Magureanu, S., Combes, R., and Proutière, A. (2014). Lipschitz Bandits: Regret lower bounds and optimal algorithms. In *Proceedings on the 27th Conference On Learning Theory*.

Mannor, S. and Tsitsiklis, J. (2004). The Sample Complexity of Exploration in the Multi-Armed Bandit Problem. *Journal of Machine Learning Research*, pages 623–648.

Ménard, P. (2019). Gradient ascent for active exploration in bandit problems. *arXiv 1905.08165*.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *J. Soc. Indust. Appl. Math.*, 5:32–38.

O'Quigley, J., Pepe, M., and Fisher, L. (1990). Continual reassessment method: A practical design for phase i clinical trials in cancer. *Biometrics*, (46):33–48.

Osband, I., Van Roy, B., and Russo, D. (2013). (More) Efficient Reinforcement Learning Via Posterior Sampling. In *Advances in Neural Information Processing Systems*.

Paladino, S., Trovò, F., Restelli, M., and Gatti, N. (2017). Unimodal thompson sampling for graph-structured arms. In *AAAI*.

Proutière, A. and Wang, P. (2019). An optimal algorithm in multiplayer multi-armed bandits. *arXiv:1909.13079*.

Qin, C., Klabjan, D., and Russo, D. (2017). Improving the expected improvement algorithm. In *Advances in Neural Information Processing Systems 30 (NIPS)*.

Riou, C. and Honda, J. (2020). Bandit algorithms based on thompson sampling for bounded reward distributions. In *Algorithmic Learning Theory (ALT)*.

Riquelme, C., Tucker, G., and Snoek, J. (2018). Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. In *6th International Conference on Learning Representations (ICLR)*.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

Robbins, H. (1970). Statistical Methods Related to the law of the iterated logarithm. *Annals of Mathematical Statistics*, 41(5):1397–1409.

Robbins, H. and Siegmund, D. (1974). The expected sample size of some tests of power one. *The Annals of Statistics*, 2(3):415–436.

Rosenski, J., Shamir, O., and Szlak, L. (2016). Multi-Player Bandits – A Musical Chairs Approach. In *International Conference on Machine Learning*, pages 155–163.

Russo, D. (2016). Simple Bayesian algorithms for best arm identification. In *Proceedings of the 29th Conference on Learning Theory (COLT)*.

Russo, D., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. (2018). A tutorial on thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96.

Réda, C., Kaufmann, E., and Delahaye-Duriez, A. (2020). Machine learning applications in drug development. *Computational and Structural Biotechnology Journal*, 18:241–252.

Shafer, G., Shen, A., Vereshchagin, N., and Vovk, V. (2011). Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101.

Shang, X., de Heide, R., Kaufmann, E., Ménard, P., and Valko, M. (2020). Fixed-confidence guarantees for bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T. P., Simonyan, K., and Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362.

Slivkins, A. (2019). Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning*, 12(1-2):1–286.

Teraoka, K., Hatano, K., and Takimoto, E. (2014). Efficient sampling method for monte carlo tree search problem. *IEICE Transactions on Infomation and Systems*, pages 392–398.

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

Tibrewal, H., Patchala, S., Hanawal, M. K., and Darak, S. J. (2019). Multiplayer multi-armed bandits for optimal assignment in heterogeneous networks. *arXiv.org:1901.03868v4*.

Tossou, A. C. Y. and Dimitrakakis, C. (2016). Algorithms for differentially private multi-armed bandits. In *13th International Conference on Artificial Intelligence (AAAI 2016)*.

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

Ville, J. (1939). *Étude critique de la notion de collectif*.

Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *Annals of Mathematical Statistics*, 16(2):117–186.

Wang, Y., Wu, X., and Hu, D. (2016). Using randomized response for differential privacy preserving data collection. In *Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference*.

Warner, S. L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, 60(309):63+.

Wilks, S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.