# Non-Parametric Algorithms for Multi-Armed Bandits

Dorian Baudry

Université de Lille, faculté des Sciences et Technologies
Ecole Doctorale des Mathématiques-Sciences du numérique et de leurs interactions

# Thèse de Doctorat

Spécialité **Informatique**

présentée par
**Dorian Baudry**

# Non-Parametric Algorithms for Multi-Armed Bandits

## Algorithmes Non-Paramétriques de Bandits Multi-Bras

sous la direction d' **Emilie Kaufmann**
et d'**Odalric-Ambrym Maillard**.

Soutenue publiquement à **Villeneuve d'Ascq**, le **05/12/2022** devant le jury composé de

| | | |
|---|---|---|
| M. Stephan **Clémençon** | Professeur, Télécom Paris | Rapporteur |
| M. Shie **Mannor** | Professeur, Technion | Rapporteur |
| M$^{me}$ Audrey **Durand** | Professeur assitant, Université Laval | Examinatrice |
| M. Vianney **Perchet** | Professeur, Crest/Ensae Paris | Président du Jury |
| M. Marc **Abeille** | Chercheur, Criteo | Invité |
| M$^{me}$ Emilie **Kaufmann** | Chargée de recherche, CNRS | Directrice de thèse |
| M. Odalric-Ambrym **Maillard** | Chargé de recherche, Inria | Directeur de thèse |

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

# Acknowledgements

Ce manuscrit vient clôturer trois années de doctorat inoubliables. Je remercie d'abord chaleureusement mes directeurs de thèse, Emilie et Odalric, pour cette expérience. Emilie, je te remercie de m'avoir fait confiance au tout début de cette aventure. J'ai appris énormément à tes côtés, et j'ai autant apprécié nos longs échanges sur tableau blanc que les moments de convivialité que nous avons été amenés à partager. Ta gentillesse, ta rigueur mathématique et ton sens de l'organisation sont un véritable modèle pour le jeune chercheur que je suis. Odalric, je te remercie également pour tout ce que tu m'as apporté, et pour m'avoir montré la voie des mathématiques artistiques. Tes connaissances encyclopédiques et ta créativité continueront également de m'inspirer pour longtemps. Je réalise la chance immense que j'ai eu d'être dirigé par deux brillants chercheurs, avec des visions très complémentaires de la recherche.

Au fil de mes travaux, j'ai également eu également la chance de pouvoir travailler avec d'excellents collaborateurs. Je remercie en premier lieu Yoan, que j'ai embarqué dans l'aventure SDA, et qui en plus d'être un ami cher s'avéra être un collègue brillant. Merci également à Romain pour m'avoir fait découvrir DSSAT lors d'un footing, et d'avoir conséquemment suscité un bon nombre des questions qui ont alimenté cette thèse. Merci enfin à tous mes collaborateurs pour votre talent et votre professionalisme: Marquinhos, Olivier, Patrick, Rémy et Rianne.

I also would like to thank a lot the members of Shimodaira lab in Kyoto University, for hosting me for these fantastic three months. Thanks a lot to Junya Honda for the warm welcome, and a very nice collaboration. I also want to address special thanks to Charles Riou and Taira Tsuchiya, who made me discover many aspects of the Japanese culture and with whom I shared great experiences and memories.

Je tiens évidemment à remercier l'ensemble de l'équipe ScooL (SequeL au début de ma thèse), pour m'avoir accueilli pendant ces trois années. Loin de l'image des chercheurs solitaires, j'ai découvert à ScooL une équipe dynamique et très soudée, ce qui semble avoir toujours fait l'identité de cette équipe. Je pense d'abord aux membres de la période pré-covid: merci de m'avoir intégré à l'équipe et d'avoir guidé mes premiers pas d'apprenti doctorant. Je n'oublierai jamais les discussions enflammées à la pause-café, le club ciné, les verres à la Capsule ou au Beerchoppe, ou encore les intenses parties d'Among Us. Je remercie particulièrement Mathieu d'être resté à Lille pendant le covid: notre amitié, notre cinéphilie, et le niveau de

nos jam en sont sortis grandis. Cette période rendit également possible une expérience sociale inédite: la délocalisation de l'équipe ScooL à Varangeville sur mer en Normandie, lorsque des doctorants exténués par les confinements à répétition décidèrent d'aller vivre en communauté aux alentours de Dieppe. Je remercie chaleureusement Edouard pour nous avoir accueilli, et Ariane, Mathieu, Moulmoul, Nathan pour ces souvenirs inoubliables. L'année 2022 permit un retour à la vie normale, et apporta son lot de nouvelles personalités/activités au labo. Ainsi, la pétanque devint l'activité phare du groupe, avec notamment Marquinhos en star et Achraf en chaotique dégommeur, et une pratique tellement assidue que même l'hiver ne parvint pas à l'interrompre. Merci enfin à Hector pour nos discussions très sérieuses autour du foot ou du Muay Thaï.

Finalement, je tiens à remercier ma famille et mes amis, sans qui je n'aurais jamais pu accomplir tout cela. Merci à ma famille pour m'avoir inculqué le goût du savoir et du dépassement de soi. Je suis très heureux d'avoir pu vous rendre fier par mon parcours. Merci à ma soeur, Mélanie, d'avoir toujours été là pour moi. Merci également à mes amis, pour avoir fait partie de ma vie tout ce temps: William, chez qui une partie de ce rapport fut écrit entre deux virées à la plage; Filito et Filita, je remercie le destin de nous avoir réuni à Lille pendant ces 3 ans; Maxime, qui a contribué à m'inculquer l'amour du RC Lens; Abeba et Renan, les meilleurs colloc-zinhos de l'univers; et Estelle pour son brin de folie. Merci également aux coachs et adhérents de Team Naja de m'avoir rendu accroc à ce délicat sport qu'est le Muay Thaï; et à Seko Fofana et Florian Sotoca pour m'avoir fait vivre des émotions incomparables à Bollaert. Merci enfin à Noémie pour l'amour qu'elle m'a prodigué au quotidien durant toutes ces années. Vivre avec un chercheur n'est pas facile, et tu t'en es accommodé avec brio.

# Résumé

Un bandit est un problème d'apprentissage dans lequel un agent choisit séquentiellement de tester une action parmi un ensemble de candidats fixé, collecte une récompense, et met en place une stratégie dans le but de maximiser son gain cumulé. Motivés par une étude de cas dans le domaine de l'agriculture, nous abordons dans cette thèse plusieurs problématiques pertinentes pour les applications réelles des bandits.

La première question que nous considérons concerne les hypothèses faites sur les distribution des récompenses. Alors qu'en théorie il est généralement commode de considérer des hypothèses paramétriques simples (par exemples, des distributions gaussiennes), le practicien peut avoir des difficultés à trouver un modèle adapté à son problème. Pour cette raison, nous étudions deux familles d'algorithmes *non-paramétriques*, dans la mesure où ils ne nécessitent pas d'hypothèses paramétriques fortes sur les distributions pour leur implémentation. Nous montrons que ces deux approches peuvent obtenir de bonnes garanties théoriques pour le problème de bandits usuel, tout en utilisant moins d'hypothèses que les méthodes précédemment proposées.

Nous proposons ensuite différentes extensions de ces algorithmes afin de faciliter leur mise en pratique. La deuxième question principalement étudiée dans nos travaux concerne la prise en compte de critères de performance alternatifs à l'espérance des récompenses cumulées, qui pourraient potentiellement mieux refléter les préférences réelles du practicien. Nous proposons notamment des algorithmes *sensibles au risque*, pour des problèmes dans lesquels l'objectif est d'identifier un bras peu risqué selon une mesure: la *Conditional-Value-at-Risk*. Nous proposons également des algorithmes efficaces pour un problème analogue au cas limite du précédent, appelé *Bandits Extrêmes*. Enfin, nous adaptons nos méthodes pour traiter des variantes usuelles du problème de bandit, avec notamment le cas de récompenses *non-stationnaires* et un exemple où les données sont collectées dans des *groupes d'observations* et non dans un cadre purement séquentiel.

# Abstract

A Multi-Armed Bandits (MAB) is a learning problem where an agent sequentially chooses an action among a given set of candidates, collects a reward, and implements a strategy in order to maximize her sum of reward. Motivated by a case study in agriculture, we tackle in this thesis several problems that are relevant towards real-world applications of MAB.

The first central question that we considered in this thesis is about the assumptions made on the distributions of rewards. While in theory it is usually convenient to consider simple parametric assumptions (e.g gaussian distributions), the practitioner may have some difficulty to find the right model fitting their problem. For this reason, we analyze two families of *non-parametric* algorithms, in the sense that they do not require strong parametric assumptions on the distributions for their implementation. We show that these two approaches can achieve strong theoretical guarantees in the standard bandit setting, improving what should be known in advance by the learner compared with previous algorithms.

Then, we analyze some extensions of these algorithms that make them more suitable for some real-world applications. A second focus of our work is to consider alternative performance metrics, that may be more suitable than the expected sum of rewards for the practitioner. We propose a *risk-aware* algorithm for a bandit problem where the learner wants to find a safe arm according to a risk metric: the *Conditional-Value-at-Risk*. We also propose efficient algorithms for a problem analogous to the limit case of this setting, known as *Extreme Bandits*. Finally, we also adapt some of our approaches for standard variant of MAB, including one with *non-stationary* rewards and one with feedback grouped into *batches* of observations.

# Foreword

## From Machine Learning to Bandits

The works presented in this thesis are part of a vast branch of computer science called *Machine Learning*, that includes any algorithm that is able to improve on a task by analyzing and drawing inferences from patterns in data. We refer to (Hastie et al., 2001) for a complete introduction of this field. In general the algorithm is presented some data, that have been collected beforehand, and then tries to learn to solve a task by training on this database. However, in some applications the data are collected *during* the training, and the algorithm needs to continuously learn from the new data feeding the database. It may even be directly responsible for the data collection process. In *Reinforcement Learning* (RL), the algorithm (called agent) *interacts with their environment* and learns to collect and *maximize rewards*. A complete introduction to RL can be found in (Sutton and Barto, 2018), some examples of applications include interactive speaker recognition (Seurin et al., 2020), or self-driving cars (Leurent, 2020).

In RL, the agent navigates in its environment, performs actions and observe a feedback (reward) associated with this action. The objective is to find the best action to do in every situation provided by the environment. When the agent repeatedly faces the same situation the problem becomes simpler, as it is reduced to the evaluation of each action: this is *Multi-Armed Bandits* (MAB). The name comes from slot machines (one-armed bandits): the agent *pulls* an *arm*, and observe a reward. Hence, each action is called an arm. The usual formulation of this problem can be traced back as far as (Thompson, 1933), and a very complete introduction to Bandits can be found in (Lattimore and Szepesvári, 2020). While this setting is relatively simple, it is powerful to model a variety of problems. Hence, Multi-Armed Bandits are still a very active research field. All the works presented in this thesis fit into this theoretical framework, and in the next section we detail a case-study that motivated our research.


**Figure 1** – Facing a complicated choice.

### Motivating application : a recommendation algorithm for agriculture

While the works presented in this thesis are mostly theoretical, many questions that we considered during these three years of research came after discussing with my supervisors and my fellow PhD student Romain Gautron about the applications of bandits for a recommendation problem in agriculture. Detailing this problem is a natural introduction for this manuscript for several reasons: it is in our opinion an interesting illustrative example of the potential applications of bandits in the real world, it is relatively easy to understand for people who are not familiar with bandits, and it allows us to introduce the main research directions that we considered in this thesis. As this manuscript is written from a mathematician's perspective we only introduce the aspects of the problem that raise the theoretical questions considered in this thesis. For an agronomic point of view on this problem we refer to (Gautron et al., 2022a).

**Experimental set-up** Farmers have been reported to primarily seek advice that reduces uncertainty in highly uncertain decision making environment (McCown, 2002; Hochman and Carberry, 2011; Evans et al., 2017). Let us consider a group of farmers who would like to collectively learn to improve their *crop management practices* for a rain-fed crop. In the context of increasing global food demand, the experiment is aimed toward small farmers under challenging weather and soil conditions, such as maize farmers in Sub-Saharian Africa. The global objective of the experiment is to help farmers find good crop-management practices, while not putting their own food security at risk. We assume that some experts can provide a learning algorithm that would meet these objectives, and that during several consecutive seasons some volunteers in the group are willing to follow the suggestions of this learning system. In the following, we try to define some of the most important aspects that need to be considered to design such algorithm, and leading to the questions that we tried to tackle during this thesis.

We can first elaborate on what defines a crop management practice. In this experiment the crop species and type of soils are fixed for the group of farmers, which still lets several factors to optimize: the planting date, that impacts the weather during crop growth, or the fertilization policy (quantity, planning, ...). Furthermore, all these elements can be defined as a set of rules to follow, allowing some adaptation to external events (e.g weather conditions). A *crop management policy* is then simply defined as the combination of all these choices (or rules). All the parameters have a direct influence on the efficiency of the policy, with very intricate effects. Hence, trying to optimize all of them at the same time is a very complex task since there is a large number of possibilities. For this reason, we propose to simplify this problem by leveraging existing expert knowledge on the field, by asking experts to design a set of *reasonable* crop management policies to try. For each policy the rules are set in advance, and the farmers will follow them each time the policy is tried.


**Figure 2** – A very basic example of crop-management policy

From the point of view of the learning system, which will be our point of view for the rest of this manuscript, all crop-management policies are in fact considered as black-boxes. We assume that the joint impact of each part of the process may be very hard to model accurately, and that the quantity of data we could collect in real life may not be enough to learn a complex dynamic. We instead assume that we are given a set of policies to try, and that we do not have enough knowledge to model their relationship. We can then define more precisely the objective of the algorithm. Ultimately, it would be to find the policy that would be the best according to the farmers' needs. However, experiments are costly: each bad trial is potentially deleterious to the farmer at all steps of the learning process. Hence, we would like to make farmers try "good" policies most of the time. Summarizing the properties we established in this first part, we obtain the first important characteristic that should be satisfied by a learning system.

> **Characteristic 1:** *The learning algorithm's objective is to recommend most often the best policy among a finite set of policies provided by experts, during the whole learning phase.*

**Evaluation of the policies** In order to characterize our problem the next step is to define how to compare policies and determine which one is the best. We first need to consider the output obtained after a farmer applies a recommendation: at the end of the season, the crops are harvested and the farmer can observe a realized *yield* for this season. The yield is simply the ratio of the quantity of crop harvested divided by the field surface (e.g in kg/ha). More sophisticated criterion can be used, for instance by taking into account the economic and environmental performance of the fertilization policy. In any case, we consider that each recommendation followed by the farmer leads to an observable numeric output at the end of the season, that we call a *reward*. We also assume that the higher a reward is, the better it is for the farmer. The main challenge for the learning strategy is that this output depends on many factors that are external to crop-management: weather conditions during the crop growing process (rainfalls, temperature,...), potential diseases, pests, extreme climate events, crops genetics ...making the outcome very uncertain. Hence, for a fixed policy we can assume that the rewards follow a *probability distribution*, and thus each policy needs to be tried many times as they can all provide both good and bad outcomes.

**Characteristic 2:** *The outcome (reward) when applying a fixed crop-management policy is uncertain, and follows a probability distribution.*

To be able to learn something from the observed data we need some minimum knowledge about these distributions. Fortunately, working with the DSSAT[1] simulator (Hoogenboom et al., 2019) can give us some intuitions. Harnessing more than 30 years of expert knowledge, this simulator is calibrated on historical field data (soil measurements, genetics, planting date...) and generates realistic crop yields. Such simulations can be used to explore crop management policies *in silico* before implementing them in the real world, where collecting enough data would take several years. Recently, Gautron et al. (2022b) implemented a Gym environment compatible with DSSAT, making its access easier for the ML community.

With DSSAT, we can have a look at the kind of distributions we can expect to observe in the real-world experiment. We implemented in the simulator conditions similar to Southern Mali for maize crops, and sampled $10^6$ observations obtained by trying 7 different planting dates. Hence in this example each one of the 7 crop-management policy is simply characterized by a unique planting date, everything else being equal. The resulting distributions incorporate historical variability as well as exogenous randomness coming from a stochastic meteorologic model. We provide the histogram obtained for each policy in Figure 3. The main remark that comes from this figure is that the distributions hardly fit the usual parametric models that are generally used in machine learning (e.g the distributions are gaussian): they are typically right-skewed, multimodal and exhibit a peak at zero corresponding to years of poor harvest. Hence, the question of the right assumptions we can make on the reward distributions is crucial to design our learning algorithm.

**Characteristic 3:** *The reward distributions do not necessarily fit a convenient parametric model (e.g gaussian). We need to consider alternative assumptions.*

Now that we raised the question of the kind of distributions that will generate the rewards we need to define the way to evaluate and rank these distributions (and hence the crop-management policies). The most intuitive criterion is certainly to compare their *expected reward*. However, in our situation it is possible that the expected reward is not a satisfying metric. Assume a context where we want to ensure food security for the community of farmers that is part of the experiment. In that case, if we consider the distributions of Figure 3 we may prefer the distribution in salmon (last one) with an average yield of $3504$ kg/ha and a relatively small probability of harvesting $0$ kg/ha than the cyan distribution ($3^{\text{rd}}$ on top) with an expected

[1] *Decision Support System for Agrotechnology Transfer* is an open-source project maintained by the DSSAT Foundation, see https://dssat.net/

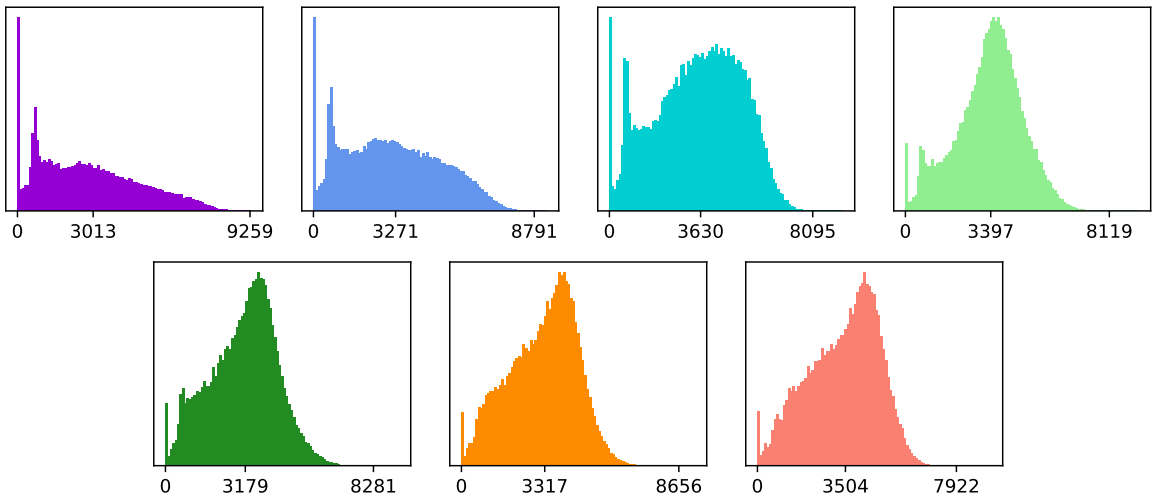


**Figure 3** – Distribution of simulated dry grain yield (kg/ha) for seven different planting dates, all other parameters being equal. Reported on the x-axis are the distribution minimum, mean and maximum values. The optimal arm is the third one (mean 3630 kg/ha) if we want to maximize the expected yield.

reward of 3630 kg/ha (overall winner in terms of expectation) but a relatively large probability of harvesting nothing. This kind of preference can be considered by the learning system by introducing a *risk-aware* performance metric, for instance.

> **Characteristic 4:** *Depending on the preferences of the farmers, the performance of crop-management policies can be evaluated using their* **expected reward** *or by considering* **alternative performance metrics** *(e.g risk-aware).*

The different elements introduced so far are already challenging from a theoretical point of view and are at the core of our main contributions, but we can also introduce some additional features of our problem that are relevant for practical implementation in real-world conditions.

**Non-Stationarity** As the experiment is going to last for several years, we can expect that some factors may change between the beginning and the end (if there is one) of the experiment. For instance climate change may have an impact on the quality of all crop-management policies, potentially making a previously sub-optimal policy become optimal at some point.

> **Characteristic 5:** *The distributions of rewards may evolve with time due to external factors.*

The literature on that topic is already quite diverse, but we will study how one of the algorithms that we developed during this thesis (to address previous points) can be adapted for this context.

**Batch learning** In the previous parts we postulated that a group of farmers is willing to participate in the experiment. We assume that we cannot control their numbers (it is given at each season), but our algorithm needs to be able to simultaneously provide recommendations to many farmers at the beginning of each season, and receive all the corresponding rewards when the season ends. We call the group of farmers participating to the experiment during a given season a *batch*. Ideally, we would like to diversify the recommendations in the early steps.

> **Characteristic 6:** *The learning algorithm needs to be adapted for a "batch" setting, and to diversify the recommendations inside the first batches of the experiment.*

This question was not central during the thesis, but we describe it for completeness in the introduction of this realistic problem. Indeed, it is clear that without the batch setting it would be impossible to collect enough data to train a learning algorithms, as a growing season can take up to one year.

We see that this relatively simply formulated problem already raises several fundamental questions, that are associated with different lines of research in *Multi-Armed Bandit*:

> **Summary of the features of our motivating example:**
>
> - Characteristics 1 and 2 describe the *Stochastic Multi-Armed Bandits* (MAB) problem in the *regret minimization* setting.
> - Characteristic 3 suggests to investigate in details the guarantees that MAB algorithms can obtain according to the *assumptions* they make and use on the distributions. We will explore the different kind of assumptions presented in the literature.
> - Characteristic 4 suggests to consider alternative criterion to the expected reward, with a focus on *risk-awareness*.
> - Characteristic 5 made us explore the vast literature on *non-stationary bandits*.
> - Characteristic 6 suggests to investigate how bandit algorithms can learn with *batch feedback*.

In Chapter 1 we introduce the theoretical formalism associated with all these points, and the related literature. After detailing the existing works on these topics, we explain our theoretical contributions.

# Table of Contents

# Chapter 1

# Introduction to some Bandit Problems

In this chapter we introduce the theoretical formalism associated with the different questions raised by the recommendation problem in agriculture that we introduced in the preliminary part of this thesis. After defining the proper mathematical formalism, we provide an overview of existing works in each of these domains. We also explain how our contributions fit into the context of existing research.

## Contents

## 1.1 Stochastic Multi-Armed Bandits

A Multi-Armed Bandit (MAB) is a sequential decision-making problem in which a learner (or bandit algorithm) sequentially samples from $K$ unknown distributions called arms. In each successive round the learner chooses an arm $A_t \in \{1, \dots, K\}$ and obtains a random reward $X_t$ drawn from the distribution of the chosen arm. The choice of arm $A_t$ depends on the strategy of the learner, that is based on the past observations $\mathcal{H}_t = (A_1, y_1, \dots, A_{t-1}, X_{t-1})$. In the standard formulation of the bandit problem this strategy aims at maximizing the expected sum of rewards obtained after a time horizon $T$. This is equivalent to minimizing the *regret*, defined as the difference between the expected total reward of an oracle strategy always selecting the arm with largest mean and the expected total reward of our strategy. In the rest of this manuscript we denote by $(\nu_k)_{k\in\{1,\dots,K\}}$ the distributions of the arms, $(\mu_1, \dots, \mu_K)$ their means, and assume that the arms belong to a family of distributions $\mathcal{F}$.

**Definition 1.1** (Regret)**.** *Consider a policy $\pi$ and a bandit problem $\nu = (\nu_1, \dots, \nu_K)$. Using the notation $\mu_\star = \max_{k\in\{1,\dots,K\}} \mu_k$, the regret after $T$ rounds is*

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

From that definition, it is clear that a policy that minimizes the regret needs to sample as often as possible arms with means that are close to $\mu_\star$. An arm $k$ with mean $\mu_k < \mu_\star$ is said to be *sub-optimal*, and we call the quantity $\Delta_k = \mu_\star - \mu_k$ the *sub-optimality gap* (often referred to as gap for simplicity). We further define the *number of pulls* of any arm $k \in \{1, \dots, K\}$ at time $T$ as $N_k(T) = \sum_{t=1}^{T} \mathbb{1}(A_t = k)$, which is simply the number of rounds at which arm $k$ has been selected by the algorithm and pulled. Using these two quantities, we can conveniently rewrite the regret as

$$\mathcal{R}_\nu(T, \pi) = \mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^{T}(\mu_\star - \sum_{k=1}^{K} \mu_k \mathbb{1}(A_t = k))\right] = \sum_{k=1}^{K} \Delta_k \mathbb{E}_{\nu,\pi}[N_k(T)]\,. \tag{1.1}$$

Thanks to this equation, it appears clearly that an algorithm with small regret has to minimize the expected number of pulls of each sub-optimal arm. To do that, it needs to balance *exploration* (gaining information about arms that have not been sampled a lot) and *exploitation* (select arms that look promising based on the available information). Before elaborating on the many approaches that have been proposed to solve this problem, it is interesting to have a look on the theoretical guarantees that are achievable by a bandit algorithm. To reformulate, we want to know what is the smallest expected number of pulls of a sub-optimal arm an

algorithm can obtain for a specific class of problem $\mathcal{F}$, that we also call *family of distributions*. In the standard bandit setting the answer to this question has been known for a long time already, with a first lower bound provided by Lai and Robbins (1985) for parametric families of distributions. Before stating their result, we recall the definition of the *Kullback-Leibler divergence* between two distributions $\nu_1$ and $\nu_2$, where $\nu_2$ is absolutely continuous with respect to $\nu_1$:

$$\mathrm{KL}(\nu_1, \nu_2) = \mathbb{E}_{\nu_1}\left[\log\left(\frac{\mathrm{d}\nu_1}{\mathrm{d}\nu_2}\right)\right]$$

**Lemma 1.2** (Lai & Robbins lower bound (Lai and Robbins, 1985))**.** *Assume that the distributions $\nu = (\nu_1, \dots, \nu_K) \subset \mathcal{F}^K$ are continuously parameterized by their means. Then under any uniformly efficient strategy $\pi$, that is any $\pi$ satisfying $\mathcal{R}_\nu(T, \pi) = o(T^\alpha)$ for any $\alpha > 0$ and any $\nu$, the number of pulls of any sub-optimal arm $k$ satisfies*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\nu,\pi}[N_k(T)]}{\log(T)} \geq \frac{1}{\mathrm{kl}(\mu_k, \mu_\star)}\ , \tag{1.2}$$

*where $\mathrm{kl}(\mu, \mu')$ is the Kullback-Leibler divergence between the distribution of mean $\mu$ and that of mean $\mu'$ in the considered family of distributions.*

This result applies to several families of distributions that are widely used in practice, such as Gaussian (with shared variance), Bernoulli, Poisson or Exponential distributions. More generally, the result holds if the distributions come from a *single-parameter exponential family* of distributions (SPEF). A family $\mathcal{F}$ is a SPEF if there exists a parameter set $\Theta$ and some functions $A : \mathbb{R} \mapsto \mathbb{R}$ and $b : \Theta \mapsto \mathbb{R}$ such that for any distribution $\nu \in \mathcal{F}$, there exists a parameter $\theta \in \Theta$ such that $\nu$ (that we denote then by $\nu_\theta$) satisfies

$$\frac{\mathrm{d}\nu_\theta}{\mathrm{d}\eta}(x) = \exp(\theta x - b(\theta))\ ,$$

for any $x$, and where $\eta$ is a reference measure. Some properties of these families of distributions can be found in (Cappé et al., 2013), among many other resources. For instance, the mean of $\nu$ is then equal to $b'(\theta)$ and the KL divergence between the distribution of parameter $\theta_1$ and the one of parameter $\theta_2$ is equal to

$$\mathrm{kl}(b'(\theta_1), b'(\theta_2)) = (\theta_1 - \theta_2)b'(\theta_1) + b(\theta_2) - b(\theta_1)\ .$$

Back to bandit algorithms, the lower bound in Equation (1.2) teaches us that a reasonable strategy can expect to pull the sub-optimal arms a logarithmic number of times, and that it is impossible to obtain a better constant before the logarithm than the one in Equation 1.2. In the first case we simply say that the algorithm has a *logarithmic regret*, while an algorithm *matching*

this lower bound[1] is said to be *asymptotically optimal*. Lemma 1.2 has later been extended by Burnetas and Katehakis (1996) for any (possibly non-parametric) family of distribution $\mathcal{F}$.

**Lemma 1.3** (Burnetas & Katehakis lower bound (Burnetas and Katehakis, 1996))**.** *Consider a bandit $\nu = (\nu_1, \dots, \nu_K) \subset \mathcal{F}^K$. Under any uniformly efficient algorithm the number of pulls of any sub-optimal arm $k$ must satisfy*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\nu,\pi}[N_k(T)]}{\log(T)} \geq \frac{1}{\mathcal{K}_{\inf}^{\mathcal{F}}(\nu_k, \mu^\star)} , \tag{1.3}$$

*where $\mathcal{K}_{\inf}^{\mathcal{F}}(\nu_k, \mu^\star) = \inf_{G\in\mathcal{F}} \{\mathrm{KL}(\nu_k, G) : \mathbb{E}_G(X) > \mu^\star\}$.*

Again, any algorithm matching this lower bound is said to be *asymptotically optimal* for the family $\mathcal{F}$. A very common example of such *non-parametric* family is the set of distributions with supports admitting a known upper bound $B$. In that case, the $\mathcal{K}_{\inf}^{\mathcal{F}}$ has been studied in depth by (Honda and Takemura, 2010), who provide its dual form and use it to derive an asymptotically optimal strategy.

**Remark 1.4** (Other notions of optimality)**.** *In this thesis we study the optimality of bandit algorithms in terms of problem-dependent guarantees. Other kind of results are studied in the literature, such as* minimax *(worst-case) optimality (see for instance Chapters 9 and 15 of Lattimore and Szepesvári (2020)). These bounds are typically in $\mathcal{O}(\sqrt{KT})$, and the constant does not depend on the instance of the problem.*

*Notably, a line of work on* ***best-of both worlds*** *consists in deriving algorithms that would achieve simultaneously optimal problem-dependent regret for stochastic bandits and optimal regret for adversarial bandits (Zimmert and Seldin, 2021; Ito et al., 2022), but with a weaker notion of problem-dependent optimality in the stochastic case. These work are out of the scope of this thesis.*

### 1.1.1 A non-exhaustive overview of multi-armed bandit algorithms

Multi-Armed Bandits have been a very active research field in the past years, and summarizing all contributions to this domain would be a tremendous task. For this reason we detail in this section a selection of bandit algorithms, with a special focus on the ones that are asymptotically optimal. We refer the reader to e.g (Lattimore and Szepesvári, 2020) for a broader survey of this research area. Motivated by the third characteristic that we raised in the use-case considered in the introduction we are specifically interested by the following questions: what type of

[1] i.e that satisfies $\liminf_{T\to\infty} \frac{\mathbb{E}_{\nu,\pi}[N_k(T)]}{\log(T)} \leq \frac{1}{\mathrm{kl}(\mu_k, \mu_\star)}$

approaches can lead to optimal algorithms? For what type of distributions do they achieve optimality? What kind of prior knowledge on the distribution do they require to be optimal?

Before detailing the main families of algorithms that can be found in the literature we define an *index policy*, which is a generic name for a bandit algorithm that (1) computes a quantity (index) for each arm using past observations for this arm only, and (2) pulls the arm with the largest index. We detail this principle in Algorithm 1.1, that will be useful in the next paragraphs since many bandit algorithms are index policies.

**Definition 1.5** (Index policy). *Consider any function $I$ that, given the time horizon $t$ and a set of observations $\mathcal{X}$ returns a scalar $I(\mathcal{X}, t)$ (the output can be random or deterministic). Then, algorithm 1.1 is called an index policy, based on $I$.*

**Input:** Horizon $T$, $K$ arms, function $I$
**for** $t \in \{1, \dots, K\}$ **do**
  Pull arm $t$, set $\mathcal{X}_t = \{X_t\}$ ; ▷ Initialize by pulling each arm once
**end**
**for** $t \in \{K+1, \dots, T\}$ **do**
  **for** $k \in \{1, \dots, K\}$ **do**
    Get $I_k = I(\mathcal{X}_k, t)$ ; ▷ Obtain the value of the index for each arm
  **end**
  Pull arm $A_t = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} I_k$ ; ▷ Pull the arm with the largest index
  Observe $X_t$ ; ▷ Collect the corresponding reward
  Update $\mathcal{X}_{A_t} = \mathcal{X}_{A_t} \cup \{X_t\}$ ; ▷ Add the reward to the history of $A_t$
**end**

**Algorithm 1.1:** Generic Index Policy

**Optimism in Face of Uncertainty** This family contains the famous Upper Confidence Bound algorithm (UCB1) (Agrawal, 1995; Auer et al., 2002a). Algorithms based on the UCB principle achieve logarithmic regret when it is possible to derive a *concentration inequality* on the empirical means, typically in a setting with bounded/sub-gaussian distributions where the support/sub-gaussianity parameter is known. For instance, UCB1 is an index policy that computes an "optimistic" value for the empirical mean of each arm. For distributions bounded in $[0, 1]$ this strategy takes the form of Algorithm 1.2

Unfortunately, this simple strategy is not asymptotically optimal. A refinement of UCB with tighter confidence intervals using the $\mathcal{K}_{\inf}$ quantity provided in the lower bound was proposed in (Cappé et al., 2013) in order to achieve asymptotic optimality: the KL-UCB algorithm is

**Input:** Data $\mathcal{Y} = (y_1, \dots, y_n)$, time $t$

**Return:** $I_{\text{UCB}} = \frac{1}{n}\sum_{i=1}^{n} y_i + \sqrt{\frac{2\log(t)}{n}}$ ; ▷ Mean of observations + Confidence Bonus

**Algorithm 1.2:** Index of UCB1 (Auer et al., 2002a) for a distribution supported in $[0, 1]$

optimal for bounded distributions with known upper bound, and kl-UCB is optimal for SPEF. Interestingly, in a recent paper (Agrawal et al., 2021a) the KL-UCB algorithm has been proved to be optimal for a third class of distributions with a bounded-moment condition, which is a usual assumption for heavy tail distributions. We report in Algorithm 1.3 a generic KL-UCB strategy, given that the family of distributions $\mathcal{F}$ is known and the function $\mathcal{K}_{\text{inf}}^{\mathcal{F}}$ can be computed.

**Input:** Empirical distribution $F_{\mathcal{Y}}$ built with $n$ observations, $t$, family $\mathcal{F}$, function $f$

**Return:** $I_{\mathcal{K}_{\text{inf}}^{\mathcal{F}}\text{-UCB}} = \max\{\mu : n\mathcal{K}_{\text{inf}}^{\mathcal{F}}(F_{\mathcal{Y}}, \mu) \leq f(t)\}$

**Algorithm 1.3:** Index of $\mathcal{K}_{\text{inf}}^{\mathcal{F}}$-UCB (Cappé et al., 2013; Agrawal et al., 2021a)

Note that for parametric family (kl-UCB) the $\mathcal{K}_{\text{inf}}^{\mathcal{F}}$ function is replaced by the $\mathrm{kl}$ of Equation (1.2). The threshold $f$ is typically of the form $f(t) = \log(t) + c\log\log(t)$, with a parameter $c > 2$ (Cappé et al., 2013). Hence, the *Optimism in Face of Uncertainty* paradigm can lead to optimal algorithms with appropriate knowledge on the family of distributions $\mathcal{F}$.

**Divergence-based algorithms** We introduce a second class of algorithms that we consider as a standard approach in Multi-Armed Bandits, relying on empirical estimates of $\mathcal{K}_{\text{inf}}^{\mathcal{F}}$ when the family $\mathcal{F}$ is known. Let us denote by $F_k(t)$ the empirical distribution of arm $k$ at time $t$, $\mu_k(t)$ its empirical mean, and define $\mu_\star(t) = \max_{k\in\{1,\dots,K\}} \mu_k(t)$. Then, the idea of these algorithms is to directly use the empirical divergence $\mathcal{K}_{\text{inf}}^{\mathcal{F}}(F_k(t), \mu_\star(t))$. Indeed, this quantity measures how far the empirical distribution of an arm is to the current best arm, and with appropriate concentration tools this is sufficient to build optimal strategies. Three approaches have been developed using this principle: MED (Honda and Takemura, 2011), DMED (Honda and Takemura, 2010) and IMED (Honda and Takemura, 2015). The first one is a randomized algorithm, where at each round the probability of sampling an arm is proportional to $\exp(-N_k(t)\mathcal{K}_{\text{inf}}^{\mathcal{F}}(F_k(t), \mu_\star(t)))$. Surprisingly this simple but intuitive strategy is proved to be optimal only for multinomial distributions (Honda and Takemura, 2010). It was however recently re-discovered in (Bian and Jun, 2022) under the name Maillard Sampling (MS) and analyzed for sub-gaussian distributions (the authors proved logarithmic regret), showing its potential. On the other hand, DMED and

IMED are both deterministic approaches that compare the empirical divergences but adding some costs to ensure sufficient exploration. As an example, we detail IMED in Algorithm 1.4. Note that IMED is not an index policy in the sense that it uses the empirical best average at the current time step for all arms. However, for simplicity we write it as an index policy, where the minus sign is due to the fact that the selected arm *minimizes* the function inside the parenthesis.

**Input:** Empirical distribution $F_{\mathcal{Y}}$ computed with $n$ data, family $\mathcal{F}$, current best empirical mean $\widehat{\mu}_\star$

**Return:** $I_{\text{IMED}} = -\left(n\mathcal{K}_{\text{inf}}^{\mathcal{F}}(F_{\mathcal{Y}}, \widehat{\mu}_\star) + \log(n)\right)$

**Algorithm 1.4:** Indexed Minimum Empirical Divergence (Honda and Takemura, 2015)

As for KL-UCB, this algorithm can be implemented for any family of distributions $\mathcal{F}$ for which $\mathcal{K}_{\text{inf}}^{\mathcal{F}}$ can be computed. In (Honda and Takemura, 2010, 2015) the authors respectively prove the asymptotic optimality of DMED and IMED for bounded distributions with known upper bound. IMED was later proved to be also optimal for light-tailed SPEF by Pesquerel et al. (2021), and works very well in practice (see our experiments in Chapter 2 and 6).

**Thompson Sampling** The last widespread category of algorithms is *Thompson Sampling (TS)*, which is a family of Bayesian algorithms named after Thompson (1933). Contrarily to UCB-based methods TS is a *randomized* algorithm, which means that if we run a step of the algorithm twice with the same data a different arm can be pulled. Indeed, the learner provides a *prior* distribution on the means of each arm to the algorithm, that then computes at each step the corresponding *posterior* distribution given the observations collected for each arm. Then, a step of TS consists in *sampling* a parameter (or to simplify, a mean) for each arm according to their posterior distribution, and to choose the arm with the largest sampled mean. Hence, TS is still an index policy, but with a randomized index. We summarize TS for a general prior/posterior distribution in Algorithm 1.5.

**Input:** Data $\mathcal{Y}$, prior distribution $p$

**if** $\mathcal{Y} = \emptyset$: sample $\widetilde{\mu} \sim p$

**else** Sample $\widetilde{\mu} \sim \mathcal{P}(\mathcal{Y}, p)$, where $\mathcal{P}(\mathcal{Y}, p)$ is the posterior distribution after observing the dataset $\mathcal{Y}$ and using the prior $p$.

**Algorithm 1.5:** Sampling step of TS for a general prior/posterior

Historically, TS was first studied for Bernoulli distributions since the application considered in (Thompson, 1933) was clinical trials. In this setting TS using a Beta-Bernoulli conjugate prior is asymptotically optimal, which was proved much later by (Agrawal and Goyal, 2013a; Kaufmann et al., 2012). This result was then extended to general SPEF in (Korda et al., 2013), using Jeffreys prior. Intuitively, the algorithm works because: (1) in the first rounds the prior distribution allows to explore each arm a sufficient number of times, and (2) when arms are sampled enough the posteriors are concentrated around the true means so that the algorithm will mostly exploit the best arm.

Recently, the principle of TS has been extended to obtain an optimal algorithm for bounded distributions with known upper bound: *Non-Parametric Thompson Sampling (NPTS)* (Riou and Honda, 2020). The idea of the authors is to build on the Beta-Bernoulli Thompson sampling. They first extend it for Multinomial distributions, where the Beta-Bernoulli prior/posteriors are naturally replaced by Dirichlet distributions (see Appendix 5.7 for a short summary of the properties of this distribution). The resulting *Multinomial TS* algorithm is asymptotically optimal when the arms have a known finite support. Then, applying the same mechanism for general bounded distributions leads to using a sample from a Dirichlet distribution of parameter $(1, \dots, 1)$ to *re-weight* the observations. We denote this distribution by $\mathcal{D}_n$ for $n$-dimensional weights. This procedure is actually known as the *Bayesian Bootstrap*, introduced by Rubin (1981). To ensure sufficient exploration the known upper bound of the support is added to the set of observations at the beginning of the experiment. We summarize NPTS in Algorithm 1.6.

1 **Input:** Data $\mathcal{Y} = (y_1, \dots, y_n)$ upper bound $B$

2 **Return:** $I_{\text{NPTS}} = \sum_{i=1}^{n} w_i y_i + w_n B$, where $w \sim \mathcal{D}_{n+1}$

**Algorithm 1.6:** Index of Non Parametric Thompson Sampling (Riou and Honda, 2020)

The fact that NPTS is asymptotically optimal is noteworthy, because contrarily to KL-UCB it does not require to explicitly compute $\mathcal{K}_{\text{inf}}^{\mathcal{F}}$. This function naturally appears in the concentration inequalities related to the Dirichlet distribution. This remarkable feature made us consider this algorithm more thoroughly, and Part II of this thesis is dedicated to extensions of NPTS.

**Re-Sampling algorithms** In the past years, there has been a surge of interest for the design of non-parametric algorithms that would *perturb* the empirical distribution of the data instead of trying to fit it in an already defined model, and are therefore good candidates for the problem in agriculture that we are considering. A first line of works explored *re-sampling* schemes (Efron and Tibshirani, 1994) to balance exploration and exploitation (Osband and Roy, 2015; Kveton

et al., 2019a,b; Wang et al., 2020). The idea is to compute the mean of a noisy version of the empirical distribution by, for instance, drawing random weights for each of their observations. In (Kveton et al., 2019b) the authors propose the term of *General Randomized Exploration* for index policies satisfying this principle.

In fact, Thompson Sampling algorithms are part of the GRE framework, but this family contains also non-Bayesian algorithms such as the non-parametric bootstrap (sampling with replacement) presented in Algorithm 1.7.

**Input:** Data $\mathcal{Y} = (y_1, \dots, y_n)$
**Return:** $I_B = \frac{1}{n} \sum_{i=1}^{n} z_i$, where $\forall i$, $z_i$ is drawn uniformly at random in $\mathcal{Y}$

**Algorithm 1.7:** Index based on sampling with replacement

Unfortunately, re-sampling in itself is not enough: an index policy based on Algorithm 1.7 trivially fails to achieve sub-linear regret in some cases. The question is then to determine both what kind of re-sampling procedure and/or what kind of modifications to the empirical data are needed to make this method work. In (Kveton et al., 2019a,b), the authors propose to perturb the empirical mean by adding fake rewards in the history (by either directly adding them or sampling them at each step). They prove that when appropriately tuned these algorithms can achieve logarithmic regret for bounded distributions. Interestingly, for the gaussian/sub-gaussian case Wang et al. (2020) proposed an algorithm with gaussian weights, that also achieves logarithmic regret. So far NPTS (Riou and Honda, 2020) is the only algorithm using this principle that achieve optimality in its setting, using the Bayesian Bootstrap. An interesting open question is to determine if non-Bayesian algorithms based on bootstrapping could achieve optimal regret.

Intuitively, the reason for the failure of the non-parametric bootstrap is the potential under-estimation of a good arm due to bad rewards in the first draws. To circumvent this issue, TS uses a convenient prior distribution while the re-sampling based algorithms introduce additional data points to improve exploration (e.g fake good samples). Interestingly, another class of algorithms consider an alternative by instead penalizing the arms that have been pulled the most so far, and hence do not require any exploration bonus: these algorithms are based on *sub-sampling*. This idea was introduced in (Baransi et al., 2014) with *Best Empirical Sample Average (BESA)*. It relies on pairwise comparisons between arms: the arm was has been less pulled (challenger) uses its empirical mean, while the other arm uses the empirical mean of a *sub-sample* of its history of the same size as the challenger's. The core idea is that if the two means are computed with the same number of samples the comparison is "fair". In BESA, the sub-sample is drawn using sampling without replacement. We summarize this comparison step in Algorithm 1.8.

**Input:** Arm 1 with $\mathcal{X} = (x_1, \dots, x_n)$, arm 2 with $\mathcal{Y} = (y_1, \dots, y_m)$, $n \geq m$
Compute $\mu_y = \frac{1}{m} \sum_{i=1}^{m} y_i$
Draw $(z_1, \dots, z_m)$ without replacement in $\mathcal{X}$, compute $\mu_z = \frac{1}{m} \sum_{i=1}^{m} z_i$
**Return:** $\text{argmax}_{y,z} \{\mu_y, \mu_z\}$ ; ▷ `Return winning arm`

**Algorithm 1.8:** pairwise comparison in BESA (Baransi et al., 2014)

When there are only 2 arms, at each round the arm winning the comparison is pulled. With more arms, Baransi et al. (2014) propose to organize a *tournament*: in successive rounds half of the arms is eliminated until only 1 arm remains. Unfortunately, the tournament is hard to analyze and Baransi et al. (2014) prove logarithmic regret only for $K = 2$. Furthermore, some of the assumptions they required may not be valid in general. However, the general idea of comparing sub-samples of the same size is interesting and differs from all existing approaches. More recently, Chan (2020) further explored this idea by proposing the *Sub-Sample Mean Comparison (SSMC)* algorithm. To analyze the algorithm for $K > 2$ arms, the author propose a more convenient *leader vs challenger* approach: the arm that has been pulled the most so far is defined as *leader*, and then competes with every other arm (called challengers) in pairwise comparisons. Then, any winning challenger or the leader (if none) is pulled. Another difference with BESA is that this time the sub-sample used is not random: it is the *worst* sequence of successive observations (in order of collection). Hence, the mean of this sub-sample is intuitively an empirical lower bound of the true mean of the leader. We summarize the comparison step of SSMC in Algorithm 1.9

**Input:** Arm 1 with $\mathcal{X} = (x_1, \dots, x_n)$, arm 2 with $\mathcal{Y} = (y_1, \dots, y_m)$, $n \geq m$
Compute $\mu_y = \frac{1}{m} \sum_{i=1}^{m} y_i$
Compute $\mu_x = \min_{j \in \{1,\dots,n-m+1\}} \frac{1}{m} \sum_{i=0}^{m-1} x_{j+i}$
**Return:** $\text{argmax}_{x,y} \{\mu_x, \mu_y\}$ ; ▷ `Return winning arm`

**Algorithm 1.9:** pairwise comparison in SSMC (Chan, 2020)

The strength of these algorithms is that they *do not use any information on the arm's distributions*. While the performance of BESA is not clear, in (Chan, 2020) the authors prove that SSMC is asymptotically optimal when arms come from the same SPEF. This is actually a very strong result, since the identity of the SPEF does not have to be specified by the learner. In the works presented in this thesis we aim at extending these results and bridging the gap between BESA and SSMC.

**Summary** In the previous paragraphs we detailed the three dominant families of bandit algorithms (UCB-based algorithms, Thompson Sampling, and Minimum Empirical Divergence), as well as alternative approaches based on *re-sampling* and *sub-sampling*. For the more standard approaches we focused on the ones achieving asymptotic optimality, in the sense that their regret matches the lower bound of Burnetas and Katehakis (1996). We motivate this choice by the practical considerations related to the recommendation problem in agriculture that we introduced in the preamble of this thesis. Indeed, we are more interested in problem-dependent guarantees, since there is no reason for the bandit problem we consider to be arbitrarily difficult in practice. Then, the experiments that we perform in the upcoming chapters actually show that the asymptotically optimal algorithms are also performing better in practice in our examples, including the ones using a realistic crop-yield simulator emulating our problem. We provide in Table 1.1 a summary of asymptotically optimal algorithms, the family of distributions for which they are optimal, and the knowledge they require to achieve these guarantees.

**Table 1.1** – Comparison of competitor bandit algorithms matching the Burnetas & Katehakis bound for various assumptions on an arm distribution $\nu$. Elements listed as parameters are considered prior knowledge and are used within the algorithm.

| Algorithm | Scope for optimality | Algorithm parameters |
|---|---|---|
| kl-UCB[1] | Single Parameter Exponential Family (SPEF) $(\nu_\theta)_{\theta\in\Theta}$ | $\mathrm{kl}(\theta,\theta')$ |
| IMED[2] | | $\mathrm{kl}(\theta,\theta')$ |
| Thompson Sampling[3] | | Prior/Posterior |
| SSMC[4] | | Non-Parametric |
| IMED[2] | $\mathrm{Supp}(\nu)\subset(-\infty,B]$ $\nu$ is light-tailed$^\star$ | $\mathcal{K}_{\inf}^{\mathcal{F}_B}$ |
| Empirical KL-UCB[1] | $\mathrm{Supp}(\nu)\subset[b,B]$ | $\mathcal{K}_{\inf}^{\mathcal{F}_B}$ |
| NPTS[5] | | $B$ |
| $\mathrm{KL}_{\inf}$-UCB[6] | $\mathcal{F}_{\varepsilon,B}=\{\nu:\mathbb{E}[\lvert X\rvert^{1+\varepsilon}]\leq B\}$ | $\mathcal{K}_{\inf}^{\mathcal{F}_{\varepsilon,B}}$ |

1. Cappé et al. (2013), 2. Honda and Takemura (2015), 3. Thompson (1933); Agrawal and Goyal (2013a); Korda et al. (2013), 4. Chan (2020), 5. Riou and Honda (2020), 6. Agrawal et al. (2021a)

$^\star$ i.e there exists some $\lambda_0\in\mathbb{R}$ such that $\mathbb{E}_{X\sim\nu}[e^{\lambda X}]<+\infty$ for all $\lambda\in[-\lambda_0,\lambda_0]$.

**Contribution** We consider light-tailed distributions, since extreme events in agriculture are more likely to lead to poor yields than exceptionally good ones. We consider the two dominant assumptions encountered in the literature: Single-Parameter Exponential Families and bounded distributions. In each setting, one approach caught our attention: NPTS for bounded distributions, as not having to compute the $\mathcal{K}_{\inf}$ function at each step is computationally appealing; and SSMC for SPEF since it does not require the knowledge of

the SPEF to be optimal. Hence, we naturally studied extensions of these two algorithms to consider the question:

*How can we design algorithms with the best theoretical guarantees with a minimum knowledge on the distributions?*.

The first part of this thesis is dedicated to *sub-sampling algorithms*, inspired by SSMC and BESA, while in the second part we study some generalizations of NPTS. Regarding the standard regret minimization problem, in Chapter 2 we propose a family of algorithms called *Sub-Sampling Dueling Algorithm* that is optimal for SPEF, but also achieves logarithmic regret on a broader class of distributions that we further characterize. In particular, we analyze the assumptions that are needed to obtain at least a logarithmic regret in a more general setting than the ones with fully-parametric assumption. Then, in Chapter 6 we study a generalization of Non-Parametric Thompson Sampling outside the family of bounded distributions with a known upper bound, from alternative families of bounded distributions to general light-tailed distributions. For the *Dirichlet Sampling* algorithm that we propose, we found a trade-off between the theoretical guarantees (optimal, logarithmic, or super-logarithmic regret) and the generality of the family of distributions considered, that can be resolved by the practitioner depending on the knowledge available on the distributions.

## 1.2 Bandits with alternative performance metrics

Over the past few years, a number of works have focused on adapting multi-armed bandit strategies to optimize another criterion than the *expected* cumulative reward. Indeed, in a large number of application domains (healthcare, agriculture, marketing,...), one needs to take into account personalized *preferences* of the practitioner that are not captured by the expected reward. For example, in the preamble of this thesis we introduced a crop-management policy recommendation problem: small farmers are typically *risk-averse* as their harvest is necessary to ensure the subsistence of their household.

Assume that the learner decides to evaluate a bandit algorithm with a metric $U$. Then, consider two datasets of $n$ collected points $\mathcal{X} = (X_1, \dots, X_n)$ and $\mathcal{Y} = (Y_1, \dots, Y_n)$ and assume that $U$ can return a numeric value for $U(\mathcal{X})$ and $U(\mathcal{Y})$ for any value of $n$. We say that the trajectory $\mathcal{X}$ is better than $\mathcal{Y}$ with respect to $U$ if $U(\mathcal{X}) > U(\mathcal{Y})$. This is sufficient to propose a natural adaptation of the expected regret for an alternative performance metric $U$.

**Definition 1.6** (U-regret). *Consider a metric $U$ that can return a numeric value for any trajectory of observations $\mathcal{Y}_T = (Y_1, \dots, Y_T)$ for any time horizon $T$, and that larger values of $U$ are preferred by the learner. Then, the $U$-regret can be defined as*

$$\mathcal{R}_\nu^U(T, \pi) = \mathbb{E}_{\nu,\pi^\star}\left[U(\mathcal{Y}_T)\right] - \mathbb{E}_{\nu,\pi}\left[U(\mathcal{Y}_T)\right] \ ,$$

*where $\pi^\star$ is the oracle optimal policy, $\pi^\star = \text{argmax}_\pi \ \mathbb{E}_{\nu,\pi}\left[U(\mathcal{Y}_T)\right]$.*

Sometimes the performance metric can only be defined for a sample $X_1, \dots, X_n$ (e.g extreme values, range), but in the majority of cases it can be defined more generally for a *probability distribution*. Being able to compute the value of a performance metric for general distributions is interesting because it allows us to evaluate the arms of a bandit problem of distributions $(\nu_1, \dots, \nu_K)$. We choose to refer to the distributions through their *cumulative distribution functions* (cdf) $F_1, \dots, F_K$, and to use the notation $U(F)$ (where $F$ is a cdf). Then, we can compare the quality of the arms by simply computing $U(F_1), \dots, U(F_K)$. This allows us to define the *sub-optimality* gap of an arm in an analogous way as for the standard expected regret.

**Definition 1.7** (Sub-optimality gaps in terms of metric U). *Consider a bandit $\nu = (\nu_1, \dots, \nu_K) \subset \mathcal{F}^K$ with respective cdf $F_1, \dots, F_K$. Consider a metric $U : \mathcal{F} \to \mathbb{R}$, then the* U-gap *of any arm $k \in \{1, \dots, K\}$ is defined as*

$$\Delta_k^U = \max_{j \in \{1,\dots,K\}} U(F_j) - U(F_k)$$

Interestingly, when $U$ is not linear (as the expectation) the U-regret of Definition 1.6 may not be convenient to obtain theoretical analysis of bandit strategies. In that case, it may be relevant to consider an alternative regret definition inspired by Equation (1.1) and involving the sub-optimality gaps, called *proxy regrey*.

**Definition 1.8** (U-proxy regret). *Consider a bandit $\nu = (\nu_1, \dots, \nu_K) \subset \mathcal{F}^K$ with respective cdf $F_1, \dots, F_K$. Denote their respective U-gaps by $\Delta_1^U, \dots, \Delta_k^U$, then the $U$-proxy regret is defined as*

$$\mathcal{R}_\nu^{U,proxy}(T, \pi) = \sum_{t=1}^{T} \Delta_k^U \ \mathbb{E}_{\nu,\pi}\left[N_k(T)\right]$$

**Remark 1.9.** *This definition of the proxy regret is different from the one in (Cassel et al., 2018), that consider another definition of sub-optimality gaps. However, as regret proofs for the proxy regret consist in upper bounding the expected number of pulls of sub-optimal arms the guarantees obtained with one definition easily translate to the other.*

### 1.2.1 Risk Metrics

In statistics, a natural way to analyze a distribution is to compute its *moments*. The standard bandit theory is based on evaluating the *expectation* of the arms, but one could want to consider moments of higher order. For example some criteria aim at penalizing arms with high variance, such as the *mean-variance* (Markowitz, 1952) and the *sharpe ratio* (Sharpe, 1994). The first one is a linear combination of the two measures, while the second is the ratio of mean and standard deviation. Both are widely used in finance due to their simplicity. However, if the distributions have a complex shape the mean and variance may not capture well their risk profile. It may be interesting to consider instead a quantile, or *Value-at-Risk (VaR)*. Hence, we denote by $\text{VaR}_\alpha$ the quantile of order $\alpha \in [0, 1]$. In order to better capture the behavior on the tail of the distribution, one can further consider the *Conditional-Value-at-Risk* (Artzner et al., 1999; Rockafellar et al., 2000). Several definitions of the CVaR exist in the literature, depending on whether the samples are considered as *losses* or as *rewards*. We consider the reward version: given a level $\alpha \in (0, 1]$, the $\text{CVaR}_\alpha$ is easily interpretable as the expected reward in the worst $\alpha$-fraction of the outcomes. It can hence capture different preferences: $\alpha = 1$ simply provides the expectation of the distribution, which expresses risk-neutrality, while making $\alpha$ very small corresponds to maximizing the expected reward in the worst-case scenarios. We illustrate this in Figure 1.3. CVaR is further a coherent spectral measure in the sense of Rockafellar et al. (2000), see (Acerbi and Tasche, 2002)). This definition entails some properties that are generally considered as desirable for a risk metric. Other general families of risk measures have been defined, such as *spectral risk measures (SRM)* (Acerbi and Tasche, 2002), and *cumulative prospect theory (CPT)* (Tversky and Kahneman, 1992). Finally, Entropic Risk is also a way to penalize distributions with heavy tails. We summarize in Table 1.2 some of the measures introduced in this paragraph with their proper definition (for a distribution of CDF $F$) and parameters.

We see that the risk metrics generally require to choose a parameter, that needs to be fixed in advance by the learner to model its level of risk-aversion. In the next paragraph we introduce the literature considering risk metrics to evaluate the performance of bandit algorithms.

**Risk-aware bandits** At a high level, the multi-armed bandit literature considering risk metrics is largely based on adapting the popular Upper Confidence Bounds (UCB) algorithms (Auer et al. (2002a)), and is hence mainly focused on deriving appropriate concentration tools for the

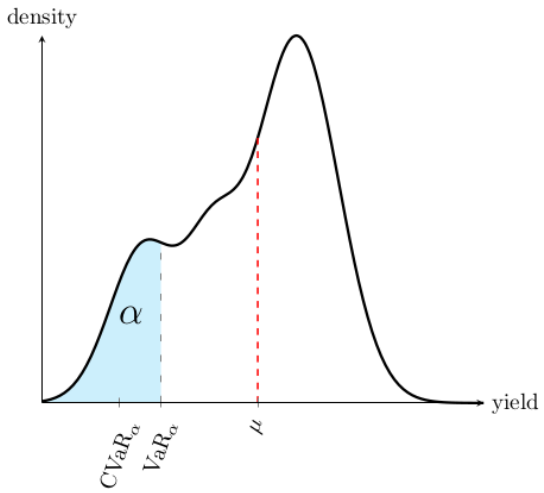


**Figure 1.1** – High risk aversion ($\alpha \approx 20\%$)

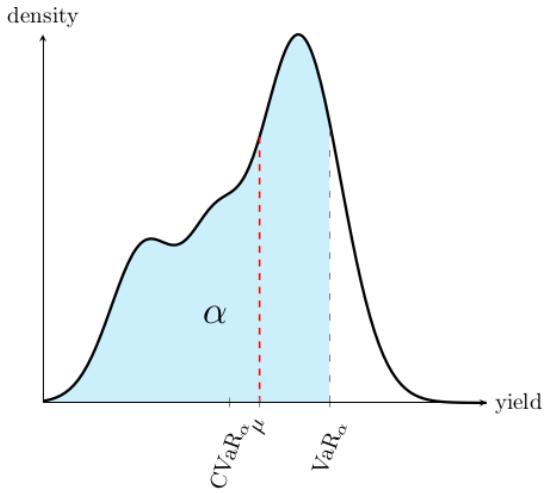


**Figure 1.2** – Low risk aversion ($\alpha \approx 80\%$)

**Figure 1.3** – The Conditional Value-at-Risk (CVaR) of level $\alpha$ is the mean value of the blue area of the distribution, that stops at $\text{VaR}_\alpha$. The red line is the average $\mu$ of the distribution.

risk metric under consideration. The first line of works on this topic followed this principle, and considered variations of the mean-variance criterion (Sani et al., 2012; Vakili and Zhao, 2015; Vakili and Zhao, 2016; Zimin et al., 2014). Szorenyi et al. (2015) study algorithms for the quantile (Value-at-Risk) criterion in both the regret minimization and the pure exploration setting, while (David and Shimkin, 2016; Zhang and Ong, 2021) investigate the second problem only. Interestingly, the algorithm proposed in (Szorenyi et al., 2015) for regret minimization implements optimism for VaR by comparing the arms using rank statistics of larger order than the one corresponding to the target quantile. Maillard (2013) focuses on the Entropic Risk and extend the KL-UCB algorithm of Cappé et al. (2013) for this risk metric under the assumption that the distributions are bounded with a known upper bound.

More recently the *Conditional Value at Risk* (CVaR) have received specific attention from the bandit community (Galichet et al., 2013; Galichet, 2015; Tamkin et al., 2020; Prashanth et al., 2020) to cite a few). These works focus on proving or refining concentration inequalities for CVaR under different assumptions on the distributions (e.g as those of Brown (2007); Thomas and Learned-Miller (2019); Prashanth et al. (2020); Holland and Haress (2020); Bhat and L.A. (2019a)), and analyzing the corresponding bandit algorithm. Interestingly, Tamkin et al. (2020) exhibits two possible approaches to implement optimism for CVaR bandits: adding directly an exploration bonus to the empirical CVaR as in MaRaB (Galichet et al., 2013; Galichet, 2015), U-UCB (Cassel et al., 2018) or Brown-UCB (Brown, 2007; Tamkin et al., 2020); or exploiting the link between the CVaR and the CDF to build an optimistic CDF as in CVaR-UCB (Tamkin et al., 2020), resorting to the celebrated Dvoretzky–Kiefer–Wolfowitz (DKW) concentration inequality (see Massart (1990)). While the two approaches are equivalent for $\alpha = 1$ (mean setting), the

**Table 1.2** – Overview of risk metrics

| Metric | Definition | Parameters |
|---|---|---|
| Expectation | $\mathbb{E}_F[X] = \int x \mathrm{d}F$ | |
| Variance | $\mathbb{V}_F[X] = \sigma_F(X) = \int (X - \mathbb{E}_F[X])^2 \mathrm{d}F$ | |
| Mean-Variance | $\mathrm{MV}_\rho = \mathbb{E}_F[X] - \rho \mathbb{V}_F[X]$ | scaling $\rho$ |
| Sharpe ratio | $\mathrm{SR}_{r_0} = \frac{\mathbb{E}_F[X] - r_0}{\sigma_F(X)}$ | reference value $r_0$ |
| Value-at-Risk | $\mathrm{VaR}_\alpha = \sup\{x \in \mathbb{R} : F(x) \leq \alpha\}$ | quantile level $\alpha$ |
| Conditional Value-at-Risk | $\mathrm{CVaR}_\alpha = \sup_{x \in \mathbb{R}} \left\{ x - \frac{1}{\alpha} \int (x - X)^+ \mathrm{d}F \right\}$ | quantile level $\alpha$ |
| Entropic Risk | $\mathrm{ER}_\theta = -\frac{1}{\theta} \log \left( \int \exp(-\theta x) \mathrm{d}F \right)$ | risk level $\theta$ |
| Spectral Risk | $\mathrm{SRM}_\phi = \int_0^1 \phi(\beta) F^{-1}(\beta) \mathrm{d}\beta$ | risk spectrum $\phi$ |

empirical results from (Tamkin et al., 2020) suggest that the second method should be preferred for $\alpha < 1$. Our empirical results in Chapter 5 confirm these findings.

Notably, Cassel et al. (2018) provided a unified UCB1-like algorithm for various risk metrics called U-UCB. They show that if the risk-metric is *quasi-convex*[2] and is a *strongly stable* performance metric the regret of an appropriately tuned U-UCB is logarithmic. The tuning depends on the metric $U$, but also on the class of distributions considered, and scales in $\sqrt{\log(T)/N_k(t)}$ in the case where the metric is Lipschitz. Strong stability is defined by the two following properties (from (Cassel et al., 2018)).

**Definition 1.10** (Strongly stable performance metric (Cassel et al., 2018))**.** *A metric is said to be strongly stable if:*

1. *There exists $b > 0$, $q > 1$ and a seminorm $||.||$ such that on $\mathcal{F}$ extended to the set of empirical distributions it holds that*

$$|U(F) - U(G)| \leq b \times (||F - G|| + ||F - G||^q) \ .$$

2. *There exists $a > 0$ such that for any $F \in \mathcal{F}$ and $x > 0$*

$$\mathbb{P}\left(||F_n - F|| \geq x\right) \leq 2 \exp(-anx^2) \ ,$$

*where $F_n$ denotes an empirical distribution corresponding to $n$ samples drawn from $F$.*

[2] i.e for two distributions $F, G$ and $\lambda \in [0, 1]$ it holds that $U(\lambda F + (1 - \lambda)G) \leq \max\{U(F), U(G)\}$.

For appropriate sets of distributions (e.g bounded), all the risk metrics we mentioned in this section can satisfy this requirement. Hence, the analysis of U-UCB is very general. Investigating if other families of algorithms could achieve this level of generality is an interesting perspective. However, investigating other methods that would be tailored for a specific problem is also still interesting since (1) the generality of U-UCB may have a cost on the empirical performance (the confidence bound may not be tight), and (2) U-UCB adapts UCB1, that is known to perform worse in practice than asymptotically optimal algorithms like TS or KL-UCB for the standard bandit problem.

**Input:** Horizon $T$, $K$ arms, performance metric $U$, $(b, q, a)$ from def 1.10
Define $I_{\text{U-UCB}} : (\mathcal{Y} = (y_1, \dots, y_n), t) \mapsto U(F_{\mathcal{Y}}) + \phi\left(\frac{\gamma \log(T)}{n}\right)$ ,
where $\phi : x \mapsto \max\left\{2b\left(\frac{x}{a}\right)^{1/2}, 2b\left(\frac{x}{a}\right)^{1/2}\right\}$
**Return:** IP($T$, $K$ arms, $I_{\text{U-UCB}}$)

**Algorithm 1.10:** U-UCB (Cassel et al., 2018)

So far the only algorithms inspired by Thompson Sampling for risk-aware bandits are restricted to the fully-parametric gaussian case: Zhu and Tan (2020) analyzed the mean-variance criterion, while Chang et al. (2020) considered a risk-constrained setting mixing expectation and CVaR. Hence, developing new TS algorithms for risk-aware bandits, especially for non-parametric settings, is an interesting perspective. Finally, the notion of *asymptotically optimal* bandit algorithm (Burnetas and Katehakis, 1996) has not been explored yet in risk-aware bandits and defining the best achievable performance in that case is an interesting question that we answer for the CVaR metric. Furthermore, this explains why divergence-based strategies have not been developed yet for this setting.

**Contribution** In Chapter 5 we introduce an optimal Thompson Sampling algorithm for CVaR bandits, building on the NPTS algorithm of Riou and Honda (2020). We chose to study the CVaR because it is an easy to interpret and widely used metric, and the choice of the quantile level $\alpha$ allows to model different possible levels of risk-awareness. Furthermore, we think it may be better adapted than the risk metrics based on moments (Mean-Variance, Sharpe ratio, ...) for the case study introduced in the foreword of this thesis, as the distributions displayed in Figure 3 suggest that the mean and variance may not be satisfying to evaluate them for this problem. We considered that the most accessible information to the practitioner is often whether or not the distribution is discrete, and for the continuous case how it is bounded. This assumption seems reasonable in applications where the reward is bounded due to physical constraints. First, we extended the lower

bound of Burnetas and Katehakis (1996) for CVaR bandits in this setting, establishing the guarantees that should be achieve by asymptotically optimal algorithms. Then, we extended the Non-Parametric Thompson Sampling algorithm under the two assumptions considered: multinomial distributions (M-CVTS) and bounded distributions with known support (B-CVTS). We prove that both algorithms are asymptotically optimal for the settings they consider, which is the first result of this type for risk-averse bandits. Finally we show empirically the benefits of the TS approaches over UCB-based algorithms in practice, on problems using both synthetic data and the DSSAT simulator.

### 1.2.2 Extreme Bandits

In the previous paragraphs we introduced the vast literature on risk-aware bandits, and in particular CVaR bandits that is the topic of Chapter 5. We explained that the CVaR relies on a quantile level $\alpha \in (0, 1]$, where a small $\alpha$ allows to model risk-averse preferences. The question we can ask is: what happens if we actually want to set $\alpha = 0$? Intuitively, this would correspond to an extremely risk-averse learner, wanting to avoid as much as possible *worst case* scenarii. While the CVaR is not defined for $\alpha = 0$, we can instead turn to *extreme statistics*, by trying for instance to maximize the expected minimum value collected by the bandit algorithm during a trajectory. A similar problem has been introduced in the literature by Cicirello and Smith (2005), and is known as *Extreme Bandits*. In this setting, the learner's objective is simply to collect the *largest* possible reward. While the objectives are rather different, the theoretical problems faced when trying to maximize the expected minimum or maximum are very similar. Hence, the algorithms proposed in the literature for Extreme Bandits could also be applied in our "extremely risk-averse" setting. For this reason in the following we introduce the literature associated with the Extreme Bandit problem, even if the risk-averse objective actually suits better our ideas of applications.

Letting $X_{k,t}$ be the reward obtained from arm $k$ at time $t$, a bandit algorithm selects an arm $I_t$ using past observations and receives the reward $X_{I_t,t}$. The rewards stream $(X_{k,t})$ is drawn i.i.d. from $\nu_k$ and independently from other rewards streams. The case where distributions have bounded support is studied by Nishihara et al. (2016), so in the following we assume that the distributions have unbounded supports. To evaluate an extreme bandit algorithm, Carpentier and Valko (2014) propose an adaptation of the regret called *extreme regret* that fits Definition 1.6 for a performance metric $U$ that returns the maximum of a set of observations,

$$\mathcal{R}_T^\pi = \max_{k \leq K} \mathbb{E}[\max_{t \leq T} X_{k,t}] - \mathbb{E}_\pi[\max_{t \leq T} X_{I_t,t}] \, . \tag{1.4}$$

This problem is hard because we are only interested in the asymptotic behavior of the right tail of the distribution: an arm could be the best for extreme bandits while providing very bad reward $99\%$ of the time, so intuitively most observations may be useless (or even misleading in some sense) to the algorithm. Furthermore, the optimal policy may change according to the time horizon. Consider a simple example with two gaussian distributions $\mathcal{N}(1,1)$ and $\mathcal{N}(0,1.7)$: for short time horizons the first arm should be preferred, for example with $T=1$ the arm with largest expectation should be preferred. On the other hand, for large time horizons the arm with the largest variance will provide the best expected maximum. We illustrate this in Figure 1.4 below.

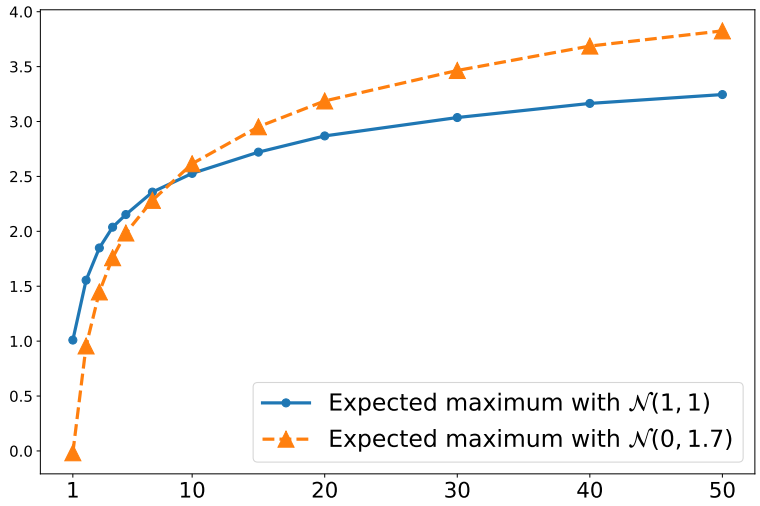


**Figure 1.4** – Average maximum on $10^4$ samples obtained from $\mathcal{N}(1,1)$ and $\mathcal{N}(1,1.7)$ for a number of observations ranging from $1$ to $50$.

In this simple example there exists some time horizon $T_0$ for which the optimal policy changes. For this reason, we will consider *asymptotic* performance guarantees with the assumption that one arm asymptotically *dominates* the others (i.e the optimal policy can only change a finite number of times). Two types of performance guarantees have been derived in previous works. Using the terminology of Bhatt et al. (2021), we introduce these two definitions below.

**Definition 1.11.** *An Extreme Bandit algorithm $\pi$ has a vanishing regret in the weak sense if*

$$\mathcal{R}_T^\pi = \underset{T\to\infty}{o}\left(\max_{k\leq K}\mathbb{E}[\max_{t\leq T} X_{k,t}]\right) \tag{1.5}$$

*and $\pi$ has a vanishing regret in the strong sense if*

$$\lim_{T\to\infty}\mathcal{R}_T^\pi = 0\,. \tag{1.6}$$

It is now understood that the peculiarities of the Extreme Bandits setting make the algorithms designed for the $K$-arm setting suboptimal. For this reason a line of works have designed algorithms specifically for this setting. Furthermore, the question of the assumptions that are made on the distributions needs to be considered. Existing algorithms for this problem can be divided into three categories:

1. Fully-parametric approaches (Cicirello and Smith, 2005; Streeter and Smith, 2006a) where the family of distributions is assumed to be known (e.g Frechet, Gumbel).

2. Semi-parametric approaches: (Carpentier and Valko, 2014; Achab et al., 2017) consider a setting where distributions satisfy a second-order Pareto assumption.

3. Distribution-free approaches (Streeter and Smith, 2006b; Bhatt et al., 2021), that do not leverage any assumption on the reward distributions. Assumptions are only required for the analysis of the algorithms.

The fully-parametric setting is not so different from the standard bandit, since the problem is reduced to using the estimated parameters to balance exploration and exploitation. The semi-parametric setting allows more flexibility. To illustrate this, we detail the *second-order Pareto* assumption used in (Carpentier and Valko, 2014; Achab et al., 2017).

**Definition 1.12** (Second order Pareto (definition 2 in (Carpentier and Valko, 2014))). *A distribution $F$ is se$(\alpha, \beta, C, C')$-second order Pareto if for $x \geq 0$:*

$$|\mathbb{P}_F(X \geq x) - Cx^{-\alpha}| \leq C' x^{-\alpha(1+\beta)} ,$$

*which implies that $\mathbb{P}_F(X \geq x) = Cx^{-\alpha} + \mathcal{O}(x^{-\alpha(1+\beta)})$.*

With this definition the tail of the distribution is asymptotically very close to the tail of a Pareto distribution, and the parameter $\beta$ controls the deviation wrt this asymptotic equivalent. In (Carpentier and Valko, 2014), weakly vanishing regret is obtained under this assumption, further assuming that a lower bound on a parameter of the distribution is known to the algorithm. Achab et al. (2017) refined this analysis and obtained strongly vanishing regret when this lower-bound is large enough. However, their approach still relies on the estimation of the parameters. The variant compared with the fully-parametric approaches is that a decreasing fraction of the samples (depending on the lower bound on $\beta$) is used to perform this estimation.

Regarding non-parametric algorithms, a first algorithm inspired by UCB1, ThresholdAscent, was proposed in (Streeter and Smith, 2006b). The principle is to compute the mean of a fixed number of the largest samples obtain for an arm, with an exploration bonus. Unfortunately, this very simple algorithm do not have theoretical guarantees, and its performance is very

sensitive to the number of observations kept for the computation. Bhatt et al. (2021) recently proposed Max-Median, an algorithm based on robust statistics that can be employed for any kind of distribution. Max-Median is proved to have weak vanishing regret for polynomial-like arms and strongly vanishing regret for exponential-like arms.

**Contribution** The recent work of Bhatt et al. (2021) suggests that non-parametric approaches could be an efficient way to tackle the Extreme Bandit problem. This can be performed by finding a right way to *compare* the arms instead of individually estimating some parameters. In Chapter 4, we propose a novel algorithm for Extreme Bandits based on *sub-sampling*. This algorithm extends the *Sub-Sampling Dueling Algorithm* introduced in Chapter 2, comparing arms with a robust estimator. We establish weakly vanishing regret for non-parametric distributions, assuming only that one tail *dominates* the others (we properly define this notion in the chapter), and refine these results for polynomial and exponential arms. We believe these guarantees to be the most general obtained so far for this problem.

## 1.3 Non-stationary Bandits

In the previous sections we considered bandit problems where the arms do not change during the experiment. However, in many practical application they are likely to evolve: for example, if we consider again the crop-management problem we can assume that if the experience lasts for several years then climate change may have an impact on the reward distribution associated with each policy.

In the following we get back to the standard setting where the learner aims at maximizing her expected sum of rewards, as in Section 1.1, and introduce the different algorithms that have been developed in the literature to handle non-stationary rewards. Before that we define the *dynamic regret*, that naturally adapts Definition 1.1.

**Definition 1.13** (Dynamic regret)**.** *For a policy $\pi$, the dynamic regret after $T$ rounds is defined as*

$$\mathcal{R}_\nu(T,\pi) = \sum_{t=1}^{T} \max_{k\in\{1,\dots,K\}} \mu_{k,t} - \mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^{T} X_t\right] ,$$

*where $\nu = (\nu_{k,t})_{k\in\{1,\dots,K\},t\in\{1,\dots,T\}}$ denotes the set of distributions corresponding to all arms at each time step, and $\mu_{k,t}$ denotes the mean of $\nu_{k,t}$ (arm $k$ at time $t$).*

The question is then : *can a bandit algorithm learn to minimize the dynamic regret?* If the distributions can change completely arbitrarily, this task seems clearly impossible: it is necessary to make some assumptions on the non stationarity. A natural idea consists in fitting a stochastic model on the dynamic, as for instance Whittle (1988) that consider a markovian dynamic for the arms. Under this assumption, changes can then be predicted by learning the model. However, in many applications the learner may not have access to such knowledge and have to rely on simpler assumptions. One of the most famous is assuming an *abruptly changing* (or *piece-wise stationary*) environment: the distributions are assumed to be stationary between *breakpoints*, which are the time instants when they can change. This number has to be limited to allow the strategy to learn between the changes, and hence the *number of breakpoints* (denoted by $\Gamma_T$) is assumed to be sub-linear in the time-horizon.

**Definition 1.14** (Number of breakpoints)**.** *The number of break-points in a piece-wise stationary model is formally defined as*

$$\Gamma_T = \sum_{t=1}^{T-1} \mathbb{1}(\exists k \in \{1, \dots, K\} : \mu_{k,t} \neq \mu_{k,t+1}) \ ,$$

*where the times* $\mathcal{T} = (t_1, t_2, \dots, t_{\Gamma_T}) \coloneqq \{t \in \{1, \dots, T\} : \mathbb{1}(\exists k \in \{1, \dots, K\} : \mu_{k,t} \neq \mu_{k,t+1})\}$ *are called the* breakpoints *and a time interval* $[t_i + 1, t_{i+1}]$ *is called a* stationary phase.

Intuitively, the existence of these breakpoints forces the bandit algorithms to explore much more than in the standard stochastic setting. Indeed, a strategy with a regret in $\mathcal{O}(\log(T))$ only marginally explores arms that have been observed as sub-optimal after a sufficient number of pulls. As breakpoints can occur at any time, an algorithm evolving in a piecewise-stationary environment has to try regularly all arms to check if they have changed or not. This intuition is formalized by a lower bound on the dynamic regret, first proved in Garivier and Moulines (2011).

**Theorem 1.15** (Theorem 31.2 in Lattimore and Szepesvári (2020))**.** *Let* $k = 2$*, and fix* $\Delta \in (0, 1)$ *and a policy* $\pi$*. Let* $\mu$ *be so that* $\mu_{i,t} = \mu_i$ *is constant for both arms and* $\Delta = \mu_1 - \mu_2 > 0$*. If the expected regret* $\mathcal{R}_\nu(T, \pi)$ *of policy* $\pi$ *on bandit* $\mu$ *satisfies* $\mathcal{R}_\nu(T, \pi) = o(T)$*, then for all sufficiently large* $T$*, there exists a non-stationary bandit* $\nu'$ *with at most two change points and* $\min_{t \in [T]} |\mu'_{1,t} - \mu'_{2,t}| \geq \Delta$ *such that*

$$\mathcal{R}_\nu(T, \pi) \geq \frac{T}{22\mathcal{R}_{\nu'}(T, \pi)} \ .$$

This result shows that no strategy can hope to obtain a better regret than $\Omega(\sqrt{T})$ for this class of problems. In particular, it shows that the standard bandit strategies with $\mathcal{O}(\log(T))$ regret in the stationary case will fail to achieve a sub-linear regret on all instances. The lower bound was later refined by Seznec et al. (2020) who included the dependency in the number of arms and number of breakpoints.

**Theorem 1.16** (Proposition 4 in Seznec et al. (2020))**.** *In a piece-wise stationary environment with at most $\Gamma_T$ breakpoints, there exists an environment such that*

$$\mathcal{R}(T, \pi) \geq \sqrt{KT\Gamma_T} \, .$$

Hence, the best achievable worst-case guarantee that can be expected from a bandit algorithm in an abruptly changing environment is of order $\mathcal{O}(\sqrt{KT\Gamma_T})$

An alternative way to model non-stationarity is to consider possibly smoother changes by introducing a *variation budget* (Besbes et al., 2014) that controls the total amplitude of arms' changes.

**Definition 1.17.** *Define a bandit $\nu = (\mu_{k,t})_{k\in\{1,\dots,K\},t\in\{1,\dots,T\}}$, its variation budget is defined as*

$$B_T \coloneqq \sum_{t=1}^{T-1} \max_{k\in\{1,\dots,K\}} |\mu_{k,t+1} - \mu_{k,t}| \, .$$

Contrarily to the piece-wise stationary model, this setting can allow for instance small changes at any time step. A special case of *slowly drifting* environment, where the amplitude of changes is controlled, was introduced recently in (Krishnamurthy and Gopalan, 2021). Besbes et al. (2014) also proved a lower bound for the regret for the variation budget setting.

**Theorem 1.18** (Theorem 1 in Besbes et al. (2014))**.** *Assume that rewards have a Bernoulli distribution. Then, there is some absolute constant $C > 0$ such that for any policy $\pi$ and for any $T \geq 1$, $K \geq 2$ and $B_T \in [1/K, T/K]$, the regret satisfies*

$$\mathcal{R}_\nu(T, \pi) \geq C(KB_T)^{1/3}T^{2/3} \, .$$

This result shows that a sub-linear regret is possible only if the variation budget is sub-linear. We can also remark that the regret is in $\Omega(T^{2/3})$ instead of $\Omega(\sqrt{T})$ for the piece-wise

stationary model. Now that we introduced those two main assumptions we can detail the different families of algorithms that have been proposed in the literature.

### 1.3.1 Algorithms for non-stationary bandits

During the past ten years, several works have considered non-stationary variants of the multi-armed bandit model, proposing methods that can be grouped into two main categories: they passively forget past information (Garivier and Moulines, 2011; Raj and Kalyani, 2017; Trovo et al., 2020), or actively try to detect modifications in the distribution of the arms with change-point detection algorithms (Liu et al., 2017; Cao et al., 2019; Auer et al., 2019; Chen et al., 2019; Besson et al., 2022).

**Passively forgetting strategy** A natural idea is to consider adaptations of standard bandit algorithms with simple mechanism to forget past data. The two standard idea for that are the use of a *sliding window* or *discounted rewards*. In the first case all the algorithm works with only the most recent collected rewards: if the window size is $\tau$ and rewards collected are denoted by $X_1, \dots, X_T$ then only rewards $X_{T-\tau+1}, \dots, X_T$ are used. In the second case all rewards are kept but the oldest rewards are discounted in order to have a reduced impact: using a fixed discount rate $\gamma$, at time $T$ the reward collected at time $t$ is associated with a discount $\gamma^{T-t}$. In Garivier and Moulines (2011) the authors analyze two variants of UCB1: SW-UCB and D-UCB, respectively implementing UCB1 with a sliding window and discounted rewards. The algorithms are proved to achieve a $\mathcal{O}(\sqrt{KT\Gamma_T \log(T)})$ dynamic regret in abruptly changing environment, with an appropriate tuning of the window/discount factor requiring the knowledge of the order of $\Gamma_T$. Similarly the celebrated Thompson Sampling algorithm (Thompson, 1933) has also been adapted to include these mechanisms, with the Discounted Thompson Sampling (DTS) (Raj and Kalyani, 2017) and the Sliding Window Thompson Sampling (SW-TS) (Trovo et al., 2020), though theoretical guarantees have been obtained only for SW-TS. In Algorithm 1.11 and 1.12, we respectively provide one time step of the generic adaptation of any bandit algorithm for stationary rewards with a sliding window and a discount factor.

For example, we could naturally adapt KL-UCB (Cappé et al., 2013) and IMED (Honda and Takemura, 2015) in the discounted computing the divergence function of an arm $k$ with $n$ rewards $y_1, \dots, y_n$ with the discounted empirical distribution $F_k(t) : x \mapsto \sum_{i=1}^{n} \rho_i \mathbb{1}(y_i \leq x)$, where the weights $\rho_1, \dots, \rho_n$ are computed following Algorithm 1.12 (and normalized) according to the time where all rewards have been collected.

Another line of work inspired by *adversarial bandits* have been studied for non stationary settings. In short, in adversarial bandits the rewards are not drawn from a probability distributions but can be any arbitrary sequence: in this harder setting an adversary can decide the next rewards, and hence try to confuse the algorithm. However, a usual assumption is that

**Input:** Sliding window $\tau$, rewards $\mathcal{X} = (X_1, \dots, X_t)$ and corresponding arms $\mathcal{K} = (k_1, \dots, k_t)$, Bandit algorithm $\pi$
**if** $t \leq \tau$ **then**
  **return** *Arm chosen by* $\pi$ *using* $\mathcal{X}$ *and* $\mathcal{K}$
**end**
**else**
  **return** *Arm chosen by* $\pi$ *using* $(X_{t-\tau+1}, \dots, X_t)$ *and* $(k_{t-\tau+1}, \dots, k_t)$
**end**

**Algorithm 1.11:** Generic sliding-window strategy

**Input:** Discount factor $\rho$, rewards $\mathcal{Y} = (y_1, \dots, X_t)$ and corresponding arms $\mathcal{K} = (k_1, \dots, k_t)$, Bandit algorithm $\pi$
**return** *Arm chosen by* $\pi$ *using* $\mathcal{Y}$ *with discounts* $\bar{\rho} = (\rho^t, \rho^{t-1}, \dots, \rho^2, \rho, 1)$ *and* $\mathcal{K}$

**Algorithm 1.12:** Generic strategy with discounted rewards

the rewards are chosen independently of the bandit algorithm: the adversary is said to be *non-adaptive*. In that case, the objective of the learner is to sample most often the action with the best trajectory of rewards for the time horizon $T$ considered: the best arm in hindsight. To make it possible to solve this problem rewards are usually assumed to be bounded in $[0, 1]$. As the adversarial setting is very general the algorithms proposed for this problem can be used to tackle non-stationary rewards, even if their guarantees are defined according to the best arm in hindsight and not the dynamic regret. For instance, the EXP3.S algorithm (Auer et al., 2002b) can be used to benchmark non-stationary bandit algorithms, and even has guarantees in the piecewise stationary setting. Furthermore, Besbes et al. (2014) proposed the Rexp3 strategy, which is simply a combination of EXP3 (Auer et al., 2002b) and scheduled restarts of the algorithm. Interestingly, Rexp3 obtains an optimal worst-case regret for the variation budget setting when an the budget $B_T$ are known.

**Change-point detection algorithms** The second main category of non-stationary bandit algorithms consist in combining a stochastic bandit algorithm with a *change-point detector* (CPD). As the name states, a CPD scans the history of rewards collected to determine if a change of distributions occurred. If this is the case, the algorithm simply erases the history and restarts. Two changepoint detection algorithms, CUSUM (Liu et al., 2017) and M-UCB (Cao et al., 2019), have been proposed using the standard UCB algorithm. They obtain optimal guarantees for respectively Bernoulli and bounded rewards in the piece-wise stationary environment, with some assumptions on the detectability of the changes and the knowledge of the number

of changes. Interestingly, a number of works have proposed algorithms that get rid of this assumption and are instead fully adaptative (Chen et al., 2019; Auer et al., 2019; Besson et al., 2022). For example, Besson et al. (2022) consider a GLR test combined with the KL-UCB algorithm (GLR-KL-UCB) and an exploration scheme depending on the number of changes already detected. GLR-KL-UCB achieves optimal guarantees in the abruptly changing setting under a detectability assumption. On the other hand, (Chen et al., 2019; Auer et al., 2019) propose algorithms based on elimination rules of empirically sub-optimal arms and scheduled replay phases for the eliminated arms. ADSWITCH (Auer et al., 2019) is optimal in abruptly changing environment, while ADA-ILTCB$^+$ is optimal in both abruptly changing and variation budget settings.

Recently, some works have tried to improve the practical performance of CPD-based algorithm by considering *most significant changes* instead of restarting every time a change is detected. Indeed, intuitively if the distributions change but the best arm remains the same we do not want our algorithm to restart completely. For this reason, considering a number of breakpoints or a variation budget may be too conservative with respect to the true complexity of the problem. For instance, Manegueu et al. (2021) proposed a change point algorithm based on the *empirical gaps* between the arms. Suk and Kpotufe (2022) further extended this idea by quantifying and trying to detect *significant shifts* at each step of the algorithm. They also remarkably avoid using knowledge on the non-stationarity by using an elaborate method to re-explore sub-optimal arms often enough, and obtain optimal guarantees for both piecewise-stationary and variation budget assumptions. Notably, an independent and parallel work of Abbasi-Yadkori et al. (2022) also obtained comparable (but slightly weaker) guarantees in the abruptly changing setting with similar ideas.

**Contribution** The family of non-stationary bandit algorithms is now very dense, and under the assumption that rewards are bounded the practitioner can choose between many options. This choice can depend on the complexity of the algorithms (the ones with the best guarantees are unfortunately difficult to implement), their theoretical guarantees, and the assumptions they need on the *non-stationarity structure*. To the best of our knowledge, no work have tried to combine a mechanism for non-stationarity and a non-parametric algorithm, able to tackle more structural changes. We study this in Chapter 3, with the combination of a simple sliding window mechanism and the Last-Block Sub-sampling Dueling Algorithm (LB-SDA) introduced in Chapter 2. We show that LB-SDA can achieve optimal guarantees in the abruptly changing environment when the order of the number of breakpoints is known. The main advantage of the resulting SW-LB-SDA algorithm is that these guarantees allow more general changes of distributions than previous approaches.

## 1.4 Batch Bandits

The formulation of batch bandits can be traced back to (Perchet et al., 2015), and this problem have been studied for instance in (Gao et al., 2019; Esfandiari et al., 2021; Jin et al., 2021). Outside of the example that we introduced in agriculture, clinical trials is a typical example where this setting applies: patients come in cohorts, and the practitioner analyzes the results for each cohort before considering the next one. The main questions considered in this literature are to determine the number of batches, and the size of each batch necessary to obtain theoretical guarantees that are as close as possible as the ones obtained in the purely sequential setting. For example, Jin et al. (2021) consider grids of exponentially increasing size and obtain a logarithmic regret (in the total number of trials) with a batch size of order $\Omega(\log\log(T))$. Kalkanli and Ozgur (2021) analyze Thompson Sampling for Gaussian rewards coupled with an algorithm that defines the size of each batch accordingly to the previous plays of TS. They prove a logarithmic regret and an expected number of batches of $\mathcal{O}(\log\log T)$. While this line of work is interesting it does not exactly correspond to the experiment introduced in the preamble of this thesis. Indeed, in our crop-management problem the number of farmers is fixed at the beginning of the experiment and we cannot decide how many farmers will participate at each season.

We propose in Theorem 5.11 an analysis of the B-CVTS algorithm introduced in Chapter 5 in the batch setting, showing no performance loss compared with the purely sequential setting. This analysis shows that this algorithm is suitable for the experiment in agriculture introduced at the beginning of this thesis.

## 1.5 Outline and Contributions

### 1.5.1 Thesis organization

After this introductory chapter, the thesis is divided in two parts corresponding to the two families of algorithms that we studied. Each part provides an initial version of the algorithm for the standard MAB setting, and then details their extensions to some of the variants of MAB introduced in the previous sections.

The first part introduces the family of *Sub-Sampling Dueling Algorithms* (SDA), which is a novel family of non-parametric bandit algorithms relying on fair pairwise comparisons between arms using sub-samples to penalize the arms that have been explored the most.

- In Chapter 2 we introduce SDA in the standard MAB setting. We present both randomized (RB-SDA, WR-SDA) and deterministic (LB-SDA) sub-sampling mechanisms. We prove logarithmic regret for both RB-SDA and LB-SDA under some conditions on the arms' distributions. In particular, we prove that for standard parametric assumptions (e.g Bernoulli, Gaussian with shared variance) these algorithms are even *asymptotically optimal*. This is interesting because these guarantees hold without the algorithm using any prior knowledge of what distributions it is facing. We further discuss the merits of each algorithm from a practical point of view.

- In Chapter 3 we extend the LB-SDA algorithm to tackle non-stationary rewards using a sliding window and some additional mechanisms, which yields the SW-LB-SDA algorithm. We prove that in abruptly changing environments the window can be tuned efficiently to make the algorithm optimal, which is on par with similar strategies in the literature. However, the non-parametric nature of SW-LB-SDA allows to obtain these guarantees while allowing a potentially broader class of possible distribution changes.

- In Chapter 4 we prove that combining LB-SDA with a robust estimator based on the upper tail of distributions can lead to efficient algorithms for the Extreme Bandit problem. This setting is an example of bandits with alternative performance metric, as in that case the learner wants to collect the largest possible reward. We illustrate the interest of non-parametric approaches in this setting, where the learner would like to make as little assumptions as possible on the tails of the distributions.

In the second part of this thesis we study algorithms inspired by the *Non-Parametric Thompson Sampling* (NPTS) algorithm of Riou and Honda (2020).

- In Chapter 5 we propose an extension of NPTS for CVaR bandit under the name *CVaR Thompson Sampling* (CVTS), with one algorithm for multinomial and one for bounded distributions. After extending the notion of *asymptotic optimality* for CVaR bandits, we

prove that the resulting M-CVTS and B-CVTS algorithms are the first algorithms to be asymptotically optimal in a risk-aware setting. We further propose an empirical evaluation of B-CVTS using the DSSAT simulator, that allows to emulate the recommendation problem in agriculture introduced in the foreword of this thesis.

- In Chapter 6 we get back to the usual setting (maximizing the expected sum of rewards) and propose extensions of Non Parametric Thompson Sampling for a broader class than bounded distributions with a known upper bound: we call these strategies *Dirichlet Sampling* (DS) algorithms. We propose different relaxations of this assumption and study the performance guarantees of DS in each case. Interestingly, we exhibit a trade-off between the level of generality of the assumptions and the theoretical guarantees that can be obtained by the algorithms. However, our experiments show that in practice all algorithms have similar performance and thus suggest that the most general algorithm, namely *Robust Dirichlet Sampling* (RDS), should be used in practice.

### List of publications

This section references the works that have been published during my PhD thesis. They are the fruit of research projects that were conducted with several collaborators, including my supervisors, other PhD students, and several researchers. They are ordered according to their date of publication, from the oldest to the most recent.

#### Publications in international conferences with proceedings

- *Sub-sampling for Efficient Non-Parametric Bandit Exploration*, by **Dorian Baudry**, Emilie Kaufmann & Odalric-Ambrym Maillard. *Advances in Neural Information Processing Systems 34* (NeurIPS 2020). See Chapter 2. (Baudry et al., 2020).

- *Optimal Thompson Sampling strategies for support-aware CVaR bandits*, by **Dorian Baudry** & Romain Gautron & Emilie Kaufmann & Odalric-Ambrym Maillard. *Proceedings of the the Thirty-eighth International Conference on Machine Learning* (ICML 2021). See Chapter 5. (Baudry et al., 2021a).

- *On Limited-Memory Subsampling Strategies for Bandits*, by **Dorian Baudry** & Yoan Russac & Olivier Cappé. *Proceedings of the the Thirty-eighth International Conference on Machine Learning* (ICML 2021). See Chapter 2 and 3. (Baudry et al., 2021b).

- *From Optimality to Robustness: Adaptive Re-Sampling Strategies in Stochastic Bandits*, by **Dorian Baudry** & Patrick Saux & Odalric-Ambrym Maillard. *Advances in Neural Information Processing Systems 35* (NeurIPS 2021). See Chapter 6. (Baudry et al., 2021c).

- *Efficient Algorithms for Extreme Bandits*, by **Dorian Baudry** & Yoan Russac & Emilie Kaufmann. *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics* (AISTATS 2022). See Chapter 4. (Baudry et al., 2022).
- *Top-Two algorithms revisited*, by Marc Jourdan & Remy Degenne & **Dorian Baudry** & Rianne de Heide & Emilie Kaufmann. *Advances in Neural Information Processing Systems 35* (NeurIPS 2022) (Jourdan et al., 2022). We do not detail this work in this thesis, as it tackles a quite different pure-exploration objective.

**Submitted works**

- *Towards an efficient and risk aware strategy for guiding farmers in identifying best crop management*, by Romain Gautron & **Dorian Baudry** & Myriam Adam & Gatien N. Falconnier & Marc Corbeels. Theorem 5.11 comes from this work.

# Part I

# Bandit Algorithms Based on Sub-Sampling

# Chapter 2

# Sub-Sampling Dueling Algorithms

In this chapter we propose a new family of bandit algorithms based on sub-sampling, called *Sub-sampling Dueling Algorithms* (SDA). Unlike most existing approaches these algorithms do not use any knowledge on the arms' distributions. Still, we prove that some instances SDA can achieve strong theoretical guarantees under some assumptions that we detail in this chapter. In particular, they are *asymptotically optimal* when distributions come from the same Single-Parameter Exponential Family, including some of the most common distributions encountered in practice. After introducing this novel family of algorithms, we analyze some instances of SDA and highlight the core properties that make this strategy work. Finally, we perform an experimental study assessing the flexibility and robustness of this promising novel approach for exploration in bandit models. This chapter unifies the results that were published in (Baudry et al., 2020) and (Baudry et al., 2021b), respectively for the RB-SDA and LB-SDA algorithms, and provides additional intuitions on this family of algorithms.

## Contents

## 2.1 Introduction

In this chapter we consider the standard Multi-Armed Bandit problem introduced in Chapter 1 (Section 1.1). We recall that a $K$-armed bandit is a sequential decision-making problem in which a learner sequentially samples from $K$ unknown distributions, called arms. In each round the learner chooses an arm $A_t \in \{1, \dots, K\}$ and obtains a random reward $X_t$ drawn from the distribution of the chosen arm, that has mean $\mu_{A_t}$. The learner should adjust her sequential sampling strategy $\pi$ (or bandit algorithm) in order to maximize the expected sum of rewards obtained after $T$ selections. This is equivalent to minimizing the *regret*, defined as

$$\mathcal{R}_\nu(T, \pi) = \max_{k \in \{1,\dots,K\}} \mu_k T - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} \left(\max_{k \in \{1,\dots,K\}} \mu_k - \mu_{A_t}\right)\right];,$$

An algorithm with small regret needs to balance exploration (gain information about arms that have not been sampled a lot) and exploitation (select arms that look promising based on the available information). In Chapter 1 we introduced the standard approaches used in bandits. We also detailed the lower bounds of (Lai and Robbins, 1985; Burnetas and Katehakis, 1996), that set the target performance that we would like to obtain with a bandit algorithm. We further discussed that most existing algorithms require knowledge on the arms' distribution, e.g to calibrate confidence bounds or use appropriate conjugate prior/posteriors with Bayesian methods to reach these guarantees. The main question we study in this chapter is:

*Under which conditions can we achieve strong theoretical guarantees with an algorithm that do not use any knowledge on the arms' distributions?*

A recent family of algorithms based on *sub-sampling* (Baransi et al., 2014; Chan, 2020), introduced in Section 1.1, seems promising to tackle this question. The *Sub-sampling Dueling Algorithms* (SDA) that we propose in this chapter combine an algorithm structure based on *pairwise comparisons* (that we call duels), inspired by SSMC, and generic sub-sampling schemes. Our objective is to bridge the gap between BESA (Baransi et al., 2014) and SSMC (Chan, 2020) by allowing both *randomized* and *deterministic* sub-sampling schemes in the same framework.

The rest of the chapter is structured as follows. In Section 2.2 we introduce the generic principle of SDA and present different instances corresponding to possible choices for the sub-sampling algorithms (that we also call *samplers*). In Section 2.3 we present general theoretical results for SDA. We provide a first regret upper bound for a category of sub-sampling algorithms that we call *Block Samplers*, only assuming that the empirical means concentrate with an exponential decay for each arm. This bound exhibits two terms: a term in $\mathcal{O}(\log(T))$, and the sum of the probabilities that the optimal arm is under-sampled (i.e sampled less than some quantity). We then further study this second term, exhibiting some sufficient conditions

on the sub-sampling algorithm and the family of distributions to obtain a logarithmic regret. In Section 2.4 we first show that the condition on the sampler is satisfied by at least two instances: RB-SDA and LB-SDA. We then propose a condition on the family of distributions ensuring that SDA is sufficient to avoid under-exploring the best arm. Finally, in Section 2.5 we present the results of an empirical study comparing several instances of SDA to asymptotically optimal parametric algorithms, and other algorithms based on re-sampling or sub-sampling. These experiments reveal the robustness of the SDA approaches, which match the performance of Thompson Sampling, without exploiting the knowledge of the distribution.

## 2.2 Sub-sampling Dueling Algorithms

In the following we define Sub-sampling Duelling Algorithms (SDA). We first introduce a few notation: for every integer $n$, we let $[n] = \{1, \dots, n\}$. We denote by $(Y_{k,s})_{s\in\mathbb{N}}$ the i.i.d. sequence of successive rewards from arm $k$, drawn from a distribution $\nu_k$ with mean $\mu_k$. For every finite subset $\mathcal{S}$ of $\mathbb{N}$, we denote by $\bar{Y}_{k,\mathcal{S}}$ the empirical mean of the observations of arm $k$ indexed by $\mathcal{S}$: if $|\mathcal{S}| > 1$, $\bar{Y}_{k,\mathcal{S}} := \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} Y_{k,i}$. We also let $\bar{Y}_{k,n}$ as a shorthand notation for $\bar{Y}_{k,[n]}$.

**A round-based algorithm** Unlike index policies, SDA relies on *rounds*, in which several arms can be played (at most once). In each round $r$ the learner selects a subset of arms $\mathcal{A}_r \subseteq \{1, \dots, K\}$, and receives $Y_{k,N_k(r)}$ for each arm $k \in \mathcal{A}_r$, where $N_k(r) := \sum_{s=1}^{r} \mathbb{1}(k \in \mathcal{A}_s)$ denotes the number of times arm $k$ was selected up to round $r$. Letting $\bar{r}_T \leq T$ be the (random) number of rounds used by the algorithm before the $T$-th arm selection, the regret of a round-based algorithm can be upper bounded as follows.

**Proposition 2.1** (Regret of a round-based algorithm)**.**

$$\begin{aligned} \mathcal{R}_T(\mathcal{A}) &= \mathbb{E}\left[\sum_{t=1}^{T}(\mu_\star - \mu_{A_t})\right] \leq \mathbb{E}\left[\sum_{s=1}^{\bar{r}_T}\sum_{k=1}^{K}(\mu_\star - \mu_k)\mathbb{1}(k \in \mathcal{A}_s)\right] \\ &\leq \mathbb{E}\left[\sum_{s=1}^{T}\sum_{k=1}^{K}(\mu_\star - \mu_k)\mathbb{1}(k \in \mathcal{A}_s)\right] = \sum_{k=1}^{K}(\mu_\star - \mu_k)\mathbb{E}\left[N_k(T)\right] . \end{aligned} \tag{2.1}$$

*Hence, upper bounding the number of pulls after $T$ rounds $\mathbb{E}[N_k(T)]$ for each sub-optimal arm provides a regret upper bound (in the usual purely sequential sense).*

Proposition 2.1 shows that adopting a round-based approach where several arms can potentially be pulled per round do not really change the way to start analyzing the algorithms.

**Sub-sampling Duelling Algorithms** SDA takes as input a *sub-sampling algorithm* $\mathrm{SP}(m, n, r)$ that depends on three parameters: two integers $n \geq m$ and a round $r$. A call to $\mathrm{SP}(m, n, r)$ at round $r$ produces a subset of $[n]$ that has size $m$, modeled as a random variable that is further assumed to be independent of the rewards generated from the arms, $(Y_{k,s})_{k\in[K],s\in\mathbb{N}^*}$. We also call *samplers* such sub-sampling algorithms.

In the first round, a SDA selects $\mathcal{A}_1 = [K]$ in order to initialize the history of all arms. Given a sampler SP, we refer to the corresponding SDA as SP-SDA. For any $r \geq 1$, at round $r+1$ SP-SDA first defines the *leader* as one of the arms that have been selected the most in the first $r$ round, namely $\ell(r) \in \mathrm{argmax}_k N_k(r)$. Ties are broken in favor of the arm with the largest mean, and if several arms are still candidate the algorithm chooses one of them at random. The set $\mathcal{A}_{r+1}$ is then initialized to the empty set and $K-1$ *duels* are performed. For each "challenger" arm $k \neq \ell(r)$, a subset $\mathcal{S}_k^r$ of $[N_{\ell(r)}(r)]$ of size $N_k(r)$ is obtained from $\mathrm{SP}(N_k(r), N_{\ell(r)}(r), r)$, and arm $k$ wins the duels if its empirical mean is larger than the empirical mean of the sub-sampled history of the leader. We can write that for any challenger $k$,

$$\bar{Y}_{k,N_k(r)} \geq \bar{Y}_{\ell(r),\mathcal{S}_k^r} \implies \mathcal{A}_{r+1} = \mathcal{A}_{r+1} \cup \{k\}\,.$$

If the leader wins all the duels, that is if $\mathcal{A}_{r+1}$ is still empty after the $K-1$ duels, we set $\mathcal{A}_{r+1} = \{\ell(r)\}$. Arms in $\mathcal{A}_{r+1}$ are then selected by the learner in a random order and are pulled. The pseudo-code of SP-SDA is given in Algorithm 2.1, for an horizon of $T$ rounds.

**Input:** $K$ arms, horizon of $T$ rounds, Sampler SP, leader definition
$\forall k\,, N_k = 1, \mathcal{H}_k = \{Y_{k,1}\}$ ; ▷ Each arm is drawn once
**for** $r \in \{1, \ldots, T\}$ **do**
  $\mathcal{A} = \{\}$, $\ell = \mathrm{leader}((N_k)_{k\in\{1,\ldots,K\}}, (\mathcal{H}_k)_{k\in\{1,\ldots,K\}}, \ell)$ ; ▷ Initialize the round
  **for** $k \neq \ell \in 1, \ldots, K$ **do**
    Draw $\bar{S}_k^r \sim \mathrm{SP}(N_k, N_\ell, r)$ ; ▷ Sub-sample of $\ell$ used for the duel with $k$
    **if** $\bar{Y}_{k,N_k} > \bar{Y}_{\ell,\bar{S}_k^r}$ **then**
      $\mathcal{A} = \mathcal{A} \cup \{k\}$ ; ▷ $k$ added to $\mathcal{A}$ if it wins the duel against $\ell$
    **end**
  **end**
  **if** $|\mathcal{A}| = 0$ **then**
    $\mathcal{A} = \{\ell\}$ ; ▷ If no challenger in $\mathcal{A}$ then $\ell$ is pulled
  **end**
  **for** $a \in \mathcal{A}$ **do**
    Pull arm $a$, observe reward $Y_{a,N_a+1}$
    $N_a = N_a + 1$, $\mathcal{H}_a = \mathcal{H}_a \cup \{Y_{a,N_a}\}$ ; ▷ Update number of pulls and history
  **end**
**end**

**Algorithm 2.1:** Generic SP-SDA

To properly define the random variable $\bar{S}_k^r$ used in the algorithm, we introduce the following probabilistic modeling: for each round $r$, each arm $k$, we define a family $(S_k^r(m,n))_{m\leq n}$ of independent random variables such that $S_k^r(m,n) \sim \mathrm{SP}(m,n,r)$. In words, $S_k^r(m,n)$ is the subset of the leader history used should arm $k$ be a challenger drawn $m$ times up to round $r$ dueling against a leader that has been drawn $n$ times. With this notation, for each arm $k \neq \ell(r)$ one has $\bar{S}_k^r = S_k^r\left(N_k(r), N_{\ell(r)}(r)\right)$. We recall that in the SDA framework, we require those random variables to be independent from the reward streams $(Y_{k,s})$ of all arms $k$. We call such sub-sampling algorithms *independent samplers*.

**Definition 2.2** (Independent Sampler). *A independent sampler satisfies*

$$\forall (k,k') \in [K], \forall r, m \geq n : \; S_k^r(m,n) \perp\!\!\!\perp (Y_{k',s})_{s\in\mathbb{N}} \; .$$

**Some instances of SDA** We now present some sub-sampling algorithms that we believe are interesting to use within the SDA framework. Intuitively, these algorithms should ensure enough *diversity* in the output subsets when called in different rounds, so that the leader cannot always look good, and challengers may win and be explored from time to time. The most intuitive candidates are random samplers like *Sampling Without Replacement* (WR) and *Random Block Sampling* (RB). But we also propose two deterministic sub-sampling algorithms: *Last Block Sampling* (LB), and *Low Discrepancy Sampling* (LDS) that uses a predefined low discrepancy sequence $(u_r)_{r\in\mathbb{N}}$ (Drmota and Tichy, 1997; Halton, 1964; Sobol, 1967). We summarize these sub-sampling algorithms in the following, considering two sample sizes $m \leq n$ and a round $r$:

- *Sampling Without Replacement (WR)* returns a subset of size $m$ selected uniformly at random without replacement in $[n]$ (i.e each element can be drawn only once).
- *Random Block Sampling (RB)* draws an integer $n_0$ uniformly at random in $[n-m]$ and returns $\{n_0, ..., n_0+m-1\}$. This is faster than WR as only one element is drawn.
- *Last Block Sampling (LB)* simply returns the last observations $\{n-m+1, \ldots, n\}$.
- *Low Discrepancy Sampling (LDS)* is a deterministic version of Random Block Sampling where $n_0$ is equal to $u_r$ at round $r$, for some low discrepancy sequence $(u_r)$.

We propose in Figure 2.1 below an example of sub-samples that could be returned for a given duel by several of the algorithms that we propose.

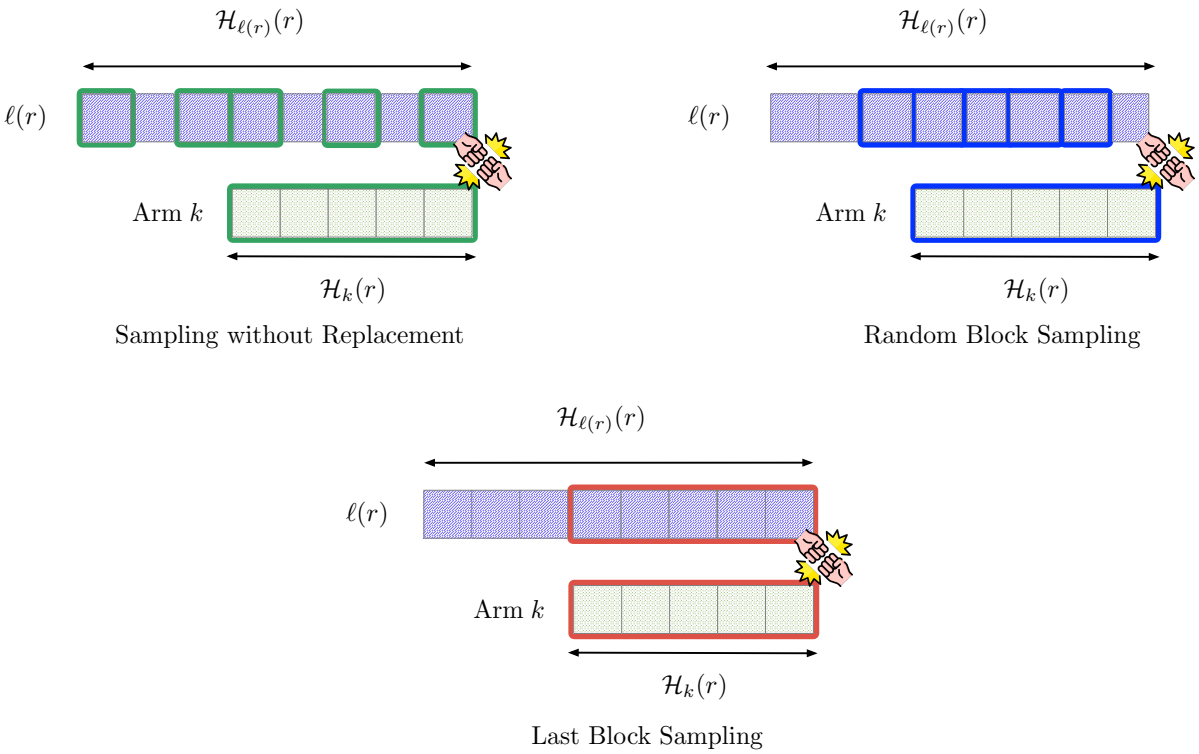


**Figure 2.1** – Illustration of a duel step for a few sub-sampling algorithms. Each box represent an observation, and in each figure the framed box are the observations selected by the sampler. For each arm $i \in \{\ell(r), k\}$, $\mathcal{H}_i$ denotes the history available for $i$ at round $r$

We proposed WR and RB samplers in hope that their randomness will naturally introduce sufficient diversity in the sub-samples. Similarly, LDS should ensure by construction the exploration of different part of the history of the leader during successive rounds. On the other hand, the reason why Last Block Sampling should work in producing diverse samples is less intuitive: after a sufficient number of rounds we expect the leader to be pulled most of the time, making the sub-sample move almost at each round in a sliding window fashion. We will prove later in this chapter that this simple mechanism is actually enough to obtain nice theoretical guarantees. Finally, we can remark that except the WR sampler most of the proposed algorithms return a set of consecutive observations, which is generally faster from a computational perspective. We call such algorithm a *block sampler*.

**Definition 2.3** (Block Sampler). *A Block Sampler is a sub-sampling algorithm that only returns sub-sets of observations that have been collected successively in the history of an arm. Given* $(m, n) \in \mathbb{N}^2$, $S_k^r(m, n) = \{j, j+1, \dots, j+m-1\}$ *for some* $j \leq n - m$.

In section 2.3 we will show that Block Samplers have convenient properties for the regret analysis of their associated SDA algorithm.

**Links with existing algorithms** The BESA algorithm (Baransi et al., 2014) with $K = 2$ coincides with WR-SDA. However beyond $K > 2$, Baransi et al. (2014) rather suggest a tournament approach, without giving a regret analysis. WR-SDA can therefore be seen as an alternative generalization of BESA beyond 2 arms, that may perform better than the tournament, as can be seen in experiments of Section 2.5. While the structure of SDA is close to that of SSMC (Chan, 2020), SSMC is not a SP-SDA algorithm. Indeed, its sub-sampling algorithm heavily relies on the rewards and is therefore not an independent sampler. It actually outputs the set $\mathcal{S} = \{n_0 + 1, \dots, n_0 + n\}$ for which $\bar{Y}_{\ell(r),\mathcal{S}}$ is the smallest, which is hence the worst possible sub-sample that can be returned by block samplers. For this reason SSMC may be more conservative than SDA algorithms, that may able to "forget" faster a sequence of bad observations that could have been drawn successively (or at least using them less often). Furthermore, our intuition is that LB-SDA could actually been seen as a simpler and possibly faster version of SSMC, as the two seem to perform very comparably in practice. Indeed, when an arm has been leader for some time the challengers can be pulled only if the newly arriving data allows to get a new sub-sample that is worse than the previous "worst sub-sample". Otherwise the comparison used by SSMC is the same as the one used in the previous round, where the challenger lost. Hence, the SDA framework is somehow unifying the ideas of BESA and SSMC as two of its instances are close, and maybe improved, variants of these algorithms.

**On the use of forced exploration** In (Chan, 2020), SSMC additionally implements some *forced exploration*: at a round $r$, each arm $k$ such that $N_k(r)$ is smaller than some value $f_r$ is added to $\mathcal{A}_{r+1}$. SSMC is proved to be asymptotically optimal for SPEF provided that $f_r = o(\log r)$ and $\log \log r = o(f_r)$. In the next sections, we show that SDA does not need forced exploration for some family of distributions: Bernoulli, Gaussian and Poisson. However, we will show that forced exploration allows to obtain more general theoretical guarantees and is not harmful in practice. Indeed, the exploration in $\sqrt{\log(r)}$ suggested by Chan (2020) only forces to draw each arm $4$ times for a time horizon $T = 10^7$.

Finally, it is interesting to mention that duel-based algorithms were an early alternative to index policies. As already mentioned by Chan (2020), the very first asymptotically optimal algorithm proposed in (Lai and Robbins, 1985) already performed pairwise comparison between the empirical mean of a leader (defined therein as the empirical best arm) and an upper confidence bound of the challenger, $\mathrm{UCB}_k(t)$, thus helping the challengers instead of disadvantaging the leader.

## 2.3 Generic Regret Analysis

In this section we present our core contributions for this chapter, which are the theoretical guarantees for the SDA family of algorithms. We proceed step by step by sequentially including new assumptions on the family of distributions and the sub-sampling algorithm, and analyzing what can be obtained at each step. These guarantees come by deriving upper bounds on the expected number of selections of each sub-optimal arm $k$, $\mathbb{E}\left[N_k(T)\right]$, which directly yields an upper bound on the regret via Equation (2.1). To ease the presentation, we assume that there is a unique optimal arm[1], that we define as arm $1$ for simplicity.

### 2.3.1 Concentration of means and sub-sample means

A first natural question is to consider whether computing empirical means is a good idea or not for the family of distributions we consider. This is the case only if these empirical estimators *concentrate* around the true means at a sufficient speed, that we formalize as follow.

**Assumption 2.4** (Concentration of empirical means)**.** *For each arm $k$, the distributions $\nu_k$ (of mean $\mu_k$) admits a good rate function $I_k : \mathbb{R} \mapsto \mathbb{R}^+$, that is*

$$\forall x > \mu_k,\ \mathbb{P}\left(\bar{Y}_{k,n} \geq x\right) \leq e^{-nI_k(x)}\ ,$$
$$\forall x < \mu_k,\ \mathbb{P}\left(\bar{Y}_{k,n} \leq x\right) \leq e^{-nI_k(x)}\ ,$$

*where $I_k$ is continuous and $I_k(x) = 0$ if and only if $x = \mu_k$.*

This assumption is satisfied by many usual families of distributions (SPEF, bounded, ...). More generally it is satisfied by any *light-tailed distribution*, that we define as any distribution $\nu$ for which

$$\exists \lambda_0 > 0\ :\ \forall |\lambda| < \lambda_0\ , \mathbb{E}_{X\sim\nu}[\exp(\lambda X)] < +\infty\ .$$

If Assumption 2.4 is satisfied it makes sense to compare sub-sample means, and we can then start considering their concentration. Considering two arms $\ell$ and $k$ and some $x < \mu_\ell$, we want to upper bound the probability $\mathbb{P}(\bar{Y}_{\ell,S_k^r} \leq x, N_k(r) \geq n_0)$ at some round $r$ and for some integer $n_0$. We now exhibit a convenient property of Block Samplers, introduced in definition 2.3.

[1] as can be seen in the analysis of SSMC Chan (2020), treating the general case only requires some additional notation.

**Lemma 2.5** (Concentration of sub-samples with Block Samplers)**.** *Let SP be a Block Sampler. Consider a round $r$ and $n_0 \in \mathbb{N}$, then it holds that*

$$\forall x > \mu_k,\ \mathbb{P}\left(\bar{Y}_{k,S_k^r(N_k(r),N_\ell(r))} \geq x, N_k(r) \geq n_0\right) \leq r^2 \sum_{n=n_0}^{r} \mathbb{P}\left(\bar{Y}_{k,n} \geq x\right) ,$$
$$\forall x < \mu_k,\ \mathbb{P}\left(\bar{Y}_{k,S_k^r(N_k(r),N_\ell(r))} \leq x, N_k(r) \geq n_0\right) \leq r^2 \sum_{n=n_0}^{r} \mathbb{P}\left(\bar{Y}_{k,n} \leq x\right) ,$$

*Proof.* Consider the first inequality. We use a union bound on the possible values of $N_\ell(r), N_k(r)$ and of the first element of the block, which provides the result by further remarking that each block of observation satisfy the same concentration properties as the first block.

$$\begin{aligned}\mathbb{P}\left(\bar{Y}_{k,S_k^r(N_k(r),N_\ell(r))} \geq x, N_k(r) \geq n_0\right) &\leq \sum_{n_\ell=1}^{r}\sum_{n_k=1}^{r}\sum_{i=1}^{r} \mathbb{P}\left(\bar{Y}_{k,i:i+n_k-1} \geq x\right) \\ &\leq r^2 \sum_{n_k=n_0}^{r} \mathbb{P}\left(\bar{Y}_{k,n_k} \geq x\right) .\end{aligned}$$

The exact same steps with $x \leq \mu_\ell$ gives the second inequality. □

To obtain this result we used that the number of blocks is at most linear in $r$. For a sampler that is not using blocks this number can be much larger. For instance, with sampling without replacement we would obtain $\binom{n_\ell}{n_k}$ sub-samples for each couple $(n_k, n_\ell)$, making the upper bound vacuous. Hence, samplers that do not use blocks require novel concentration tools. For that reason, the results presented in the rest of this chapter are only valid for Block Slampers.

Unfortunately, the upper bound presented in lemma 2.5 is not sufficient in some parts of the upcoming proofs. However, it can be improved under the additional event that the arm $k$ by summing on the consecutive rounds $s = 1, \dots, r$.

**Lemma 2.6** (Concentration of sub-samples inside a trajectory of a bandit algorithm)**.** *Let SP be a Block Sampler. Consider a round $r$ and $n_0 \in \mathbb{N}$, then it holds that*

$$\forall x > \mu_k,\ \sum_{s=1}^{r} \mathbb{P}\left(\bar{Y}_{k,S_k^s(N_k(s),N_\ell(s))} \geq x, N_k(s) \geq n_0, k \in \mathcal{A}_{s+1}\right) \leq r \sum_{n=n_0}^{r} \mathbb{P}\left(\bar{Y}_{k,n} \geq x\right) ,$$
$$\forall x < \mu_k,\ \sum_{s=1}^{r} \mathbb{P}\left(\bar{Y}_{k,S_k^s(N_k(s),N_\ell(s))} \leq x, N_k(s) \geq n_0, k \in \mathcal{A}_{s+1}\right) \leq r \sum_{n=n_0}^{r} \mathbb{P}\left(\bar{Y}_{k,n} \leq x\right) ,$$

The obtained bound is surprisingly better than when considering an individual sub-sample.

*Proof.* We first write that

$$\begin{aligned}(\mathrm{A}) \coloneqq &\sum_{s=1}^{r} \mathbb{P}\left(\bar{Y}_{k,S_k^s(N_k(s),N_\ell(s))} \geq x, N_k(s) \geq n_0, k \in \mathcal{A}_{s+1}\right) \\ =&\mathbb{E}\left[\sum_{s=1}^{r} \mathbb{1}\left(\bar{Y}_{k,S_k^s(N_k(s),N_\ell(s))} \geq x, N_k(s) \geq n_0, k \in \mathcal{A}_{s+1}\right)\right] .\end{aligned}$$

Now let's consider the sum inside of the expectation. It represents the count of the number of times a block of $\ell$ of size $N_k(r)$ has led to arm $k$ being pulled. The rest of the proof is based on the remark that each possible block of observations (for any size and any location in the history of $\ell$) can only have a cost of 1 in this sum: if $k$ wins the duel its number of pulls will increase. Hence, the sum is upper bounded by the total number of blocks with a sub-sample mean satisfying the inequality, and we directly have

$$\begin{aligned}(\mathrm{A}) \leq&\mathbb{E}\left[\sum_{s=1}^{r}\sum_{n_k=n_0}^{r} \mathbb{1}(\bar{Y}_{k,s:s+n_k-1} \geq x)\right] \\ \leq&r\sum_{n_k=n_0}^{r} \mathbb{P}\left(\bar{Y}_{k,n_k} \geq x\right)\end{aligned}$$

□

The results presented so far are sufficient to obtain a first upper bound on the number of pulls of each sub-optimal arm, for any SDA equipped with a block sampler.

**Lemma 2.7** (First upper bound)**.** *Consider a block sampler SP and a bandit $\nu = (\nu_1, \dots, \nu_k) \in \mathcal{F}^K$ where $\mathcal{F}$ is a family of distributions satisfying Assumption 2.4. Then, for any suboptimal arm $k \neq 1$, the expected number of pulls of each sub-optimal arm $k$ under SP-SDA is upper bounded by*

$$\mathbb{E}[N_k(T)] \leq \frac{1+\varepsilon}{I_1(\mu_k)}\log(T) + C_k(\nu,\varepsilon) + 9\sum_{r=1}^{T}\mathbb{P}\left(N_1(r) \leq C_1\log(r)\right) ,$$

*where $C_k(\nu,\varepsilon)$ and $C_1$ are both problem-dependent constants.*

This result is interesting because it shows that Assumption 2.4 along with elementary properties of block samplers are enough to exhibit the logarithmic term that we hope to be dominant in the regret upper bound.

The proof is inspired from the analysis of SSMC (Chan, 2020), using the concentration inequality of Lemma 2.6. The details can be found in Appendix D of (Baudry et al., 2020). We

refer the interested reader to Chapter 3 where we provide the proof of this result for a variant of LB-SDA using a limited memory size, that is valid for Lemma 2.7 by setting the limit to $+\infty$.

The rest of the analysis consists in further upper bounding the quantity $\mathbb{P}(N_1(r) \leq C_1 \log(r))$ for a given round $r$. This probability corresponds to the case where arm 1 have been under-explored. In the next part we examine the conditions that should be met to make this probability small under an instance of SDA.

### 2.3.2 Avoiding under-exploration of the best arm

It is clear that whether the best arm is sure to be explored or not depends jointly on the sub-sampling algorithm and the properties of arms' distributions. First, we remark that if the number of pulls of arm 1 stays below $C_1 \log(r)$ then it has necessarily lost a lot of duels. We will further prove that it has even lost a large number of duels against the *same leader*, and while being stuck with *some fixed sample* the entire time. The question is then: has it been given a fair chance and lost a lot of *diverse* duels, or has it been provided too few different duels? We introduce the definition of *non-overlapping* sub-samples as a way to formalize this question.

**Definition 2.8** (Non-overlapping sets). *Two sets of integers $\mathcal{I}_1 \subset \mathbb{N}$ and $\mathcal{I}_2 \subset \mathbb{N}$ are non-overlapping if $\mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset$. Furthermore, the sets in $\mathcal{J} = (\mathcal{I}_1, \mathcal{I}_2, \dots, \mathcal{I}_M)$ are said to be* pairwise *non-overlapping if for any pair $(\mathcal{I}_k, \mathcal{I}_{k'}) \in \mathcal{J}^2$ it holds that $\mathcal{I}_k \cap \mathcal{I}_{k'} = \emptyset$.*

Definition 2.8 allows to quantify more precisely what kind of diversity we care about: we want sub-sampling algorithms to provide enough *pairwise non-overlapping* sub-samples during a run of SDA. Note that we use the expressions non overlapping sets or non-overlapping sub-samples indifferently due to the isometry between a sub-sample and the positions of its items in the history of the arms.

Let us now consider the case for which the sampler was able to provide a large number of pairwise non-overlapping sub-samples. If arm 1 has not been drawn, its sample mean was smaller than the mean of *each* sub-sample from the leader. If the samples were non-overlapping, this event can be written using the *balance function*, introduced by Baransi et al. (2014) and that we recall here.

**Definition 2.9** (Balance function). *Letting $\nu_{k,j}$ denote the distribution of the sum of $j$ independent variables drawn from $\nu_k$, and $F_{\nu_{k,j}}$ its corresponding CDF, the balance function of arm $k$ is*

$$\alpha_k(M, j) = \mathbb{E}_{X \sim \nu_{1,j}} \left( \left( 1 - F_{\nu_{k,j}}(X) \right)^M \right).$$

For a fixed $M$ and $j$, $\alpha_k(M, j)$ corresponds exactly to the probability that arm 1 loses $M$ independent duels with its $j$ first observations.

Then, using the result of Lemma 2.7, Definition 2.8 and Definition 2.9 we can formulate the following definition of *sufficient diversity* for a given sampler SP.

**Definition 2.10** (Sufficient Diversity)**.** *Consider a sampler SP, a forced exploration $f_r$, and a bandit problem $\nu = (\nu_1, \dots, \nu_K)$. We say that SP introduces sufficient diversity in the sub-samples if for any round $r$ large enough SP-SDA satisfies*

$$\mathbb{P}\left(N_1(r) \leq C_1 \log(r)\right) = \sum_{k=2}^{K} \sum_{j=f_r}^{r} \alpha_k(M_{j,r}, j) + o(r^{-2}) \ ,$$

*for some sequence $M_{j,r}$ that is linear in $r$ up to poly-log terms.*

*In other words, SP provides enough pairwise non-overlapping sub-samples so that if arm 1 is not pulled enough it is (with high probability) because it has lost a large number of diverse duels.*

An interesting remark from this definition is that, if SP provides sufficient diversity, then the guarantees of SP-SDA depends only on the family of distributions of the arms through their balance function. Hence, we disentangled in the analysis the impact of the sampler and of the family of distributions in the performance of an instance of SDA.

Furthermore, Definition 2.10 allows to naturally establish the condition we want to check for a family of distributions to determine whether SP-SDA can work or not for this family.

**Definition 2.11** (Balance condition)**.** *The balance condition is respected for a family $\mathcal{F}$ and a forced exploration $f_r$ if for any sequence $M_{j,r}$ satisfying $M_{j,r} = \Omega\left(\frac{r}{(\log r)^a}\right)$ for a large enough it holds that*

$$\sum_{r=1}^{T} \sum_{j=f_r}^{C_1 \log(r)} \alpha_k\left(M_{j,r}\right) = o(\log(T)) \ .$$

Note that the level of forced exploration at round $r$ is included in these definitions, hence some distributions may satisfy the balance condition for some values of $f_r$ but not for others. We can now provide the main theorem of this chapter, that is using Lemma 2.7 and all the definitions presented in this section. The statement remains as general as possible, and we refer to the next section for its application to specific instances of SDA and families of distributions.

**Theorem 2.12** (Logarithmic Regret for SP-SDA)**.** *Consider a bandit $\nu = (\nu_1, \dots, \nu_K) \in \mathcal{F}^K$, where $\mathcal{F}$ is a family of distributions satisfying Assumption 2.4 (concentration of the means). Let SP be a **block sampler** providing **sufficient diversity**. Further assume that all the arms drawn from $\mathcal{F}$ satisfy the **balance condition**. Then for any $\varepsilon > 0$ the regret of SP-SDA is upper bounded by*

$$\mathcal{R}_\nu(T, \pi^{\text{SP-SDA}}) \leq \sum_{k=2}^{K} \frac{1+\varepsilon}{I_k(\mu_1)} \log(T) + o_\varepsilon(\log T) \ .$$

*Proof.* This result is a direct combination of Lemma 2.7, and the definition of sufficient diversity and the balance condition. □

The challenge is now to determine which samplers and family of distributions satisfy these properties. However, these two investigations can be made separately. Before that, to provide the reader further intuitions let us consider some simple examples of failure for SDA.

*Example 1 (bad sampler):* Consider a sampler returning the *first* successive observations of the leader. Then, the regret of SDA is linear.

↪ The sampler fails to provide enough diversity by returning always the same sub-sample.

*Example 2 (distributions unadapted to sub-sampling):* Consider an interval $[a, b] \subset \mathbb{R}$, and a two-armed bandit where arm 1 is the best arm, $\mathbb{P}_{\nu_1}([a, b)) = p$, and $\nu_2$ is supported on $(b, +\infty)$. Then, the regret of SDA is at least in $\Omega(p^{f_T} T)$: if the first $f_T$ observations from $\nu_1$ are in $[a, b)$ the arm is never pulled outside of forced exploration.

↪ The worst rewards are more likely for the best arm than for the other, in that case the sub-sampling is not enough to recover from them.

In the following we formalize some sufficient conditions to avoid these cases.

## 2.4 Theoretical guarantees for RB-SDA and LB-SDA

In this part we instantiate the results from previous section for some instances of samplers and some families of distributions. We will first prove that both *Random Block* and *Last Block* samplers ensures sufficient diversity. Then we will provide a general technique to upper bound balance functions and some examples of distributions satisfying the balance condition.

### 2.4.1 Analysis of LB and RB samplers

In this section we further analyze the Last Block (LB) and Random Block (RB) samplers. Both are Block Samplers, as defined in 2.3, but the way they provide diverse sub-samples is very different: while RB explores a random part of the history at each step, LB simply returns the last observations. Our striking result is that both methods manage to ensure sufficient diversity.

**Lemma 2.13** (Random Block Sampler)**.** *The Random Block sub-sampling algorithm ensures sufficient diversity, as defined in 2.10.*

In the following we sketch the main results that allow to prove Lemma 2.13. We introduce the random variable $X_{m,H,j}$ that denotes the maximal number of pairwise non-overlapping subsets obtained in $m$ i.i.d. samples from $\mathrm{RB}(j, H)$. We aim at upper bounding

$$\mathbb{P}\left(X_{m,H,j} \leq x\right)$$

for some values of $m$, $H$, $j$, and some specific value of $x$ that will be specified later. Intuitively, this probability is decreasing in $H$ and $m$, and increasing in $j$. Furthermore, if $m$ is large compared to $H$ we can reasonably expect to visit most of the history. Then, the proof of Lemma 2.13 mainly relies on the results we propose in Lemma 2.14, that are written with $m = H$. This is sufficient in our proof since during a run of RB-SDA we expect both the history size of the leader and the number of duels played to be linear in $r$, so we can set the two values equal to their minimum using that $X_{a,b,j} \geq X_{a,a,j}$ if $a \leq b$.

**Lemma 2.14** (Maximal number of non-overlapping subsets with RB-SDA)**.** *Consider a history size $H$, a number of draws that is also $H$, a sub-sample size $j$, and some $\alpha \in (0, 1)$. Then, it holds that*

$$\mathbb{P}\left(X_{H,H,j} \leq \left\lceil \frac{\alpha H}{j} \right\rceil\right) \leq \mathbb{P}\left(X_{H,H,1} \leq \alpha H\right) .$$

*Furthermore, if $\alpha$ is small enough it holds that*

$$-\log\left(\mathbb{P}\left(X_{H,H,1} \leq \alpha H\right)\right) = \mathcal{O}(H) .$$

*Proof.* The first part of Lemma 2.14 consists in showing that it is sufficient to study $X_{H,H,1}$ to get a result for any sub-sample size $j$. This result comes from block sampling: we can focus on the number of unique elements drawn by RB (starting observation for each block). If $M$ such unique elements are available, then in the worst case these observations are successive in the

history, and we have to skip $j$ blocks for 1 observation to avoid overlaps. This result directly yields the first statement of the lemma.

We now consider $X_{H,H,1}$. In this case the distribution of the points collected is the one of sampling with replacement in a set of size $H$. It has been well studied in the literature, that allows to obtain the following upper bound.

**Lemma 2.15.** *Combining the exact expression of $\mathbb{P}(X_{m,H,1} = k)$ from (Mendelson et al., 2016) and an upper bound on Stirling numbers of the second kind from (Rennie and Dobson, 1969), we obtain that for any $k \leq H$*

$$\mathbb{P}(X_{H,H,1} = k) \leq \frac{1}{2}\left(\frac{k}{H}\right)^{H-k}\binom{H}{k}.$$

As $k$ should be small compared to $H$ it is then natural to use $\binom{H}{k} \leq \frac{H^k}{k!}$. We then bound $1/k!$ by its equivalent given by Stirling formula and introduce a multiplicative constant $c$ to pay for the approximation,

$$\binom{H}{k} \leq c\frac{H^k}{\sqrt{2\pi k} \times k^k}e^k.$$

Refactoring and assuming that $k \leq H - 2k$ we obtain

$$\mathbb{P}(X_{H,H,1} = k) \leq \frac{c}{2}\frac{e^k}{\sqrt{2\pi k}}\left(\frac{k}{H}\right)^{H-2k} \leq \frac{c}{2\sqrt{2\pi}}\left(\frac{ke}{H}\right)^{H-2k}.$$

To obtain the upper bound of Lemma 2.14 we need to consider this for $k \leq \alpha H$. We assume that $\alpha < 1/e$, so we can finally obtain

$$\begin{aligned}
\mathbb{P}(X_{H,H,1} \leq \alpha H) &\leq \frac{c}{2\sqrt{2\pi}}\sum_{k=0}^{\lfloor \alpha H \rfloor}(\alpha e)^{H-2k} \\
&= \frac{c}{2\sqrt{2\pi}}(\alpha e)^H\frac{(\alpha e)^{-2\lfloor \alpha H \rfloor} - 1}{(\alpha e)^{-2} - 1} \\
&= \mathcal{O}\left((\alpha e)^{H(1-2\alpha)}\right),
\end{aligned}$$

which is exponentially decreasing in $H$ for small enough $\alpha$, and hence gives the result. □

The complete proof of Lemma 2.13, that relies on Lemma 2.14, can be found in Appendix 2.6. It remains to prove that if arm 1 is not pulled then it has been stuck with its first $j$ observations for an interval of size $\Omega(r/\log(r))$ in which the leader already had at least $\Omega(r/\log(r))$ observations (corresponding to $m$ and $H$ here).

Hence, the proof of sufficient diversity with Random Block relied on probabilistic arguments due to the randomized sampling. For Last Block, on the contrary, all arguments will be purely deterministic.

**Lemma 2.16** (Last Block Sampler). *LB-SDA ensures sufficient diversity in the sub-samples.*

*Proof.* The main trick to prove this result is to show that for any trajectory of the bandit algorithm the leader (whatever arm it is) will be played *many times*, making sure that the sub-samples presented to the challengers will be diverse. To measure this, we define the number of rounds for which the *current* leader was pulled up to a round $r$,

$$W_r = 1 + \sum_{s=1}^{r-1} \mathbb{1}(\mathcal{A}_{s+1} = \{\ell(s)\}) \ ,$$

where we added 1 for the first round where every arm (including the arbitrary leader) is pulled once. For any trajectory, we show that this quantity is equal to the number of pulls of the leader, that we know to be linear in $r$.

**Proposition 2.17.** *For any round $r \geq 2$, under LB-SDA it holds that*

$$W_r = N_{\ell(r)}(r) \geq r/K \ .$$

This proposition, formally proved in (Baudry et al., 2021b), comes from considering each phase between changes in leadership, and that any time the leader gets ahead in terms of number of pulls this arm must be drawn *while being leader*. We want to use this results on the rounds for which arm 1 has already "enough" samples thanks to the sampling obligation. We refer to Appendix 2.6.1, where we prove that if $f_r = (\log r)^{1/k}$ for $k > 1$ there exists a round $a_r = o(r)$ for which $N_1(r) \geq f_r - 1$. For the rest of the proof we consider the number of duels lost by arm 1 after $a_r$ against unique sub-samples of a sub-optimal leader. The number of duels won by the leader between the rounds $a_r$ and $r$ is equal to $W_r - W_{a_r}$. Out of those duels, at most $C_1 \log(r)$ can be won by the optimal arm. Consequently, there is at least $W_r - W_{a_r} - C_1 \log(r)$ duels won by a suboptimal leader between rounds $a_r$ and $r$. Using Lemma 2.17 and $W_{a_r} \leq a_r$

we can then prove that for any $\beta \in (0,1)$, there exists a round $r_{\beta,K}$ from which

$$W_r - W_{a_r} - C_1 \log(r) \geq \beta \frac{r}{K} \,. \tag{2.2}$$

Under $N_1(r) \leq C_1 \log(r)$ we are sure that there exists some $j \in \{1, ..., \lfloor C_1 \log(r) \rfloor\}$ such that a fraction $1/(C_1 \log(r))$ of the duels counted above have been played with $N_1(r) = j$. Let us denote $\widetilde{W}_r = W_r - W_{a_r} - C_1 \log(r)$ and show this by contradiction. Out of those duels, we denote $\widetilde{W}_{r,j}$ the number of duels played with $N_1(r) = j$. If we assume that for all $j \leq \lfloor C_1 \log(r) \rfloor$, there is strictly less than $\frac{\beta}{C_1 \log(r)} \frac{r}{K}$ duels played with $N_1(r) = j$. The following would hold,

$$\widetilde{W}_r \coloneqq W_r - W_{a_r} - C_1 \log(r) = \sum_{j=1}^{\lfloor C_1 \log(r) \rfloor} \widetilde{W}_{r,j} < \sum_{j=1}^{\lfloor C_1 \log(r) \rfloor} \frac{\beta}{C_1 \log(r)} \frac{r}{K} < \beta \frac{r}{K} \,.$$

There is a contradiction with Equation (2.2) and so there is necessarily some $j \leq \lfloor C_1 \log(r) \rfloor$ and $\beta r/(C_1 \log(r) K)$ duels such that arm 1 competes using exactly its first $j$ samples.

Furthermore, with the same argument we are sure that a fraction $1/(K-1)$ of these duels is played against the same leader $k \in \{2, \dots, K\}$. We would now like to obtain duels with non-overlapping blocks. Even if the blocks are all consecutive, waiting for $j$ steps is enough to ensure that they are not overlapping. Hence, taking a fraction $1/j$ of the duels from the previous subsets is a conservative estimate of the true number of non-overlapping blocks they contain. Finally, we conclude that for any $\beta \in (0,1)$ there exists a constant $r_{\beta,K}$ such that for any round $r > r_{\beta,K}$, under the event $\{N_1(r) \leq C_1 \log r\}$ there exists some $k \in \{2, \dots, K\}$ and some $j \in \{\lfloor f_r - 1 \rfloor, \lfloor C_1 \log r \rfloor\}$ such as arm 1 lost at least $\beta \frac{r}{K(K-1)(C_1 \log r) j}$ duels against non-overlapping blocks of arm $k$ while $k$ is the leader and 1 has exactly $j$ observations. This term correspond exactly to the balance function $\alpha_k(M, j)$ from Definition 2.9, with $M = \beta \frac{r}{K(K-1)(C_1 \log(r)) j}$. This concludes the proof, as we obtain that LB-SDA satisfies

$$\mathbb{P}(N_1(r) \leq C_1 \log(r)) \leq r_{\beta,K} + \sum_{k=2}^{K} \sum_{j=1}^{C_1 \log(r)} \alpha_k \left( \beta \frac{r}{K(K-1)(C_1 \log r) j}, j \right) \,,$$

which satisfies Definition 2.10 of *sufficient diversity*. □

To conclude this section on sub-sampling algorithms we explain why the tournament structure of BESA (Baransi et al., 2014) may not work due to problems related to the diversity of sub-samples drawn. With the following example, we explain that the tournament (for $K \geq 4$) may fail because an under-sampled best arm will be required to win against other arms that have not been sampled a lot too, potentially repeatedly playing the same unsuccessful duels.

**Example 2.18** (Failure of the tournament structure of Baransi et al. (2014)). *Let us consider a 4-armed bandit with $\mu_1 > \mu_2 > \mu_3 = \mu_4$. Assume that $N_1(l) = C\log(r)$ for some $0 < C < C_1$ (the constant in Lemma 2.7), and that $\bar{Y}_{1,j} \leq \mu_4 - \varepsilon$ for some $\varepsilon > 0$. If all other arms are well estimated ($|\bar{Y}_{k,N_k(r)} - \mu_k| \leq \varepsilon$ for $k \geq 2$), then 2 should be sampled most of the time (be the leader) and 3 and 4 have $\mathcal{O}(\log(r))$ samples. Further assume that 3 and 4 have roughly $C\log(r)$ samples too. Then, the best arm may sometimes win a duel against arm 2 thanks to diversity of sub-samples and Assumption 2.20. However, after that it will most likely lose its duel against arm 3 or 4, and hence take an unreasonably long time to be pulled because of its duels against other challenger arms.*

### 2.4.2 A deeper look at the balance condition

In this section we consider properties of families of distributions, that are independent of the run of SDA and the sampler it uses. We first exhibit a general property that ensures that the balance condition is satisfied with some amount of forced exploration, and then show that several families including SPEF actually satisfy it. We finally prove that several families of distributions (Bernoulli, Gaussian, Poisson) satisfy the balance condition without forced exploration, and on the other hand that some amount of forced exploration is necessary for exponential distributions.

We first propose a general way to upper bound the balance function. We use the notation $F_{k,j}$ for the cdf of the distribution of the sum of $j$ i.i.d samples drawn from any arm $k$, and $G_{k,j} = 1 - F_{k,j}$. Then, the intuition is rather simple: take a reference value $u$, if arm 1 lost the $M$ duels then either (1) the sum of all the samples collected from arm 1 is smaller than $u$, or (2) the sum of each of the $M$ sub-samples drawn from the history of arm $k$ is larger than $u$.

**Proposition 2.19.**

$$\alpha_k(M, j) \leq F_{1,j}(u) + G_{k,j}(u)^M \ , \ \forall u \in \mathbb{R} \ . \tag{2.3}$$

The idea is then to consider values of $u$ of the form $u = G_{2,j}^{-1}(1 - \frac{(\log M)^a}{M})$, that allows to upper bound the second term by $M^{-a}$. We then need to control $F_{1,j}\left(F_{2,j}^{-1}\left(\frac{(\log M)^a}{M}\right)\right)$. Interestingly, only having $F_{1,j}(x) \leq F_{2,j}(x)$ for $x$ small enough is not sufficient as $\frac{\log(M)^a}{M}$ is not small enough for the balance condition to hold. Interestingly, it is however enough to ensure a sub-linear regret, more precisely a poly-logarithmic regret. In the following we show that a slightly stronger condition is sufficient to ensure a logarithmic regret.

### 2.4.3 A general family satisfying the balance condition

In this section we will prove that assuming the best distribution to have a "better" lower tail than the others is enough to prove the balance condition. We formalize this as follow.

**Assumption 2.20** (Dominant left tail)**.**

$$\forall k \geq 2 \ : \ \exists y_k \in \mathbb{R}, \ c_k \in (0,1) : \forall x \leq y_k \ , \ \frac{\mathrm{d}\mathbb{P}_{\nu_1}}{\mathrm{d}\mathbb{P}_{\nu_k}}(x) \leq c_k \ .$$

Under this assumption the worst rewards have to be more likely for bad arms than for good arms, to avoid the best arm getting stuck for a long time with a "bad" sample. The parameter $c \leq (0,1)$ is added to ensure that the two tails are distinguishable enough for our proof to work. In the following we discuss this condition and show that it is satisfied by many families of distributions. We now prove that it is enough to make the balance condition hold, thanks to the following result.

**Proposition 2.21.** *If assumption 2.20 is satisfied with some parameters $(c_k, y_k)_{k\geq 2}$, then for all $k \geq 2, j \geq 1$ it holds that for all $x \leq y_k$:*

$$F_{1,j}(x) \leq c_k^j F_{k,j}(x) \ .$$

*Proof.* For any $x \leq y_k$ The computation of $F_{1,j}(x)$ directly provides the result,

$$\begin{aligned} F_{1,j}(x) &= \int_{-\infty}^{x} \int_{-\infty}^{x-x_1} \cdots \int_{-\infty}^{x-x_1-\cdots-x_{j-1}} \mathrm{d}\mathbb{P}_{\nu_1}(x_1) \ldots \mathrm{d}\mathbb{P}_{\nu_1}(x_j) \\ &\leq \int_{-\infty}^{x} \int_{-\infty}^{x-x_1} \cdots \int_{-\infty}^{x-x_1-\cdots-x_{j-1}} c_k^j \mathrm{d}\mathbb{P}_{\nu_k}(x_1) \ldots \mathrm{d}\mathbb{P}_{\nu_k}(x_j) \\ &= c_k^j F_{k,j}(x) \end{aligned}$$

□

This result allows to use Equation 2.3 to provide an upper bound on the balance function.

**Lemma 2.22** (Balance condition under Assumption 2.20)**.** *Under Assumption 2.20, the balance function of any arm $k \geq 2$ can be upper bounded by*

$$\alpha_k(M,j) \leq \gamma \times c_k^j \frac{\log(M)}{M} + \frac{1}{M^\gamma} \ ,$$

*for any $M \in \mathbb{N}$ satisfying $\log(M)/M \leq y_k$, any $\gamma > 1$, and $j \in \mathbb{N}$. Furthermore, the balance condition holds for any forced exploration $f_r$ satisfying $\frac{f_r}{\log\log(r)} \to +\infty$.*

*Proof.* Using Proposition 2.21 and Equation 2.3, for any $x \leq y_k$ the balance function of any arm $k \geq 2$ can be upper bounded by

$$\alpha_k(M,j) \leq c_k^j F_{k,j}(x) + G_{k,j}(x)^M = c_k^j F_{k,j}(x) + (1 - F_{k,j}(x))^M$$

Interestingly, to provide the smallest possible upper bound on $\alpha_k(M,j)$ we can simply consider minimizing the function $z \mapsto c_k^j z + (1-z)^M$. For simplicity we use that $(1-z)^M \leq \exp(-Mz)$, and then consider the value $z_M = \frac{\gamma \log(M)}{M}$ for some $\gamma > 1$ (which is not exactly the minimizer but of the same order). This provides the first statement for $M$ large enough such that $z_M \leq y_k$.

Now, consider a sequence $M_{j,r} = \frac{(\log r)^a}{r}$ for some $a > 1$ (we omit potential constants for simplicity of notations). We consider the $\sum_{r=1}^{T} \sum_{j=f_r}^{C_1 \log(r)} \alpha_k(M_{j,r})$. We use the upper bound we obtained on $\alpha_k(M_{j,r})$, and first remark that for any $\gamma > 1$ it holds that

$$\begin{aligned}
\sum_{r=1}^{T} \sum_{j=f_r}^{C_1 \log(r)} \frac{1}{M_{j,r}^{\gamma}} &= \sum_{r=1}^{T} \sum_{j=f_r}^{C_1 \log(r)} \frac{(\log(r))^{a\gamma}}{r^{\gamma}} \\
&= \sum_{r=1}^{T} C_1 \frac{(\log(r))^{a\gamma+1}}{r^{\gamma}} \\
&= \mathcal{O}(1) \,,
\end{aligned}$$

for any value of $f_r$. Hence, the difficulty comes from upper bounding the term resulting from the first part of the upper bound of $\alpha_k(M,j)$,

$$\begin{aligned}
\sum_{r=1}^{T} \sum_{j=f_r}^{C_1 \log(r)} c_k^j \frac{\log(M_{j,r})}{M_{j,r}} &= \sum_{r=1}^{T} \sum_{j=f_r}^{C_1 \log(r)} c_k^j \frac{\log(r/(\log(r)^a))}{r/(\log(r)^a)} \\
&\leq \sum_{r=1}^{T} \sum_{j=f_r}^{C_1 \log(r)} c_k^j \frac{\log(r)^{a+1}}{r} \\
&\leq \sum_{r=1}^{T} \frac{\log(r)^{a+1}}{r} \sum_{j=f_r}^{+\infty} c_k^j \\
&= \sum_{r=1}^{T} \frac{\log(r)^{a+1}}{r} \frac{c_k^{f_r}}{1-c_k} \\
&= \frac{1}{1-c_k} \sum_{r=1}^{T} \frac{\log(r)^{a+1} c_k^{f_r}}{r} \,.
\end{aligned}$$

We now investigate the possible values of $f_r$ to make this sum converge. For any $\varepsilon > 0$, this is the case if

$$\log(r)^{a+1} c_k^{f_r} = \exp((a+1)\log\log(r) - f_r \log(1/c_k)) \leq \log(r)^{-(1+\varepsilon)} ,$$

which holds if $f_r \geq \frac{a+2+\varepsilon}{\log(1/c_k)} \log\log(r)$. Summarizing these results, we obtain that the balance condition holds for $f_r$ satisfying $\frac{f_r}{\log\log(r)} \to +\infty$. □

**Examples** We provide a few examples of problems for which Assumption 2.20 holds. The first one is Single Parameter Exponential Families (SPEF), that contain some usual families of distributions.

**Proposition 2.23** (Any SPEF satisfies Assumption 2.20)**.** *Assume that for a family of distribution $\mathcal{F}$ there exists some functions $g, \psi$ and a set $\Theta$ such that for any $\nu \in \mathcal{F}$ there exists $\theta \in \Theta$ satisfying*

$$\frac{d\mathbb{P}_\nu}{d\eta} = e^{g(\theta)y - \psi(\theta)} ,$$

*for a reference measure $\eta$ and with $g(\theta)$ that is increasing with the mean of $\nu$. Then any bandit problem with arms in $\mathcal{F}$ satisfy Assumption 2.20.*

*Proof.* According to the definition and properties of SPEF for any $x \leq \mu_k$ it holds that

$$\begin{aligned}
\frac{d\mathbb{P}_{\nu_1}}{d\mathbb{P}_{\nu_k}}(x) &= \exp\left((g(\theta_1) - g(\theta_k))x - \Psi(\theta_1) + \Psi(\theta_k)\right) \\
&\leq \exp\left((g(\theta_1) - g(\theta_k))\mu_k - \Psi(\theta_1) + \Psi(\theta_k)\right) \\
&\leq \exp\left(-\mathrm{kl}(\mu_k, \mu_1)\right) ,
\end{aligned}$$

with $\mathrm{kl}(\mu_k, \mu_1) > 0$ as $\mu_k < \mu_1$. This is exactly what we need for Assumption 2.20 with $y_k = \mu_k$ and $c_k = \exp(-\mathrm{kl}(\mu_k, \mu_1))$ □

This result combined with Lemma 2.22 show that SPEF are well adapted to sub-sampling based algorithms, but we can find other type of assumptions on the distributions that would make Assumption 2.20 hold. For instance, a quite common assumption in bandits is to consider that rewards are generated by adding a noise to their mean, i.e any reward $X^k$ received from arm $k$ can be written as

$$X^k = \mu_k + \eta ,$$

where $\mu_k$ is the mean of arm $k$ and $\eta$ is a noise, drawn from some fixed probability distribution of mean zero. For instance, $\eta$ can be a be a *sub-gaussian* or *sub-exponential* noise. In that case, if the noise admits a density $f$ then the densities of arm $1$ and arm $k$ in any $x$ are respectively $f(x-\mu_1)$ and $f(x-\mu_k)$. Hence, Assumption 2.20 is valid if there exists $c_k \in (0,1)$ such that for $x$ small enough

$$\frac{f(x-\Delta_k)}{f(x)} \leq c_k \ .$$

If the noise is supported on $\mathbb{R}$ this depends of the asymptotic behavior of $f$. This will be typically true if $f$ has an exponential or gaussian asymptotic equivalent.

Another question is to determined whether or not forced exploration is necessary. Indeed, the upper bound on the balance function may not be tight and it may be possible to improve it for specific distributions. In Appendix 2.7 we actually prove that forced exploration is not necessary for Gaussian (with shared variance), Poisson and Bernoulli distributions. On the other hand, we discuss that the balance condition cannot be satisfied without at least some forced exploration for exponential distributions. Finding other characterizations of families of distributions that can satisfy the balance condition and the level of forced exploration they need is an interesting future research direction.

### 2.4.4 Summary of our results

We provide a short summary of the results we obtained in previous sections:

- With Theorem 2.12, we proved that SP-SDA has a regret of order $\left(\sum_{k=2}^{K} I_1(\mu_k)^{-1}\right)\log(T)$ if the family of distribution $\mathcal{F}$ satisfies Assumption 2.4 (concentration of means with exponential decay) for some rate functions $(I_k)_{k\in\{1,\dots,K\}}$, if SP is a *block sampler* bringing *sufficient diversity*, and if any bandit problem from $\mathcal{F}$ satisfies the *balance condition*.
- Lemma 2.13 and Lemma 2.16 show that both the *Random Block* and *Last Block* samplers bring a sufficient diversity of sub-samples to their respective algorithms.
- Lemma 2.22 shows that the balance condition is satisfied if $\mathcal{F}$ satisfies Assumption 2.20 with forced exploration $f_r$ satisfying $f_r/\log\log(r) \to +\infty$. Furthermore, Proposition 2.23 shows that this assumption holds if $\mathcal{F}$ is a SPEF.

Hence, we can conclude this part with two results that showcase the potential of the algorithms based on sub-sampling.

**Corollary 2.24** (of Theorem 2.12, Lemma 2.13, 2.16 and Lemma 2.22)**.** *Let $\nu = (\nu_1, \dots, \nu_K) \in \mathcal{F}^K$ be a bandit problem satisfying assumption 2.20 (dominant tail for the best arm), with arms satisfying assumption 2.4 (concentration of empirical means) with rate functions $I_1, \dots, I_K$. Then,*

*if the forced exploration $f_r$ satisfies $\frac{f_r}{\log\log(r)} \to +\infty$ and $f_r = o(\log(r))$ the number of pulls of each sub-optimal arm $k > 2$ for LB-SDA and RB-SDA can be upper bounded by*

$$\mathbb{E}[N_k(T)] \leq \frac{1+\varepsilon}{I_1(\mu_k)} \log(T) + \mathcal{O}_\varepsilon(1) \ ,$$

*for any $\varepsilon > 0$.*

As the LB and RB samplers are proved to be sufficient, the learner only needs to check that the family of distributions satisfy the two assumptions we consider. We recall that assumption 2.4 is always valid if the distributions are *light-tailed*, which is in general a knowledge that is affordable in practice. The other assumption may be more difficult to check but is easy to re-formulate as: *worst-case scenarios need to be less probable for better arms*. We showed that this is true for SPEF (e.g Bernoulli, Gaussian with shared variance, Poisson, ...), that are very common distributions. Furthermore, Corollary 2.24 can be further refined for SPEF by remarking that these families of distributions satisfy $I_1(\mu_k) = \mathrm{kl}(\mu_k, \mu_1)$ (the proof simply relies on Chernoff inequality), so that the regret upper bound actually matches the lower bound of Lai and Robbins (1985).

**Corollary 2.25** (Asymptotic optimality for SPEF). *Let $\nu = (\nu_1, \dots, \nu_K) \in \mathcal{F}^K$ be a bandit problem, where $\mathcal{F}$ is a SPEF satisfying Assumption 2.4. Then, if the forced exploration $f_r$ satisfies $\frac{f_r}{\log\log(r)} \to +\infty$ and $f_r = o(\log(r))$ the number of pulls of each sub-optimal arm $k > 2$ for LB-SDA and RB-SDA can be upper bounded by*

$$\mathbb{E}[N_k(T)] \leq \frac{1+\varepsilon}{\mathrm{kl}(\mu_k, \mu_1)} \log(T) + \mathcal{O}_\varepsilon(1) \ ,$$

*for any $\varepsilon > 0$. Both RB-SDA and LB-SDA are then* asymptotically optimal.

Again, a striking result is that these theoretical guarantees are simultaneously achieved for very different examples of distribution: some are bounded (Bernoulli), un-bounded (Gaussian, Poisson, Exponential), discrete or continuous, ...and SDA works without using any of those information in its implementation. The only algorithm with the same guarantees is SSMC (Chan, 2020), that inspired SDA, and we think that we demonstrated further interesting properties of algorithms based on sub-sampling. First, we bridged the gap between BESA and SSMC by proposing algorithms that are their close variants (WR-SDA and LB-SDA respectively) in the same framework, with LB-SDA being simpler and more computationally efficient than SSMC. Then, we showed that both *randomized* and *deterministic* samplers can achieve strong theoretical guarantees, for families of distributions satisfying a generic property that we defined. Both approaches can have their interest according to additional constraints faced by the learner.

Finally, we also showed that forced exploration is not necessary for some distributions (Bernoulli, Gaussian, Poisson), but it seems necessary to some extent for others (Exponential).

We further remark that RB-SDA and LB-SDA being asymptotically optimal for binomial distributions is noteworthy. Indeed, this ensures that it is possible to build on these algorithms to propose a bandit algorithm that has logarithmic regret for distributions that are bounded in a known support. To do so, we can use the binarization trick already proposed by Agrawal and Goyal (2013a) for Thompson Sampling, and run RB-SDA or LB-SDA on top of a binarized history $\mathcal{H}'_k$ for each arm $k$ in which a reward $Y_{k,s}$ is replaced by a binary pseudo-reward is $Y'_{k,s}$ generated from a Bernoulli distribution with mean $Y_{k,s}$. The resulting algorithm inherits the regret guarantees of these algorithms applied to Bernoulli distributions.

An interesting future work consists in finding a variant of SDA that could work when Assumption 2.20 is not satisfied. For bounded distributions this can be done with the binarization trick introduced in previous paragraph. However, the right way to achieve this goal is much less clear for unbounded distributions.

## 2.5 Experiments

In this section, we perform experiments on simulated data in order to illustrate the good performance of the four instances of SDA algorithms introduced in Section 2.2 for various distributions. The Python code used to perform these experiments is available on Github.

**Bernoulli and Gaussian arms** First, in order to illustrate Corollary 2.25, we investigate the performance of SDA for both Bernoulli and Gaussian distributions (with known variance 1). Our first objective is to check that for a finite horizon the regret of SDA is comparable with the regret of Thompson Sampling (with respectively a beta and improper uniform prior), which efficiently use the knowledge of the distribution. Our second objective is to empirically compare different variants of SDA to other non-parametric approaches based on sub-sampling (BESA, SSMC) or on re-sampling. For Bernoulli and Gaussian distribution, Non-Parameteric TS coincides with Thompson Sampling, so we focus our study on algorithms based on history perturbation. We experiment with PHE (Kveton et al., 2019a) for Bernoulli bandits and ReBoot (Wang et al., 2020) for Gaussian bandits, as those two algorithms are guaranteed to have logarithmic regret in each of these settings. As advised by the authors, we use a parameter $a = 1.1$ for PHE and $\sigma = 1.5$ for ReBoot. As results in Appendix 2.7 show that SDA do not require forced exploration with Bernoulli and Gaussian arms we set $f_r = 1$.

We ran experiments on 4 different Bernoulli bandit models, that we present in Table 2.1. The objective of the experiments is to test different cases that can happen with Bernoulli distributions: large mean, high variance ($\mu$ around $0.5$), low mean, and many similar arms.

**Table 2.1** – Experiments with Bernoulli arms

| Name | Number of arms | Means $\{\mu_1, \dots, \mu_K\}$ |
|---|---|---|
| **xp 1B** | $K = 2$ | $\mu = \{0.8, 0.9\}$ |
| **xp 2B** | $K = 2$ | $\mu = \{0.5, 0.6\}$ |
| **xp 3B** | $K = 10$ | $\mu_1 = 0.1, \mu_{2:4} = 0.01, \mu_{5:7} = 0.03, \mu_{8:10} = 0.05$ |
| **xp 4B** | $K = 8$ | $\mu = \{0.9, 0.85, \dots, 0.85\}$ |

We then propose 3 experiments with gaussian arms $\mathcal{N}(\mu_k, 1)$, that we present in Table 2.2.

**Table 2.2** – Experiments with Gaussian arms with variance $\sigma^2 = 1$

| Name | Number of arms | Means $\{\mu_1, \dots, \mu_K\}$ |
|---|---|---|
| **xp 1G** | $K = 2$ | $\mu = \{0.5, 0\}$ |
| **xp 2G** | $K = 4$ | $\mu = \{0.5, 0, 0, 0\}$ |
| **xp 3G** | $K = 4$ | $\mu = \{1.5, 1, 0.5, 0\}$ |

For each experiment with Bernoulli and Gaussian arms, Table 2.3 and Table 2.4 report an estimate of the regret at time $T = 20000$ based on 5000 independent runs, as well as standard deviation across al trajectories. The best performing algorithms are highlighted in bold. In Figure 2.2 we further plot the regret of several algorithms as a function of time (in logarithmic scale) for $t \in [15000; 20000]$ for one Bernoulli and one Gaussian experiment respectively. We also add the Lai and Robbins lower bound (Lai and Robbins, 1985). These figures aim at showing that the first order term of the empirical regret matches the lower bound for SDA (and other) algorithms.

**Table 2.3** – Regret and at $T = 20000$ for Bernoulli arms, with standard deviation

| | Benchmark | | | | SDA | | | |
|---|---|---|---|---|---|---|---|---|
| xp | TS | PHE | BESA | SSMC | RB | WR | LB | LDS |
| 1B | **11.2** (10.) | 25.9 (87.9) | **11.7** (12.1) | **12.3** (7.3) | **11.5** (10.1) | **11.6** (10.2) | **12.2** (7.4) | **11.4** (9.0) |
| 2B | **22.9** (29.2) | **24.0** (22.0) | **22.1** (25.2) | **24.3** (38.2) | **22.0** (34.5) | **21.5** (17.3) | **24.0** (24.6) | **21.8** (24.5) |
| 3B | **94.2** (15.8) | 248.1 (25.5) | **88.1** (89.2) | 100.1 (20.0) | **89.0** (19.8) | **86.9** (21.7) | 100.7 (21.3) | **89.2** (21.8) |
| 4B | **108.1** (45.1) | 216.5 (89.8) | 147.5 (209.8) | 119.9 (40.8) | **105.1** (41.1) | **106.9** (42.1) | 119.6 (42.7) | **106.8** (47.7) |

In all of these experiments, we notice that SDA algorithms are indeed strong competitors to Thompson Sampling (with appropriate prior) for both Bernoulli and Gaussian bandits. Figure 2.2 further show that they are empirically matching the Lai and Robbins' lower bound

**Table 2.4** – Regret and at $T = 20000$ for Gaussian arms, with standard deviation

| | Benchmark | | | | SDA | | | |
|---|---|---|---|---|---|---|---|---|
| xp | TS | ReBoot | BESA | SSMC | RB | WR | LB | LDS |
| **1G** | **24.4** (17.1) | 92.2 (23.4) | **25.3** (27.1) | **26.9** (52.8) | **25.6** (62.8) | **24.7** (20.6) | **25.1** (17.9) | **26.5** (140.2) |
| **2G** | **73.5** (107.8) | 277.1 (41.3) | 122.5 (585.5) | **74.8** (34.7) | **71.0** (152.2) | **71.1** (50.2) | **74.6** (35.1) | **69.0** (50.4) |
| **3G** | **49.7** (26.9) | 190.9 (29.6) | 72.1 (410.3) | **51.3** (23.7) | **50.4** (156.5) | **50.0** (33.3) | **51.2** (22.4) | **48.6** (41.6) |

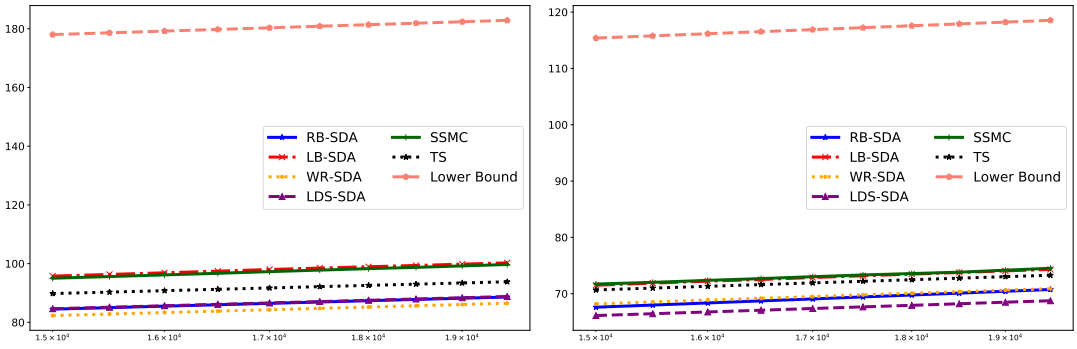


**Figure 2.2** – Regret as a function of time for xp 3B and xp 2G (Right), for $T = 2 \times 10^4$ and 5000 simulations. The $y$ axis is in logarithmic scale, and the $x$ axis starts at $T = 15 \times 10^4$ to illustrate the "asymptotic" regime of the algorithms (parallel straight lines correspond to a logarithmic regret with the same constant before the log).

on two instances, just like SSMC, which can be seen from the parallel straight lines with the $x$ axis in log scale. The fact that the lower bound is above shows that it is really asymptotic and only captures the right first order term. The same observation was made for all experiments, but is not reported due to space limitation. Even if we only established the asymptotic optimality of RB-SDA and LB-SDA, these results suggest that the other SDA algorithms considered in this chapter may also be asymptotically optimal. Compared to SDA, re-sampling algorithms based on history perturbation seem to be much less robust. Indeed, in the Bernoulli case, PHE performs very well for experiment 2, but is significantly worse than Thompson Sampling on the three other instances. In the Gaussian case, ReBoot always performs significantly worse than other algorithms. Finally, we notice that the standard deviations are comparable for SDA algorithms and TS. However, for gaussian arms some of them may experience large standard deviations for some experiments. This may be due to the fact that the constant upper bounding the balance function is larger than for bernoulli distributions, and may be reduced by the asymptotically negligible forced exploration in $f_r = \sqrt{\log(r)}$ (which would correspond to only 4 samples in our case). However, we notice that even TS suffers from this problem (xp 2G)

Turning our attention to algorithms based on sub-sampling, we first notice that WR-SDA seems to be a better generalization of BESA with 2 arms than the tournament approach currently proposed, as in experiments with $K > 2$, WR-SDA often performs significantly better than BESA. Then we observe that SSMC and SDA algorithms have similar performance. Looking a bit closer, we see that the performance of SSMC is very close to that of LB-SDA as we intuited in previous sections, whereas SDA algorithms based on "randomized" (or pseudo-randomized for LDS-SDA) samplers tend to perform slightly better on average, at the cost of larger variance in the results.

**Truncated Gaussian** Theorem 2.24 suggests that RB-SDA and LB-SDA may attain logarithmic regret beyond exponential families. As an illustration, we present the results of experiments performed with Truncated Gaussian distributions (in which the distribution of arm $k$ is that of $Y_k = 0 \vee (X_k \wedge 1)$ where $X_k \sim \mathcal{N}(\mu_k, \sigma^2)$). These distributions trivially satisfy Assumption 2.20 since the probability of obtaining 0 is strictly smaller for the best arm than for the other arms. We present in 2.5 the 4 problems we considered in this setting, and report in Table 2.6 the regret at time $T = 20000$ (estimated over 5000 runs) of various algorithms on four different problem instances:

**Table 2.5** – Experiments with Truncated Gaussian arms

| Name | Number of arms | Means $\{\mu_1, \dots, \mu_K\}$ and Std |
|---|---|---|
| **xp 1TG** | $K = 2$ | $\mu = \{0.5, 0.6\}$, $\sigma = 0.1$ |
| **xp 2TG** | $K = 2$ | $\mu = \{0, 0.2\}$, $\sigma = 0.3$ |
| **xp 3TG** | $K = 2$ | $\mu = \{1.5, 2\}$, $\sigma = 1$ |
| **xp 4TG** | $K = 4$ | $\mu = \{0.4, 0.5, 0.6, 0.7\}$, $\sigma = 1$ |

We include Non-Parametric TS which is known to be asymptotically optimal in this setting (while TS which uses a Beta prior and a binarization trick is not), PHE, and all algorithms based on sub-sampling. We again observe the good performance of SSMC and SDA algorithms across all experiments. They even outperform NPTS in some experiments, which suggests SDA algorithms may be also asymptotically optimal for this class of distributions.

**Exponential Arms** In Appendix 2.7, we show that exponential actually requires some level of forced exploration. In this section we propose to check the performance of our algorithms without and with forced exploration to see if it seems really necessary in practice or if forced exploration is just a proof artifact. We performed 6 experiments, with mean parameters reported in Table 2.7.

It is interesting to remark that the standard deviation of an exponential distribution is equal to its mean, so with similar gaps problems are harder when the means are high. We first report in Table 2.8 our result without forced exploration.

**Table 2.6** – Regret at $T = 20000$ for Truncated Gaussian arms

| | Benchmark | | | | | SDA | | | |
|---|---|---|---|---|---|---|---|---|---|
| xp | TS | NPTS | PHE | BESA | SSMC | RB | WR | LB | LDS |
| **1TG** | 21.9 (20.4) | 4.2 (0.6) | 22.3 (2.6) | **1.4** (1.7) | **1.5** (0.7) | **1.4** (1.1) | **1.4** (0.8) | **1.5** (0.7) | **1.4** (0.8) |
| **2TG** | 13.3 (7) | 8 (1.8) | 19.5 (3.8) | **4.6** (3.3) | **4.7** (2.3) | **4.4** (4.6) | **4.5** (3.1) | **4.6** (2.4) | **4.3** (2.9) |
| **3TG** | 9.7 (10.1) | **7.8** (4.5) | 48.5 (217.8) | **7.8** (9.4) | **7.6** (5) | **7.1** (10) | **7.7** (13.4) | 8.2 (27.5) | **7.1** (5.8) |
| **4TG** | 86.6 (57.8) | **70** (39.4) | 86 (53.7) | 76.5 (113.9) | **69.5** (40.9) | **64.9** (60.5) | **64.8** (43.9) | **68.7** (39.1) | **63.2** (51.1) |

**Table 2.7** – Experiments with Exponential arms

| Name | Number of arms | Means $\{\mu_1, \dots, \mu_K\}$ and Std |
|---|---|---|
| **xp 1E** | $K = 2$ | $\mu = \{1.5, 1\}$ |
| **xp 2E** | $K = 2$ | $\mu = \{0.2, 0.1\}$ |
| **xp 3E** | $K = 2$ | $\mu = \{11, 10\}$ |
| **xp 4E** | $K = 4$ | $\mu = \{4, 3, 2, 1\}$ |
| **xp 5E** | $K = 4$ | $\mu = \{0.4, 0.3, 0.2, 0.1\}$ |
| **xp 6E** | $K = 4$ | $\mu = \{5, 4, 4, 4\}$ |

First, we notice that the performance of the SDA in terms of the average regret are reasonable, although less impressive than with the other distributions we tested. IMED is almost always the best algorithm in these experiments, and SSMC performs pretty well on many examples (note that we left $f_r = \sqrt{\log(r)}$ for SSMC). We remark that there is much more variability in the results of RB-SDA, WR-SDA and LDS-SDA than before, where they performed quite similarly. For instance, we notice that on example 3, LDS-SDA and RB-SDA are much worse than WR-SDA. A look at the quantile table for this experiment (Table 2.9), which displays the empirical quantiles of the regret estimated over 5000 runs, shows that this is due to a small number of "bad" trajectories for these algorithms:

**Table 2.8** – Average Regret with Exponential Arms (with std) without forced exploration

| xp | TS | IMED | BESA | SSMC | RB | WR | LB | LDS |
|---|---|---|---|---|---|---|---|---|
| **1E** | 48.2 (191.8) | **40.0** (78.4) | 45.7 (114.1) | **41.9** (84.2) | 44.8 (121.4) | 45.4 (134.4) | 46.6 (176.8) | 45.5 (109.7) |
| **2E** | 3.8 (9.9) | **3.4** (3.6) | 4.2 (25.1) | **3.6** (41.9) | 4.1 (14.3) | 3.9 (13.4) | 3.9 (8.7) | 5.4 (49.5) |
| **3E** | 832.8 (1065.1) | **779.9** (896.9) | 820.5 (1304.6) | 856.9 (1111.0) | 848.4 (1533.3) | **778.4** (1118.7) | 846.7 (1150.1) | 877.7 (1708.7) |
| **4E** | 258.3 (519.6) | **234.6** (126.6) | 525.4 (2115.1) | **251.3** (328.3) | 272.6 (692.2) | 262.1 (524.4) | 263.8 (477.9) | 258.4 (599.0) |
| **5E** | **25.6** (51.2) | **24.0** (33.6) | 55.7 (219.9) | **25.6** (23.6) | **25.5** (46.7) | **25.0** (24.0) | 26.5 (36.8) | **24.7** (37.6) |
| **6E** | **618.7** (672.3) | **603.6** (576.8) | 1184.2 (3096.4) | **627.9** (755.6) | **595.7** (790.7) | **616.0** (780.2) | 652.6 (685.3) | **605.9** (871.4) |

**Table 2.9** – Quantiles of the distribution of empirical regret at $T = 10^4$ for Experiment 3 with exponential arms, over $5000$ runs.

| % of runs | TS | IMED | SSMC | RB | WR | LB | LDS |
|---|---|---|---|---|---|---|---|
| 20% | 319.8 | 336.0 | 335.0 | 261.0 | 290.0 | 326.0 | 261.8 |
| 50% | 626.0 | 650.0 | 661.0 | 532.0 | 568.5 | 642.0 | 536.0 |
| 80% | 1122.0 | 1080.0 | 1142.0 | 1006.0 | 1019.0 | 1143.2 | 1020.2 |
| 95% | 1924.1 | 1704.0 | 1846.0 | **2199.0** | 1817.2 | 1869.1 | **2134.1** |
| 99% | 4209.4 | 2632.9 | 3536.8 | **6813.1** | 4146.0 | 3762.3 | **7396.7** |

We see that up to the $80\%$ quantile, RB-SDA and LDS-SDA are even significantly better than IMED. This is very different when we look at the $95\%$ and $99\%$ quantiles, which are much greater for our 2 algorithms (even $2.5$ times greater for the $99\%$ quantile).

We believe that this very high variability prevents SDA to have a logarithmic regret for exponential arms. Still, the regret is not as bad as being linear, and a closer look at the balance function allows to prove a $\mathcal{O}((\log(T))^2)$ regret, with the term $\mathbb{P}(N_1(r) < C_1 \log(r))$ becoming the first order term of the regret. Hence, the experiments seem to confirm that the asymptotically negligible forced exploration is actually necessary for exponential arms. In Table 2.10 we run the same experiments but choose $f_r = \sqrt{\log(r)}$ as suggested in (Chan, 2020)

We remark that adding forced exploration results in a noticeable improvement for SDA algorithms, with RB-SDA, WR-SDA and LDS-SDA becoming competitive with IMED on most examples. Furthermore, LB-SDA becomes again comparable with SSMC. Considering all these

**Table 2.10** – Average regret with exponential arms: SDA with forced exploration

| xp | RB-SDA | WR-SDA | LB-SDA | LDS-SDA |
|---|---|---|---|---|
| **1E** | 44.9<br>(167.3) | **42.5**<br>(107.4) | **42.4**<br>(60.5) | 45.0<br>(176.0) |
| **2E** | **3.6**<br>(9.2) | **3.4**<br>(2.2) | 4.0<br>(27.9) | **3.6**<br>(11.2) |
| **3E** | 837.5<br>(1466.1) | **788.5**<br>(1222.1) | 827.7<br>(1055.3) | 832.3<br>(1514.6) |
| **4E** | **244.8**<br>(403.3) | **238.9**<br>(250.8) | 251.7<br>(248.5) | **246.0**<br>(323.4) |
| **5E** | **23.6**<br>(33.4) | **25.1**<br>(41.0) | **25.4**<br>(23.4) | **24.9**<br>(42.2) |
| **6E** | **578.9**<br>(651.9) | **595.1**<br>(561.3) | 631.2<br>(446.4) | **577.8**<br>(652.7) |

results and the fact that a forced exploration of $f_r = \sqrt{\log(r)}$ is not harmful in practice we recommend to always run SDA with this level of exploration.

**Bayesian Experiments** So far we tried our algorithms on specific instances of the distributions we considered. It is also interesting to check the robustness of the algorithms when the means of the arms are drawn at random according to some distribution. In this section we consider two examples: Bernoulli bandits where the arms are drawn uniformly at random in $[0, 1]$, and Gaussian distributions with the mean parameter of each arm itself drawn from a gaussian distribution $\mu_k \sim \mathcal{N}(0, 1)$. In both cases we draw $10000$ random problems with $K = 10$ arms and run the algorithms for a time horizon $T = 20000$. We experiment with TS, SSMC, RB-SDA and WR-SDA and IMED. We do not add LDS-SDA and LB-SDA as they are similar to RB-SDA and SSMC, respectively. In the Bernoulli case, we also run the PHE algorithm, which fails to compete with the other algorithms. This is not in contradiction with the results of Kveton et al. (2019a) as in the Bayesian experiments of this paper, arms are drawn uniformly in $[0.25, 0.75]$ instead of $[0, 1]$. Actually, we noticed that PHE with parameter $a = 1.1$ has some difficulties when several arms are close to 1.

**Table 2.11** – Average Regret on 10000 random experiments with Bernoulli Arms

| T | TS | IMED | PHE | SSMC | RB | WR |
|---|---|---|---|---|---|---|
| $10^2$ | 13.8 | 15.1 | 16.7 | 16.5 | 14.8 | 14.3 |
| $10^3$ | 27.8 | 31.9 | 39.5 | 34.2 | 31.8 | 30.9 |
| $10^4$ | 45.8 | 51.2 | 72.3 | 55.0 | 51.1 | 50.6 |
| $2.10^4$ | 52.2 | 57.6 | 85.6 | 61.9 | 57.7 | 57.3 |

**Table 2.12** – Average Regret on 10000 random experiments with Gaussian Arms

| T | TS | IMED | WR | RB | SSMC |
|---|---|---|---|---|---|
| $10^2$ | 41.2 | 45.1 | 38.3 | 38.1 | 40.6 |
| $10^3$ | 76.4 | 82.1 | 72.7 | 70.4 | 76.2 |
| $10^4$ | 118.5 | 124.0 | 115.8 | 111.8 | 120.1 |
| $2.10^4$ | 132.6 | 138.1 | 130.2 | 125.7 | 135.1 |

Results reported in Tables 2.11 and 2.12 show that RB-SDA and WR-SDA are strong competitors to TS and IMED for both Bernoulli and Gaussian bandits. Recall that these algorithm operate without the need for a specific tuning for each distribution, unlike TS and IMED. Moreover, observe that in the Bernoulli case, TS further uses the same prior as that from which the means are drawn.

**Computational aspects** To choose a sub-sampling based algorithm, numerical consideration can be taken into account. First, compared to algorithms like UCB1 or Thompson Sampling the main drawback of sub-sampling based algorithms is that they require to store the full history of rewards. This motivated our study in Chapter 3 about a variant of LB-SDA with limited memory. On the other hand, the computation cost of sub-sampling varies across algorithms: block samplers are generally more efficient than WR-SDA as the latter requires to draw a random subset while the formers only need at most to draw the random integer starting the block and compute the sub-samples' mean. However, for distributions with finite supports WR-SDA can be made as efficient as TS using multivariate geometric distributions. LDS-SDA could be preferred to RB-SDA to avoid randomization, as it uses a deterministic sequence. Finally, LB-SDA has the smallest computational cost in the SDA family and while its performance is very close to that of SSMC, it can avoid the cost of scanning all the sub-sample means in this algorithm. The computational cost of these two algorithms depends on what happened during the round: their update can be made very efficient when the leader does not change and is pulled. Indeed, in that case one only needs to update $K-1$ means by replacing the oldest observation by the last collected. However, the alternative case is costly for SSMC as it requires to perform a complete scan of the leaders' history. Under such scenario LB-SDA is much more efficient, and does not cost more than a step of RB-SDA. We complete this discussion in next chapter.

## 2.6 Appendix A: Sufficient diversity for RB-SDA (Lemma 2.13)

We start with a decomposition that follows the steps of Baransi et al. (2014) for BESA with 2 arms that we generalize for $K$ arms. Furthermore, in the following we first write the proof for a forced exploration $f_r = 1$, as the notations are already heavy, and we detail in the next section 2.6.1 why introducing an asymptotically negligible forced exploration does not change the proof.

We first denote by $r_j$ the round of the $j^{th}$ play of arm 1 with $r_0 = 0$ and let $\tau_j = r_{j+1} - r_j$. We notice that $\tau_0 \leq K$ as all arms are initialized once. Then:

$$\begin{aligned}\mathbb{P}\left(N_1(r) \leq C_1 \log(r)\right) &\leq \mathbb{P}\left(\exists j \in \{1, ..., C_1 \log(r)\} : \tau_j \geq r/C_1 \log(r) - 1\right) \\ &\leq \sum_{j=1}^{C_1 \log(r)} \mathbb{P}\left(\tau_j \geq r/C_1 \log(r) - 1\right)\end{aligned}$$

Indeed, if we assume that $\forall j\ \tau_j \leq r/C_1 \log(r) - 1$ then $t_{C_1 \log(r)} = \sum_{j=0}^{C_1 \log r} \tau_j < r$, which yields $N_\ell(r) > C_1 \log r + 1$. We now fix $j \leq C_1 \log(r)$ and upper bound the probability of the event

$$\mathcal{E}_j := \left\{\tau_j \geq r/C_1 \log(r) - 1\right\}.$$

On this event arm 1 lost at least $r/C_1 \log(r)$ consecutive duels between $r_j + 1$ and $r_{j+1}$ (either as a challenger of as the leader) which yields

$$\begin{aligned}\mathbb{P}(\mathcal{E}_j) \leq \mathbb{P}(\forall s \in \{r_j + 1, ..., r_j + \lfloor r/C_1 \log(r) - 1 \rfloor\} : \\ \{\bar{Y}_{1,j} \leq \bar{Y}_{\ell(s), S_1^s(N_{\ell(s)}(s), j)}, N_1(s) = j, N_{\ell(s)}(s) \geq j\} \cup \{\ell(s) = 1, N_1(s) = j\})\end{aligned}$$

The important change compared to the proof of Baransi et al. (2014) is that with $K > 2$, we don't know the identity of the leader and the leader is not necessarily pulled if it wins its duel against 1. We then notice that when $r$ is large, the time range considered in each $\mathcal{E}_j$ includes a large number of rounds. By looking at the second half of this time range only, we can ensure that the leader has been drawn a large number of times. More precisely, introducing the two intervals

$$\begin{aligned}\mathcal{M}_{j,r}^1 &= \left[r_j + 1, r_j + \left\lfloor \frac{r/(C_1 \log(r)) - 1}{2} \right\rfloor\right] \\ \mathcal{M}_{j,r}^2 &= \left[r_j + \left\lceil \frac{r/(C_1 \log(r)) - 1}{2} \right\rceil, r_j + \lfloor r/(C_1 \log(r)) \rfloor - 1\right]\end{aligned}$$

it holds that

$$\mathbb{P}(\mathcal{E}_j) \leq \mathbb{P}(\forall s \in \mathcal{M}^2_{j,r} : \{\bar{Y}_{1,j} \leq \bar{Y}_{\ell(s),S^s_1(N_{\ell(s)}(s),j)}, N_1(s) = j, N_{\ell(s)}(s) \geq j\} \\ \cup \{\ell(s) = 1, N_1(s) = j\}) \ .$$

But we know that on $\mathcal{M}^2_{j,r}$ the leader must has been selected at least $\frac{1}{K}\left(j + \left\lceil \frac{r/(C_1 \log(r))-1}{2} \right\rceil\right)$ times. Let $r_K$ be the first integer such that $C_1 \log(r) < \frac{1}{K-1}\left\lceil \frac{r/(C_1 \log(r))-1}{2} \right\rceil$, for every $r \geq r_K$, as $j \leq C_1 \log(r)$, the leader has been selected strictly more than $j$ times, which prevents arm 1 from being the leader for any round in $\mathcal{M}^2_{j,r}$. Hence, for $r \geq r_K$, for all $j \leq C_1 \log(r)$,

$$\mathbb{P}(\mathcal{E}_j) \leq \mathbb{P}\left(\forall s \in \mathcal{M}^2_{j,r} : \{\bar{Y}_{1,j} \leq \bar{Y}_{\ell(s),S^s_1(N_{\ell(s)}(s),j)}, N_1(s) = j, N_{\ell(s)}(s) \geq j\}\right) \ .$$

To remove the problem of the identity of the leader we would like to find a way to fix our attention on one arm. To this extent, we notice that during an interval of length $|\mathcal{M}^2_{j,r}|$, if there are only $K-1$ candidates for the leader then one of them must have been leader at least $m_r := |\mathcal{M}^2_{j,r}|/(K-1) - 1$ times during this range. We also know that at any round in $\mathcal{M}^2_{j,r}$, the leader satisfies $N_{\ell(s)}(s) \geq (t_j + \lfloor \frac{r/C_1 \log(r)-1}{2} \rfloor)/K - 1 \geq (\lfloor \frac{r/C_1 \log(r)-1}{2} \rfloor)/K - 1 = \frac{|\mathcal{M}^1_{j,r}|}{K} - 1 := c_r$. Observe that $m_r > c_r$. Finally, we introduce the notation

$$I^k_{j,r} = \{s \in \mathcal{M}^2_{j,r} : \ell(s) = k\}$$

for the set of rounds in $\mathcal{M}^2_{j,r}$ in which a particular arm $k$ is leader. From the above discussion, we know that there exists an arm $k$ such that $|I^k_{j,r}| \geq m_r$.

To ease the notation, we introduce the event

$$\mathcal{W}^k_{s,j} = \left\{\left\{\bar{Y}_{1,j} < \bar{Y}_{k,S^s_1(N_k(s),j)}\right\}, N_k(s) \geq c_r, N_1(s) = j\right\}$$

and write

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_j) &\leq \mathbb{P}\left(\bigcap_{s \in \mathcal{M}^2_{j,r}} \bigcup_{k=2}^{K} \{\ell(s) = k, 1 \notin \mathcal{A}_s)\}\right) \\
&\leq \mathbb{P}\left(\bigcap_{k=2}^{K} \bigcap_{s \in I^k_{j,r}} \mathcal{W}^k_{s,j}\right) \\
&\leq \mathbb{P}\left(\bigcup_{k=2}^{K} \left\{|I^k_{j,r}| > m_r, \bigcap_{s \in I^k_{j,r}} \mathcal{W}^k_{s,j}\right\}\right) \\
&\leq \sum_{k=2}^{K} \mathbb{P}\left(|I^k_{j,r}| > m_r, \bigcap_{s \in I^k_{j,r}} \mathcal{W}^k_{s,j}\right) \ .
\end{aligned}$$

Finally, we define for any integer $M$ the event that we can find $M$ pairwise non-overlapping sub-samples in the set of the sub-samples of arm $k$ drawn in rounds $s \in I_{j,r}^k$:

$$\mathcal{F}_{j,M}^{k,r} = \left\{ \exists i_1, ..., i_M \in I_{j,r}^k : \forall m < m' \in [M], S_1^{i_m}(N_k(i_m), j) \cap S_1^{i_{m'}}(N_k(i_{m'}), j) = \emptyset \right\}$$

Introducing $H_{j,r}^k = \min_{s \in I_{j,r}^k} N_k(s)$, the minimal size of the history of arm $k$ during rounds in $I_{j,r}^k$ (which is known to be larger than $c_r$ as $k$ is leader in these rounds), one has

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_j) &\leq \sum_{k=2}^{K} \mathbb{P}\left( |I_{j,r}^k| > m_r, \cap_{s \in I_{j,r}^k} \mathcal{W}_{s,j} \cap \{ \mathcal{F}_{j,M}^{k,r} \cup \bar{\mathcal{F}}_{j,M}^{k,r} \} \right) \\
&\leq \sum_{k=2}^{K} \mathbb{P}\left( |I_{j,r}^k| \geq m_r, H_{j,r}^k \geq c_r, \bar{\mathcal{F}}_{j,M}^{k,r} \right) + \sum_{k=2}^{K} \mathbb{P}\left( |I_{j,r}^k| > m_r, \cap_{s \in I_{j,r}^k} \mathcal{W}_{s,j} \cap \mathcal{F}_{j,M}^{k,r} \right) \qquad (2.4)
\end{aligned}$$

**Upper bound on the first term in (2.4)** The probability $\mathbb{P}\left( |I_{j,r}^k| \geq m_r, H_{j,r}^k \geq c_r, \bar{\mathcal{F}}_{j,M}^{k,r} \right)$ can be upper bounded by

$$\mathbb{P}\left( \#\left\{ \text{pairwise non-overlapping subsets in } (S_1^s(N_k(s), j))_{s \in I_{j,r}^k} \right\} < M \middle| \left\{ |I_{j,r}^k| > m_r, H_{j,r}^k \geq c_r \right\} \right) .$$

This probability can be related to some intrinsic properties of the sampler $\mathrm{SP}(H, j)$. To formalize this, we introduce the following definition.

**Definition 2.26.** *For every integers $N, H, j$ such that $H > j$, $X_{N,H,j}$ is a random variable which counts the maximum number of non-overlapping subsets among $N$ i.i.d. samples from $SP(H, j)$.*

Letting $H_1, \ldots, H_{m_r}$ be integers that are all larger than $c_r$, and letting $S_1, \ldots, S_{m_r}$ be independent subsets such that $S_i \sim \mathrm{SP}(H_i, j)$, the above probability is upper bounded by

$$\mathbb{P}\left( \#\left\{ \text{pairwise non-overlapping subsets in } (S_i)_{i=1}^{m_r} \right\} < M \right)$$

which is itself upper bounded by $\mathbb{P}\left( X_{m_r, c_r, j} < M \right)$.

This last inequality is quite intuitive: if one draws subsets of size $j$ from histories that may be larger than $c_r$, there is more "room" for non-overlapping subsets than if we always draw them from the same history of size $c_r$. For Random Block sampling, where the drawn subset is fully determined by the random position of its first element, to formalize this intuition it is sufficient to prove that if $X_i, Y_i$ are two sequences of random variables such that $X_i$ is uniform in $[H_i - j]$ and $Y_i$ is uniform in $[H - j]$, where $H_i \geq H$, the random variable that counts the maximal number of elements in the sequence $(Y_i)$ whose pairwise distance are larger than $j$ is stochastically dominated by that the same random variable but for the sequence $(X_i)$.

**Upper bound on the second term in (2.4)** On the event $\left(|I^k_{j,r}| > m_r, \cap_{s\in I^k_{j,r}} \mathcal{W}_{s,j} \cap \mathcal{F}^{k,r}_{j,M}\right)$, one can define $\tilde{\imath}_1, \dots, \tilde{\imath}_M$ the first $M$ rounds in $I^k_{j,r}$ for which the subsets $\tilde{S}_m := S^{\tilde{\imath}_m}(N_k(\tilde{\imath}_m), j)$ are pairwise non-overlapping and we get

$$\mathbb{P}\left(|I^k_{j,r}| > m_r, \cap_{s\in I^k_{j,r}} \mathcal{W}_{s,j} \cap \mathcal{F}^{k,r}_{j,M}\right) \leq \mathbb{P}\left(\forall m \in [M], \bar{Y}_{1,j} \leq \bar{Y}_{k,\tilde{S}_m}\right) .$$

By definition the subsets $\tilde{S}_m$ are pairwise non-overlapping, hence the sub-samples $\bar{Y}_{k,\tilde{S}_m}$ are independent. We prove that this probability can be in fact upper bound by the *balance function* we defined in section 2.3.

Indeed, introducing $X \sim \nu_{1,j}$ and an independent i.i.d. sequence $Z_i \sim \nu_{k,j}$, one can write

$$\begin{aligned}
\mathbb{P}\left(|I^k_{j,r}| > m_r, \cap_{s\in I^k_{j,r}} \mathcal{W}_{s,j} \cap \mathcal{F}^{k,r}_{j,M}\right) &\leq \mathbb{P}(X < \min_{i\in[M]} Z_i) \\
&= \mathbb{E}_{\substack{X\sim\nu_{1,j} \\ Z\sim\nu_{k,j}^{\otimes j}}}\left[\prod_{i=1}^{M} \mathbb{1}_{X\leq Z_i}\right] \\
&= \mathbb{E}_{X\sim\nu_{1,j}}\left[\mathbb{E}_{Z\sim\nu_{k,j}^{\otimes j}}\left[\prod_{i} \mathbb{1}_{X\leq Z_i}\middle| X\right]\right] \\
&= \mathbb{E}_{X\sim\nu_{1,j}}\left[(1-F_{k,j}(X))^M\right] \\
&= \alpha_k(M,j).
\end{aligned}$$

**Conclusion** Putting things together, we have proved that

$$\mathbb{P}(\mathcal{E}_j) \leq (K-1)\mathbb{P}\left(X_{m_r,c_r,j} < M\right) + \sum_{k=2}^{K} \alpha_k(M,j),$$

where $X_{N,H,j}$ and $\alpha_k(M,j)$ are introduced in Definition 2.26 and 2.9 respectively. If we replace $M$ by the sequence $\beta_{r,j}$ we have

$$\begin{aligned}
\sum_{r=1}^{T} \mathbb{P}(N_1(r) \leq C_1\log(r)) &\leq r_K + \sum_{r=r_K}^{T}\sum_{j=1}^{C_1\log(r)}\left[(K-1)\mathbb{P}\left(X_{m_r,c_r,j} < \beta_{r,j}\right) + \sum_{k=2}^{K}\alpha_k(\beta_{r,j},j)\right] \\
&\leq r_K + \sum_{r=r_K}^{T}\sum_{j=1}^{C_1\log(r)}\left[(K-1)\mathbb{P}\left(X_{c_r,c_r,j} < \beta_{r,j}\right) + \sum_{k=2}^{K}\alpha_k(\beta_{r,j},j)\right] ,
\end{aligned}$$

as $c_r \leq m_r$. We then conclude by using Lemma 2.14, introduced earlier in this chapter. The quantity $\mathbb{P}\left(X_{c_r,c_r,j} < \beta_{r,j}\right)$ admits an exponential upper bound in $c_r$, which ensures that

$$\sum_{r=r_K}^{T}\sum_{j=1}^{C_1\log(r)} (K-1)\mathbb{P}\left(X_{c_r,c_r,j} < \beta_{r,j}\right) = \mathcal{O}(1) ,$$

which concludes the proof that RB-SDA provides sufficient diversity according to definition 2.10.

### 2.6.1 Adding Forced Exploration in the proof

The idea is to use the same proof sketch as without forced exploration. We consider any sequence $f_r$ of the form $f_r = (\log r)^{\frac{1}{k}}$ for some $k > 1$ for simplicity. The following proof sketch can be adapted to other sequences. Let us denote for simplicity $f^{-1}(x) = \inf\{r \in \mathbb{N} : f_r \geq x\}$ for any $x \in \mathbb{R}$.

Let us consider the round $a_r = f^{-1}(f_r - 1)$. At this round, the value of exploration function is $f_r - 1 = \log(r)^{1/k} - 1$, which means that the number of pulls of arm 1 is at least $\lfloor (\log r)^{1/k} - 1 \rfloor$.

Now we aim at proving that the number of rounds in the interval $r - a_r$ is very close to $r$ when $r$ is large. To do that, we use that

$$\begin{aligned} a_r &= f^{-1}(f_r - 1) \\ &= \exp(((\log r)^{\frac{1}{k}} - 1)^k) \ . \end{aligned}$$

We then compare the exponent with $\log r$. For $\eta \in (0, 1)$ and $r$ large enough it holds that

$$\begin{aligned} \log(r) - ((\log r)^{\frac{1}{k}} - 1)^k &= \log(r) \left(1 - \left(1 - \frac{1}{(\log(r))^{1/k}}\right)^k\right) \\ &\geq \log(r) \left(1 - \exp\left(-\frac{k}{(\log r)^{1/k}}\right)\right) \\ &\geq \log(r) \left(1 - \left(1 - \eta \frac{k}{(\log r)^{1/k}}\right)\right) \\ &= \eta k \log(r)^{1-\frac{1}{k}} \\ &\longrightarrow +\infty \ , \end{aligned}$$

so $a_r = o(r)$, and we can conclude that for any $\gamma \in (0, 1)$, there exists some round $r_\gamma$ such that for $r \geq r_\gamma$,

$$r - a_r \geq \gamma r \ .$$

This means that after the round $a_r$ arm 1 faces a linear amount of duels, and has an history of at least $j = \lfloor (\log r)^{1/k} - 1 \rfloor$ samples. Introducing $b_r$ the random variable giving the first time when $N_1(b_r) = \lfloor (\log r)^{1/k} - 1 \rfloor$, we necessarily have $b_r \leq a_r$.

Then, using the exact proof as in the previous section we finally obtain a result of the form

$$\sum_{r=1}^{T}\mathbb{P}(N_1(r)\leq C_1\log(r))\leq r'_K+\sum_{r=r'_K}^{T}\sum_{j=\lfloor f_r\rfloor-1}^{(\log r)^2}\left[(K-1)\mathbb{P}(X_{c_r,c_r,j}<M'_{r,j})+\sum_{k=2}^{K}\alpha_k(M'_{r,j},j)\right] \quad (2.5)$$

for a new sequence $M_{r,j}$ smaller than the previous one but of the same order in $(r,j)$, and a new constant $r'_K$. This result concludes this part.

## 2.7 Appendix B: Further Analysis of the Balance Function of some distributions

In this section we detail the upper bound on the balance function that allows to show that gaussian distributions do not require forced exploration. This proof can be found in Appendix G of (Baudry et al., 2020), where we also prove similar results for the Bernoulli and Poisson distributions. We choose to provide the results for Gaussian distributions as an example of how such bound can be derived. Then, we provide a simple result that shows that exponential distributions require forced exploration to some extent.

In the next parts we use the notation $G(x)=1-F(x)$ where $F$ if the CDF of the distribution considered. For some arm distribution $\nu_i$ the distribution of the sum of $j$ independent observations drawn from $\nu_i$ is denoted by $\nu_{i,j}$. With this notation, for two arms 1 and 2 we write

$$\alpha(M,j)=\mathbb{E}_{Z\sim\nu_{1,j}}[G_{2,j}(Z)^M];.$$

### 2.7.1 Gaussian Distribution

For the Gaussian distribution we leverage the fact that both the PDF and CDF can be expressed with the PDF and CDF of the standard normal distribution. We use the notations $f$ and $F$ for the PDF and CDF of the $\mathcal{N}(0,1)$ distribution, write $G=1-F$, $\Delta$ for the gap between the two arms, and compute the expectation

$$\begin{aligned}
\alpha(M,j)&=\int_{-\infty}^{+\infty}f_{1,j}(x)G_{2,j}(x)^M dx\\
&\leq\int_{-\infty}^{z}f_{1,j}(x)G_{2,j}(x)^M dx+G_{2,j}(z)^M,\forall z\in\mathbb{R}\\
&\leq\int_{-\infty}^{z}f\left(\sqrt{j}(x-\mu_1)\right)G\left(\sqrt{j}(x-\mu_2)\right)^M dx+G_{2,j}(z)^M\\
&\leq\frac{1}{\sqrt{j}}\int_{-\infty}^{\sqrt{j}(z-\mu_2)}f\left(y-\sqrt{j}\Delta\right)G(y)^M dy+G_{2,j}(z)^M\,.
\end{aligned}$$

At this step we use that $f(x-a) = e^{-a^2/2+ax} f(x)$ for all $a, x$, and that the function $h : x \to (M+1)f(x)G(x)^M$ is a probability distribution of CDF $x \to 1 - G(x)^{M+1}$. These properties allow to write

$$\begin{aligned}
\alpha(M,j) &\leq \frac{1}{\sqrt{j}} \frac{e^{-j\frac{\Delta^2}{2}}}{M+1} \int_{-\infty}^{\sqrt{j}(z-\mu_2)} e^{\sqrt{j}\Delta y} h(y) dy + G_{2,j}(z)^M \\
&\leq \frac{1}{\sqrt{j}} \frac{e^{-j\frac{\Delta^2}{2}}}{M+1} e^{j\Delta(z-\mu_2)} \left(1 - G\left(\sqrt{j}(z-\mu_2\right)^{M+1}\right) + G\left(\sqrt{j}(z-\mu_2)\right)^M \\
&\leq \frac{1}{\sqrt{j}} \frac{e^{-j\left(\frac{\Delta^2}{2} - \Delta(z-\mu_2)\right)}}{M+1} + G\left(\sqrt{j}(z-\mu_2)\right)^M \ .
\end{aligned}$$

As the inequality is true for all $z \in \mathbb{R}$, it holds that

$$\forall y \in \mathbb{R}, \ \ \alpha(M,j) \leq \frac{e^{-j\frac{\Delta^2}{2}}}{M+1} e^{\sqrt{j}\Delta y} + G\left(y\right)^M \ .$$

Now let $y_M$ be such as $G(y_M) = 1 - \frac{1}{\sqrt{M}}$. This value ensures that the second term satisfies $G(y_M)^M \leq e^{-\sqrt{M}} = o(M^{-2})$. Observe that $y_M = F^{-1}(\frac{1}{\sqrt{M+1}})$. Using the following equivalent of the quantile function of the normal distribution when the quantile is small (see for instance Ledford and Tawn (1997)):

$$F^{-1}(p) = -\sqrt{\log \frac{1}{p^2} - \log\log \frac{1}{p} + \log 2\pi} + o_{p\to 0}(1) \ ,$$

there exists a constant $C \in \mathbb{R}$ such that $y_M \leq -C\sqrt{\log M - \log\log M + \log 4\pi}$. This yields

$$\alpha(M,j) \leq \frac{e^{-j\frac{\Delta^2}{2}}}{M+1} e^{-C\sqrt{j}\Delta\sqrt{\log M - \log\log M + \log 4\pi}} + e^{-\sqrt{M}} \ .$$

We then remark that for all $k \in \mathbb{N}^*$,

$$k \log\log M = o(C\sqrt{j}\Delta\sqrt{\log M - \log\log M + \log 4\pi})$$

and as a consequence that

$$\alpha(M,j) = o\left(\frac{e^{-j\frac{\Delta^2}{2}}}{(M+1)(\log M)^k}\right) \ ,$$

for any $k \in \mathbb{N}^*$.

This is sufficient to prove that the Gaussian distribution is balanced. Indeed, with $M = \mathcal{O}(r/(C_1 \log r))$ the series in $r$ is convergent ($\mathcal{O}\left(\frac{1}{r(\log r)^k}\right)$ for some $k > 1$).

### 2.7.2 Exponential Distribution

For $j = 1$, a direct calculation yields

$$\alpha(M, 1) = \frac{1}{1 + \left(\frac{\mu_1}{\mu_2}\right) M}.$$

Using this and $M = \beta \frac{r}{\log r}$ we then obtain

$$\begin{aligned}
\sum_{r=1}^{T} \sum_{j=1}^{\lfloor C_1 \log(r) \rfloor} \alpha_k(\lfloor \beta r / \log r \rfloor, j) &\geq \sum_{t=1}^{T} \alpha_k(\lfloor \beta r / \log r \rfloor, 1) \\
&= \sum_{t=2}^{T} \frac{1}{1 + \left(\frac{\mu_1}{\mu_k}\right) \lfloor \beta r / log r \rfloor} \\
&\geq \sum_{t=2}^{T} \frac{1}{1 + \left(\frac{\mu_1}{\mu_k}\right) \beta r / \log r} \\
&\geq C \sum_{r=2}^{T} \frac{\log(r)}{r} = \Omega(\log(T)^2),
\end{aligned}$$

where $C$ is some small enough constant that depend on $\mu_1, \mu_k$ and $\beta$.

# Chapter 3

# LB-SDA with Limited-Memory

In the previous chapter we introduced the family of *Sub-Sampling Dueling Algorithms*, based on the principle of pairwise comparisons between arms with sub-samples of the same size. We proposed various sub-sampling schemes and derived theoretical guarantees along with general intuitions on why this principle works in bandits. Strikingly, we showed that a very simple algorithm returning the *Last Block* of observations achieve strong theoretical guarantees. In this chapter we propose to extend this *Last Block Sub-Sampling Dueling Algorithm (LB-SDA)*, in two directions. First, we prove that its guarantees hold when limiting the algorithm memory to a polylogarithmic function of the time horizon. Then, we consider non-stationary scenarios in which the arm distributions evolve over time. We propose a natural variant of the algorithm in which only the most recent observations are used for sub-sampling, achieving optimal regret guarantees under the assumption of a known number of abrupt changes. Numerical simulations highlight the merits of this approach, particularly when the changes are not only affecting the means of the rewards.

## Contents

## 3.1 Introduction

In this chapter we still consider the Multi-Armed Bandit problem introduced in Chapter 1, as well as the algorithms based on *sub-sampling* that we introduced in Chapter 2. More specifically, we build on the Last-Block Sub-sampling Dueling Algorithm (LB-SDA), that we introduced and analyzed, and that is particularly attractive because of its simplicity and computational efficiency compared to other instances of SDA. Our first contribution in this chapter is to show that the theoretical guarantees of LB-SDA still hold, without additional changes, for a variant of the algorithm using a *limited memory* of the observations of each arm. We prove in particular that storing a poly-logarithmic amount of observations (instead of linear) in the number of rounds played is sufficient to maintain the theoretical guarantees, making the algorithm more tractable for larger time horizons. This is interesting since the main drawback of the algorithms analyzed in Chapter 2 is the requirement to store all $T$ rewards.

Furthermore, building a sub-sampling algorithm based on the most recent observations makes it an ideal candidate for a passively forgetting policy in a *non-stationary* environment. We presented a short introduction to the vast literature on non-stationary bandits in Section 1.3 of Chapter 1, detailing some known theoretical results in this setting along with common approaches to tackle this problem. Our second contribution is to propose a natural extension of the LB-SDA strategy to non-stationary environments by using a *sliding window*. By limiting the extent of the time window in which sub-sampling is allowed to occur, one obtains a passively forgetting non-parametric bandit algorithm, which we refer to as Sliding Window Last Block Sub-sampling Duelling Algorithm (SW-LB-SDA). To analyze the performance of this algorithm, we assume an abruptly changing environment in which the reward distributions change at unknown time instants called *breakpoints*. We show that SW-LB-SDA guarantees a regret of order $\mathcal{O}(\sqrt{\Gamma_T T \log(T)})$ for any abruptly changing environment with at most $\Gamma_T$ breakpoints, thus matching the lower bounds from (Garivier and Moulines, 2011), up to logarithmic factors. The only required assumption is that, during each stationary phase, the reward distributions satisfy Assumption 2.4 and 2.20, introduced in the previous chapter.

Due to its non-parametric nature, this algorithm can thus be used in many scenarios of interest beyond the standard bounded-rewards / change-in-the-mean framework. We discuss some of these scenarios in Section 3.5, where we validate numerically the potential of the approach by comparing it with a variety of state-of-the-art algorithms for non-stationary bandits. Hence, our contribution is not about providing novel insights on how non-stationarity can be handled by a bandit algorithm, but about analyzing the adaptation of a non-parametric algorithm to include a well-known passively forgetting strategy, that can tackle settings that are potentially not covered by existing approaches in terms of the family of distributions and changes that are allowed.

## 3.2 Preliminaries

Just as in Chapter 2, the algorithms to be presented below are designed for the *stochastic K-armed bandit* problem. We briefly re-introduce the two variants of this basic model that will be considered in the chapter: *stationary* and *abruptly changing* environments.

**Stationary environments** When the environment is stationary, we recall that the $K$ arms are characterized by the reward distributions $(\nu_k)_{k\leq K}$ and their associated means $(\mu_k)_{k\leq K}$, with $\mu^\star = \max_{k\in\{1,\dots,K\}} \mu_k$ denoting the highest expected reward. We denote by $(Y_{k,s})_{s\in\mathbb{N}}$ the i.i.d. sequence of rewards from arm $k$. We further recall that LB-SDA operates in successive rounds, whose length varies between 1 and $K$ time steps. At each round $r$, the *leader* denoted $\ell(r)$ is defined and $(K-1)$ duels with the remaining arms called *challengers* are performed. Denoting by $N_k(r)$ the number of pulls of arm $k$ up to the round $r$, the leader is the arm that has been most pulled: $\ell(r) \in \text{argmax}_{k\in\{1,\dots,K\}} N_k(r)$. When several arms are candidate, the one with the largest sum of rewards is chosen. If this is still not sufficient to obtain a unique arm, the leader is chosen at random among the arms maximizing both criteria. At round $r$, a subset $\mathcal{A}_r \subset \{1,\dots,K\}$ is selected by the learner based on the outcomes of the duels against $\ell(r)$. Next, all arms in $\mathcal{A}_r$ are drawn, yielding $Y_{k,N_k(r)}$ for $k \in \mathcal{A}_r$, where $N_k(r) = \sum_{s=1}^{r} \mathbb{1}(k \in \mathcal{A}_s)$.

We refer the reader to definition to Section 1.1 for the definition of the *regret* and the lower bounds of Lai and Robbins (1985) for parametric families of distributions.

**Abruptly changing environments** In Section 3.4, we consider abruptly changing environments that we introduced in Section 1.3. We recall that the number of breakpoints up to time $T$, denoted $\Gamma_T$, is defined by

$$\Gamma_T = \sum_{t=1}^{T-1} \mathbb{1}\{\exists k, \nu_{k,t} \neq \nu_{k,t+1}\},$$

and that the time instants $(t_1,\dots,t_{\Gamma_T})$ associated to these breakpoints define $\Gamma_T + 1$ stationary phases where the reward distributions are fixed. In such environments, letting $\mu_t^\star$ denote the best arm at time $t$, the performance of a policy is measured through the *dynamic regret* defined as

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T}(\mu_t^\star - \mu_{A_t})\right].$$

In the non-stationary case, the lower bound for the regret takes a different form: for any strategy, there exists an abruptly changing instance such that $\mathcal{R}_T = \Omega(\sqrt{T\Gamma_T})$ (Garivier and Moulines, 2011; Seznec et al., 2020). In this chapter, we only consider this type of non-stationary environments, and hence target this theoretical performance.

## 3.3 LB-SDA with Limited Memory in Stationary Environments

In this section we recall the general principle of LB-SDA and analyze a variant with *limited memory*, that we call LB-SDA-LM.

### 3.3.1 Last Block Sampling

We recall the general principle of SDA: at each round $r$, the algorithm (1) selects a leader $\ell(r)$, (2) makes this leader compete against every other arm (called challenger) in *duels* (pairwise comparisons), and (3) at the end of the round each winning challenger (if any) is pulled, otherwise the leader is pulled. This general principle is detailed in Algorithm 2.1, and in the following Algorithm 3.1 we simply recall the duel step of LB-SDA. In the rest of this chapter we assume that LB-SDA uses a forced exploration $f_r = \sqrt{\log(r)}$, as suggested in Chapter 2.

**Input:** 2 arms $\ell$ (leader) and $k$ (challenger), History $\mathcal{H}_\ell = (Y_{\ell,1}, \dots, Y_{\ell,N_\ell})$ and $\mathcal{H}_k = (Y_{k,1}, \dots, Y_{k,N_k})$, set of arms to pull $\mathcal{A}$.

Define $\bar{Y}_{k,N_k} := \frac{1}{N_k}\sum_{i=1}^{N_k} Y_{k,i}$ and $\bar{Y}_{\ell,N_\ell-N_k+1:N_\ell} := \frac{1}{N_k}\sum_{j=1}^{N_k} Y_{\ell,N_\ell-N_k+j}$

**if** $N_k \leq \sqrt{\log(r)}$ *or* $\bar{Y}_{k,N_k} \geq \bar{Y}_{\ell,N_\ell-N_k+1:N_\ell}$ ; ▷ `Using last block for the leader`

**then**

  Add $k$ to $\mathcal{A}$

**end**

**Algorithm 3.1:** Duel step of LB-SDA

We finally recall that LB-SDA works because it provides a *sufficient diversity* of sub-samples so that the challenger has a fair chance of winning. This is due to the fact that the leader will be pulled a linear number of times during a run of the algorithm. According to Lemma 2.16 and Definition 2.10, the performance of LB-SDA depends only on the family of distributions of the arms, that needs to satisfy the *balance condition* (Definition 2.11) to ensure a logarithmic regret. For example, *Single Parameter Exponential Families* (SPEF) satisfy this condition with $f_r = \sqrt{\log(r)}$, and LB-SDA is even *asymptotically optimal* in that case as the upper bound of the regret matches the lower bound of Lai and Robbins (1985).

In the rest of this chapter we present our results assuming that the arms come from the same SPEF, as our focus is on the mechanisms to ensure the performance of LB-SDA in two settings with limited memory. We refer the interested reader to Chapter 2 for discussions on the assumptions that can be made on the arms' distributions to ensure the performance of LB-SDA. All the results we are going to present are valid for distributions satisfying the more general Assumptions 2.4 and 2.20.

### 3.3.2 Memory-Limited LB-SDA

One of our main motivations for further studying LB-SDA is its simplicity and efficiency. Yet, all existing subsampling algorithms (Baransi et al., 2014; Chan, 2020; Baudry et al., 2020), including the vanilla version of LB-SDA (see Chapter 2), have to store the entire history of rewards for all the arms. In this section, we explain how to modify LB-SDA to reduce the storage cost while preserving the theoretical guarantees.

Consider a family of distributions for which LB-SDA has logarithmic regret. When $T$ is large, the arm with the largest mean is the leader with high probability, and all the challengers should have a number of pulls that is of order $\mathcal{O}(\log T)$ only. With duels based on the last block, this would mean in particular that only the last $\mathcal{O}(\log T)$ observations from the optimal arm should be stored and that very old observations will *never* be used again in practice. Based on this intuition, one might think that keeping only $\log(T)/(\mu^\star - \mu_k)^2$ observations is enough for LB-SDA. However, this could only be done with the knowledge of the true gaps.

We propose instead to limit the storage memory of each arm at round $r$ to a quantity

$$m_r = \Omega\left(\log(r)^{1+\gamma}\right) ,$$

for some $\gamma > 0$. Following the definition of Agrawal and Goyal (2012b), we then define the set of *saturated arms* at a round $r$ as

$$\mathcal{S}_r = \{k \in \{1, \dots, K\} : N_k(r) \geq m_r\} .$$

The only modification of LB-SDA is the following: at each round $r$, if a saturated arm is pulled then the newly collected observation replaces the oldest observation in its history. The following result shows that LB-SDA-LM keeps the same asymptotical performance as LB-SDA for $m_r$ satisfying $m_r/\log(r) \to +\infty$. We stated this result for SPEF for simplicity, but the guarantees under Assumption 2.20 translate similarly.

**Theorem 3.1** (Asymptotic optimality of LB-SDA with Limited Memory). *For any bandit model $\nu = (\nu_1, \dots, \nu_K) \subset \mathcal{P}_\Theta^K$ where $\mathcal{P}_\Theta$ is any single parameter exponential family of distributions satisfying Assumption 2.4, if $m_r/\log(r) \to \infty$, the regret of memory-limited LB-SDA satisfies, for all $\varepsilon > 0$,*

$$\mathcal{R}_T \leq \sum_{k:\mu_k<\mu^\star} \frac{1+\varepsilon}{\mathrm{kl}(\mu_k, \mu^\star)} \log(T) + C'(\nu, \varepsilon, \mathcal{M}) ,$$

*where $\mathcal{M} = (m_1, m_2, \dots, m_T)$ denotes the sequence $(m_r)_{r\in\mathbb{N}}$ and $C'(\nu, \varepsilon, \mathcal{M})$ is a problem-dependent constant.*

We detail the proof of this result, making explicit all the constant terms. Furthermore, we highlight that it also works for Lemma 2.7 for LB-SDA by setting $m_r = +\infty$ or $m_r = r$ in some parts of the proof (that we will point out).

*Proof.* We introduce a sequence $m_r$ of allowed memory for each arm at a round $r$. In the beginning of the proof we do not make any assumption on the sequence $(m_r)_{r\geq 1}$ except that $m_r/\log(r) \to +\infty$, which is required in the statement of Theorem 3.1. We further assume that $m_r$ is an integer for any round $r$, which does not change anything for the algorithm but simplifies the notation for the proof. Without loss of generality, we assume that the arm 1 is the unique optimal arm, $\mu_1 = \max_{k\in[K]} \mu_k$. We also recall that the arms are assumed to come from the same SPEF for simplicity, so Assumption 2.4 is satisfied for some rate functions $(I_k)_{k\in\{1,\dots,K\}}$. In terms of notation, we remark that if $N_k(r) \geq m_r$ and $\ell(r) \neq k$ then the duel between $k$ and $\ell(r)$ is the comparison between $\bar{Y}_{k,N_k(r)-m_r+1:N_k(r)}$ and $\bar{Y}_{\ell(r),N_{\ell(r)}(r)-m_r+1:N_{\ell(r)}(r)}$. Otherwise, if $N_k(r) \leq m_r$ and $\ell(r) \neq k$ then the duel is the comparison between $\bar{Y}_{k,N_k(r)}$ and $\bar{Y}_{\ell(r),N_{\ell(r)}(r)-N_k(r)+1:N_{\ell(r)}(r)}$, which is the same as for the vanilla LB-SDA.

We recall that the set of *saturated arms* at round $r$ is defined as $\mathcal{S}_r = \{k : N_k(r) \geq m_r\}$.

To simplify the notation for each arm $k$ we define the real number $x_k = \frac{\mu_1+\mu_k}{2} \in (\mu_k, \mu_1)$, and write $\omega_k = \min(I_1(x_k), I_k(x_k))$. Hence, we will write most of our results using concentration with this value $\omega_k$ for arm $k$.

Our first step decomposes the number of pulls of arm $k$ depending of if arm 1 is the leader or not, and if it is the case whether $k$ is saturated or not, which gives

$$\mathbb{E}[N_k(T)] \leq 1 + \underbrace{\mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(\ell(r) \neq 1)\right]}_{\mathcal{Z}_r} + \underbrace{\mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, k \notin \mathcal{S}_r, \ell(r) = 1)\right]}_{\mathcal{G}_r}$$
$$+ \underbrace{\mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, k \in \mathcal{S}_r, \ell(r) = 1)\right]}_{\bar{\mathcal{G}}_r}.$$

We first study $\bar{\mathcal{G}}_r$, and start the sum on the rounds at $2m_1$ because two arms cannot be saturated before this round is reached. We upper bound $\bar{\mathcal{G}}_r$ by taking a union bound on the sample size of each arm, and that if two random variables $(X, Y)$ satisfy $X > Y$ then either $X \geq \xi$ or $Y \leq \xi$ for any real value $\xi$. We then obtain

$$\bar{\mathcal{G}}_r \leq \sum_{r=2m_1}^{T-1} \mathbb{P}\left(\ell(r) = 1, k \in \mathcal{A}_{r+1}, N_k(r) \geq m_r, N_1(r) \geq m_r\right)$$

$$
\begin{aligned}
&\leq \sum_{r=2m_1}^{T-1} \mathbb{P}\left(N_1(r) \geq N_k(r) \geq m_r, \bar{Y}_{k,N_k(r)-m_r+1:N_k(r)} \geq \bar{Y}_{1,N_1(r)-m_r+1:N_1(r)}\right) \\
&\leq \sum_{r=2m_1}^{T-1} \sum_{n_k=m_r}^{r} \mathbb{P}\left(\bar{Y}_{k,n_k-m_r+1:n_k} \geq x_k, N_k(r) = n_k\right) \\
&\quad + \sum_{r=2m_1}^{T-1} \sum_{n_1=m_r}^{r} \mathbb{P}\left(\bar{Y}_{1,n_1-m_r+1:n_1} \leq x_k, N_1(r) = n_1\right) \\
&\leq 2 \sum_{r=2m_1}^{T-1} r e^{-m_r \omega_k} \ ,
\end{aligned}
$$

where we used the concentration inequality of Assumption 2.4, that is satisfied for SPEF.

We then consider $\mathcal{G}_r$ and distinguish two cases, whenever $N_k(r) \leq n_0(T)$ or not at each round, for some $n_0(T)$ that will be specified later. First, we have that

$$
\mathcal{G}_r \leq n_0(T) + \mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, k \notin \mathcal{S}_r, \ell(r) = 1, N_k(r) \geq n_0(T))\right] \ .
$$

We then use that on the event $k \notin \mathcal{S}_r$ the duels played between $k$ and 1 will be the classical duel with the last block: $k$ will compete with its empirical mean and 1 with the mean of its last block of size $N_k(r)$. We define some $\eta_k \in (\mu_k, \mu_1)$ and write

$$
\begin{aligned}
\mathcal{G}_r &\leq n_0(T) + \mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, k \notin \mathcal{S}_r, \ell(r) = 1, N_k(r) \geq n_0(T))\right] \\
&\leq n_0(T) + \mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, \bar{Y}_{k,N_k(r)} \geq \bar{Y}_{1,N_1(r)-N_k(r)+1:N_1(r)}, N_k(r) \geq n_0(T))\right] \ .
\end{aligned}
$$

We then use the same trick as for $\bar{\mathcal{G}}_r$ to separate the means of the two arms, before taking a union bound on the sample size for arm $k$, and use Lemma 2.6 for the leader. These steps give

$$
\mathcal{G}_r \leq n_0(T) + \sum_{n_k=n_0(T)}^{T-1} \mathbb{P}\left(\bar{Y}_{k,n_k} \geq \eta_k\right) + \sum_{n_k=n_0(T)}^{T-1} \sum_{n_1=n_0(T)}^{T-1} \mathbb{P}\left(\bar{Y}_{1,n_1-n_k+1:n_1} \leq \eta_k\right) \ .
$$

for any $\eta_k \in \mathbb{R}$. We remark that the sum on $r$ disappeared for arm $k$. Indeed, the fact that arm $k$ is pulled ensures that $\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, N_k(r) = n_k) \leq 1$. We finally obtain

$$\mathcal{G}_r \leq n_0(T) + \frac{e^{-n_0(T)I_k(\eta_k)}}{1-e^{-I_k(\eta_k)}} + T\frac{e^{-n_0(T)I_1(\eta_k)}}{1-e^{-I_1(\eta_k)}} \ .$$

We then calibrate $n_0(T)$ and $\eta_k$ in order to makes these terms converge properly. We define $\varepsilon > 0$ and state $n_0(T) = \frac{1+\varepsilon}{I_1(\mu_k)} \log T$. We then use the continuity of the rate functions on $(\mu_k, \mu_1)$ to state that for any $\varepsilon > 0$ there exists $\eta_k \in (\mu_k, \mu_1)$ satisfying $I_1(\eta_k) \geq \frac{I_1(\mu_k)}{1+\varepsilon}$. That translates in our result to $Te^{-n_0(T)I_1(\eta_k)} \leq 1$, and the remaining term is upper bounded by a constant that depends on $\varepsilon$ (typically, in $\mathcal{O}(\varepsilon^{-2})$). Hence, for any $\varepsilon > 0$ it holds that

$$\mathcal{G}_r \leq \frac{1+\varepsilon}{I_1(\mu_k)} \log T + C_{k,\varepsilon} \ ,$$

where $C_{k,\varepsilon}$ is a constant. Combining these results we can write a first upper bound on $\mathbb{E}[N_k(T)]$ as

$$\mathbb{E}[N_k(T)] \leq 1 + \frac{1+\varepsilon}{I_1(\mu_k)} \log T + 2\sum_{r=2m_1}^{T-1} re^{-m_r\omega_k} + C_{k,\varepsilon} + \sum_{r=2m_1}^{T-1} \mathbb{P}(\ell(r) \neq 1) \ . \tag{3.1}$$

We remark that without the memory limit (equivalently $m_r = +\infty$) the first sum vanishes. Indeed this expression provides an explicit dependence in $m_r$, that justifies the condition in Theorem 3.1 for $m_r$ ( namely, $m_r/(\log r) \to +\infty$) to make the sum converge.

We then work on upper bounding $\mathcal{Z}_r$. As in the proof of Chan (2020) this part is the most technically challenging. In the next steps we will consider the same events as in the original proof, but the storage limitation will add some complexity to the task. We first recall that at each round the leader satisfies

$$\ell(r) = k \Rightarrow N_k(r) \geq \left\lceil \frac{r}{K} \right\rceil \ .$$

However, adding the storage constraint we have that for any $r$ satisfying $r \geq Km_r$ the leader has necessarily more than $m_r$ observations. For this reason, its history will be always truncated to the $m_r$ last observations. For $r$ is reasonably large, $m_r$ is still large enough to guarantee a good concentration of the empirical mean of the saturated arms. We define $a_r = \lceil \frac{r}{4} \rceil$, and write

$$\mathbb{P}\left(\ell(r) \neq 1\right) = \mathbb{P}\left(\{\ell(r) \neq 1\} \cap \mathcal{D}^r\right) + \mathbb{P}\left(\{\ell(r) \neq 1\} \cap \bar{\mathcal{D}}^r\right) \ . \tag{3.2}$$

We define $\mathcal{D}^r$ the event under which the optimal arm has been leader at least once in $[a_r, r]$.

$$\mathcal{D}^r = \{\exists u \in [a_r, r] \text{ such that } \ell(u) = 1\}.$$

We then upper bound the term in the left hand side of Equation (3.2). As Chan (2020), we use that if arm 1 is not the leader and $\mathcal{D}_r$ holds, a *leadership takeover* happened (arm 1 was leader, arm $k$ becomes the next leader). In that case, some arm $k'$ had at some point a better empirical average as the leader while having the same sample size. We denote this event by $\mathcal{D}_r^{k'}$. For any $r_0 \in \mathbb{N}$ it holds that

$$\sum_{r=r_0}^{T-1} \mathbb{P}\left(\mathcal{D}_r^{k'}\right) \leq \mathbb{E}\left[\sum_{r=r_0}^{T-1}\sum_{u=a_r}^{r} \mathbb{1}(\bar{Y}_{1,N_1(u)} \leq \bar{Y}_{k',N_{k'}(r)}, N_1(u) = N_{k'}(u), N_1(u) \leq m_u)\right] + \mathbb{E}\left[\sum_{r=r_0}^{T-1}\sum_{u=a_r}^{r} \mathbb{1}(\bar{Y}_{1,N_1(u)-m_r:N_1(r)} \leq \bar{Y}_{k',N_{k'}(r)-m_r:N_{k'}(r)}, N_1(u) = N_{k'}(u), N_1(u) \geq m_u)\right] .$$

We remark that if $r_0$ is large enough such that for $r \geq r_0$ it holds that

$$\frac{a_r}{K} - 1 \leq m_{a_r} ,$$

then the first term is equal to 0, and the leader is necessarily saturated. It remains to upper bound the second term. Similarly as for the upper bound of $\bar{\mathcal{G}}_r$ a union bound on the sample sizes provides

$$\begin{aligned}\sum_{r=r_0}^{T-1} \mathbb{P}\left(\mathcal{D}_r^{k'}\right) &\leq \mathbb{E}\left[\sum_{r=r_0}^{T-1}\sum_{u=a_r}^{r} \mathbb{1}\left(N_1(u) = N_{k'}(u), \bar{Y}_{1,N_1(r)-m_r+1:N_1(r)} \leq \bar{Y}_{k',N_{k'}(r)-m_r+1:N_{k'}(r)}\right)\right] \\ &\leq r_0 + \sum_{r=r_0}^{T-1}\sum_{u=\max(a_r,2m_1)}^{r} 2ue^{-m_u\omega_{k'}} \\ &\leq r_0 + 2\sum_{r=r_0}^{T-1} r^2 e^{-m_{a_r}\omega_{k'}} .\end{aligned}$$

We first use this result without commenting its dependence in the sequence $(m_r)_{r\geq 1}$. Summing on all suboptimal arms $k'$ we obtain

$$\sum_{r=r_0}^{T-1} \mathbb{P}\left(\{\ell(r) \neq 1\} \cap \mathcal{D}^r\right) \leq 2\sum_{k'=2}^{K}\sum_{r=r_0}^{T-1} r^2 e^{-m_{a_r}\omega_{k'}} . \tag{3.3}$$

Again, the constraint $m_r/\log(r) \to +\infty$ is sufficient to ensure a proper convergence of this sum to a constant with the same arguments as before, because $a_r$ is still linear in $r$.

**Remark 3.2.** *In that case we can again set $m_r = +\infty$ to obtain the proof without limited memory, so this last term would be 0. However, our argument to discard the "unsaturated leader" case is not valid anymore. Deriving an upper bound for this term would give $2\sum_{k'=2}^{K}\sum_{r=1}^{T-1} e^{-b_r\omega_{k'}}$ instead, which is smaller.*

*Hence, the limited memory makes sub-optimal leadership takeovers more likely (which is intuitive).*

We then consider the event for which arm 1 has never been the leader between $a_r$ and $r$. The idea in this part is to leverage the fact that if the optimal arm is not leader between $\lfloor r/4 \rfloor$ and $r$, then it has necessarily lost a lot of duels against the current leader at each round. We then use the fact that when the leader has been drawn "enough", concentration prevents this situation with large probability. We introduce

$$\mathcal{L}^r = \sum_{u=a_r}^{r} \mathbb{1}(\mathcal{C}^u) \ ,$$

with $\mathcal{C}^u$ defined as $\mathcal{C}^u = \{\exists k \neq 1, \ell(u) = k, 1 \notin \mathcal{A}_{u+1}\}$. The following holds

$$\mathbb{P}(\ell(r) \neq 1 \cap \overline{\mathcal{D}}^r) \leq \mathbb{P}(\mathcal{L}^r \geq r/4) \ . \tag{3.4}$$

This result comes from (Chan, 2020), along with the direct use of the Markov inequality to provide the upper bound

$$\mathbb{P}(\mathcal{L}^r \geq r/4) \leq \frac{\mathbb{E}(\mathcal{L}^r)}{r/4} = \frac{4}{r}\sum_{u=a_r}^{r} \mathbb{P}(\mathcal{C}^u) \ . \tag{3.5}$$

We further decompose $\mathbb{P}(\mathcal{C}^u)$ in two parts according to the number of selections of arm 1. For a constant $C > 0$, we write

$$\begin{aligned}\sum_{r=r_0}^{T-1} \mathbb{P}(\{\ell(r) \neq 1\} \cap \overline{\mathcal{D}}^r) \leq & \underbrace{\sum_{r=r_0}^{T-1} \frac{4}{r}\sum_{u=a_r}^{r} \mathbb{P}\left(N_1(u) \leq \frac{C}{4}\log(u)\right)}_{\bar{D}_1} \\ & + \underbrace{\sum_{r=r_0}^{T-1} \frac{4}{r}\sum_{u=a_r}^{r} \mathbb{P}\left(\mathcal{C}^u, N_1(u) \geq \frac{C}{4}\log(u)\right)}_{\bar{D}_2} \ .\end{aligned}$$

We consider $\bar{D}_2$. Again, we decompose it according to if the optimal arm is saturated or not. We also introduce $\mathcal{C}_k^u = \{\ell(u) = k, 1 \notin \mathcal{A}_{u+1}\}$ for any $k \in \{2, \dots, K\}$. We first upper bound

$$\begin{aligned}
\bar{D}_{k,1} &\coloneqq \sum_{r=r_0}^{T-1} \frac{4}{r} \sum_{u=a_r}^{r} \mathbb{P}\left(\mathcal{C}_k^u, N_1(u) \geq \frac{C}{4}\log(u), 1 \in \mathcal{S}_u\right) \\
&\leq \sum_{r=r_0}^{T-1} \frac{8}{r} \sum_{u=a_r}^{r} u e^{-m_{a_u}\omega_k} \\
&\leq 8 \sum_{r=r_0}^{T-1} \sum_{u=a_r}^{r} e^{-m_{a_u}\omega_k} \\
&\leq 8 \sum_{r=r_0}^{T-1} r e^{-m_{a_r}\omega_k} \ .
\end{aligned}$$

Then, we consider the alternative

$$\begin{aligned}
\bar{D}_{k,2} &\coloneqq \sum_{r=r_0}^{T-1} \frac{4}{r} \sum_{u=a_r}^{r} \mathbb{P}\left(\mathcal{C}_k^u, N_1(u) \geq \frac{C}{4}\log(u), 1 \notin \mathcal{S}_u\right) \\
&\leq \sum_{r=r_0}^{T-1} \frac{4}{r} \sum_{u=a_r}^{r} \mathbb{P}\left(\bar{Y}_{k,N_k(u)-N_1(u)+1:N_k(u)} > \bar{Y}_{1,N_1(u)}, N_k(u_) \geq N_1(u) \geq \frac{C}{4}\log(u), 1 \notin \mathcal{S}_u\right) \\
&\leq \sum_{r=r_0}^{T-1} \frac{4}{r} \left[\frac{1}{1-e^{-\omega_k}} e^{-\frac{C}{4}\log(a_r)\omega_k} + \frac{r}{1-e^{-\omega_k}} e^{-\frac{C}{4}\log(a_r)\omega_k}\right] \\
&\leq \sum_{r=r_0}^{T-1} \frac{4(r+1)}{r(1-e^{-\omega_k})} e^{-\frac{C}{4}\log(a_r)\omega_k} \\
&\leq \sum_{r=r_0}^{T-1} \frac{6}{1-e^{-\omega_k}} e^{-\frac{C}{4}\log(a_r)\omega_k} \ .
\end{aligned}$$

So finally,

$$\bar{D}_2 \leq \sum_{k=2}^{K} \left[8 \sum_{r=r_0}^{T-1} r e^{-m_{a_r}\omega_k} + \sum_{r=r_0}^{T-1} \frac{6}{1-e^{-\omega_k}} e^{-\frac{C}{4}\log(a_r)\omega_k}\right] \ .$$

At this step we need to choose the constant $C$ large enough in order to make this sum converge, which is possible since $C$ is only a parameter of the analysis. Furthermore, the first sum vanishes if we set $m_r = +\infty$, and again converges if $m_r/\log(r) \to +\infty$.

We then consider the term $\bar{D}_1$. We transform the double sum in a simple sum by simply counting the number of times each term is included.,

$$\bar{D}_1 = \sum_{r=r_0}^{T} \frac{4}{r} \sum_{u=a_r}^{r} \mathbb{P}\left(N_1(u) \leq \frac{C}{4}\log(u)\right) = \sum_{r=r_0}^{T} \left(\sum_{t=a_{r_0}}^{r} \frac{4}{t} \mathbb{1}(t \in [r, 4r])\right) \mathbb{P}\left(N_1(r) \leq \frac{C}{4}\log(r)\right) \ .$$

If we remark that $\sum_{t=1}^{r} \frac{4}{t}\mathbb{1}(t \in [r, 4r]) \leq 4\log(4r/(r-1)) \leq 9$ for $r \geq 2$, we finally get:

$$\sum_{r=r_0}^{T} \mathbb{P}(\{\ell(r) \neq 1\} \cap \overline{\mathcal{D}}^r) \leq r_0 + 9\sum_{r=r_0}^{T} \mathbb{P}\left(N_1(r) \leq \frac{C}{4}\log(r)\right) + \mathcal{O}(1). \tag{3.6}$$

Combining (3.3) and (3.6) yields

$$\sum_{r=r_0}^{T} \mathbb{P}\left(\ell(r) \neq 1\right) \leq r_0 + 9\sum_{r=r_0}^{T} \mathbb{P}\left(N_1(r) \leq \frac{C}{4}\log(r)\right) + \mathcal{O}(1)\ .$$

Hence, the storage limit may introduce larger constant terms in the proof, but asymptotically the dominant terms are the same as in the proof of the vanilla LB-SDA algorithm.

**Remark 3.3.** *This step concludes the proof of Lemma 2.7 for LB-SDA by setting $m_r = +\infty$ when the memory limit is used. Note that the proof only requires slight adaptations for any block sampler. Details can be found in (Baudry et al., 2020).*

The last step is to show that we can upper bound the remaining term with the same results as the ones we used in Chapter 2. We only need to prove that if $r_0$ is large enough and $\{N_1(r) \leq C/4\log(r)\}$, then the arm 1 has not been saturated for a long time. Indeed, in that case the saturation would have no impact in the way to upper bound this term. Defining $m^{-1}(x) = \inf\{r : m_r \geq x\}$, and knowing that if $C$ and $r \geq r_0$ (we can increase $r_0$) are large enough $m_r \geq C\log r$ and so $m^{-1}(x) \leq \exp(x/C)$, we have $m^{-1}(C/4\log r) \leq \exp(C/4\log(r)C^{-1}) = r^{1/4}$.

Hence, after the round $r_0$ we are sure that arm 1 has never been saturated since the round $r^{1/4}$, but also that the sub-sample required by LB-SDA-LM at each step will not be altered by the memory limit too. In this scenario LB-SDA-LM does exactly the same as LB-SDA. As LB-SDA ensures sufficient diversity (Lemma 2.16) we conclude that if the arms come from the same SPEF,

$$\sum_{r=r_0}^{T} \mathbb{P}\left(N_1(r) \leq \frac{C}{4}\log(r)\right) = \mathcal{O}(1)\ .$$

□

### 3.3.3 Storage and Computational Cost

To the best of our knowledge, LB-SDA-LM is the only bandit algorithm based on the idea of sub-sampling that does not require to store the full history of rewards. In Table 3.1 we compare estimates of the computational cost of LB-SDA-LM and the other algorithms based on sub-sampling. We consider best and worst cases since SSMC and LB-SDA can be updated se-

quentially when the leader does not change and is pulled. The estimates we provide correspond to the cost of a single step of the algorithm at time $T$.

**Table 3.1** – Memory and computational costs at round $T$ for existing subsampling algorithms.

| Algorithm | Memory | Computational cost Best-Worst case |
|---|---|---|
| BESA Baransi et al. (2014) | $\mathcal{O}(T)$ | $\mathcal{O}((\log T)^2)$ |
| SSMC Chan (2020) | $\mathcal{O}(T)$ | $\mathcal{O}(1)$-$\mathcal{O}(T)$ |
| RB-SDA Baudry et al. (2020) | $\mathcal{O}(T)$ | $\mathcal{O}(\log T)$ |
| LB-SDA **(this chapter)** | $\mathcal{O}(T)$ | $\mathcal{O}(1)$-$\mathcal{O}(\log T)$ |
| LB-SDA-LM **(this chapter)** | $\mathcal{O}((\log T)^2)$ | $\mathcal{O}(1)$-$\mathcal{O}(\log T)$ |

**Efficient updates for LB-SDA** Let us detail the possible scenarios for a given leader and challenger. Assume that at round $r$ the leader was using the sample mean $\bar{Y}_{N-n+1:N}$ for some $(n, N) \in \mathbb{N}^2$. If it is pulled then $Y_{N+1}$ is collected, and at next round it will use $\bar{Y}_{N-n+2:N+1}$ against the same challenger. An efficient update consists in computing

$$\bar{Y}_{N-n+2:N+1} = \bar{Y}_{N-n+1:N} + \frac{1}{n}(Y_{N+1} - Y_{N-n+1}) \ ,$$

which comes at almost no cost. Furthermore, an efficient update can also be performed if the challenger is pulled. In that case, the leader is not pulled and needs to use $\bar{Y}_{N-n:N}$, that can be computed as $\bar{Y}_{N-n:N} = \frac{n}{n+1}\bar{Y}_{N-n+1:N} + \frac{1}{n+1}Y_n$. Hence, the most costly scenario is when the leader changes, and new sub-sample means have to be computed. However, our proof shows that the number of changes of leadership is expected to be finite if there a single best arm.

**Updating SSMC** We can compare these results to possible updates of SSMC. We denote by $Y_r^- = \min_{j\in[N-n+1]} \bar{Y}_{j:j+n-1}$ the sub-sample mean used at a given round against some challenger. In the first scenario we described (when the leader is pulled), we need to compute $Y_{r+1}^- = \min\{Y_r^-, \bar{Y}_{N-n+2:N+1}\}$ at next round. If the value of $Y_r^-$ and of the sub-sample mean of the last block (as for LB-SDA) are kept in memory, then the update costs the same as for

LB-SDA: only the last block can change the outcome of the duel. In the second and third scenario, however, the leader needs to perform a screening of its entire history. This comes at a linear cost in terms of the sample size of the leader, which is itself linear in the round $r$. Contrarily to the number of leadership changes, the number of rounds when a challenger is pulled is expected to be logarithmic in the time horizon.

**Details for other algorithms** The computational cost can be broken into two parts: (a) the sub-sampling cost and (b) the computation of the means of the samples. We assume that drawing a sample of size $n$ without replacement has $\mathcal{O}(n)$ cost (independently of the size of the set) and that computing the mean of this sub-sample costs another $\mathcal{O}(n)$. Furthermore, at round $T$, each challenger to the best arm has about $\mathcal{O}(\log T)$ samples. This gives an estimated cost of $O\left((\log T)^2\right)$ for BESA (Baransi et al., 2014). For RB-SDA (see Chapter 2) the estimated cost is $\mathcal{O}(\log(T))$, because the sampling cost for random block sampling is $\mathcal{O}(1)$ but a sub-sample mean is recomputed at each round.

For the three deterministic algorithms (namely SSMC (Chan, 2020), LB-SDA, LB-SDA-LM), when the leader arm wins all its duels, we assume that the sequential update costs $\mathcal{O}(1)$. This is the *best case* in terms of computational cost. However, we detailed that in some cases SSMC requires a full screening of the leader's history, with $\mathcal{O}(T)$ cost, while LB-SDA and LB-SDA-LM need at most the computation of the mean of the last $\mathcal{O}(\log T)$ samples from the leader.

## 3.4 LB-SDA in Non-Stationary Environments

In stationary environments, LB-SDA achieves optimal regret rates, even when its decisions are constrained to use at most (for instance) $O((\log T)^2)$ observations. One might think that this argument itself is sufficient to address non-stationary scenarios as the duels are performed mostly using recent observations. For instance, if the distribution of the best arm changes we can hope that LB-SDA will "adapt" to this change relatively fast. However, there are other cases where LB-SDA is not sufficient. For instance, if an arm has been bad for a long period of time and suddenly becomes the best arm, adapting to the change would still be prohibitively slow. For this reason, LB-SDA has to be equipped with an additional mechanism to perform well in non-stationary environments.

### 3.4.1 SW-LB-SA: LB-SDA with a Sliding-Window

We keep a *round-based* structure for the algorithm, where, at each round $r$, duels between arms are performed and the algorithm subsequently selects the subset of arms $\mathcal{A}_r$ that will be pulled. In contrast to Section 3.3.2, where we introduced a constraint on the number of observations

kept in memory for each arm, we propose here to use a sliding window of length $\tau$ to limit the historical data available to the algorithm to that of the last $\tau$ rounds. We highlight the fundamental difference between the two approaches: the sliding window will make LB-SDA forget old observations for all arms, whether they have been sampled a lot or not.

**Modified leader definition** The introduction of a sliding window requires a new definition for the *leader*. By analogy with the stationary case, the leader could be defined as the arm that has been pulled the most during the $\tau$ last rounds. However, with the inclusion of the sliding window, a new phenomenon, which we call *passive leadership takeover*, can occur. Let us define $N_k^\tau(r) = \sum_{s=r-\tau}^{r-1} \mathbb{1}\left(k \in \mathcal{A}_{s+1}\right)$, the number of times arm $k$ has been pulled during the last $\tau$ rounds and consider a situation with 3 arms $\{1, 2, 3\}$. Assume that the leader is arm 1 and at a round $(r-1)$ we have $N_1^\tau(r-1) = N_2^\tau(r-1)$. If the leader has been pulled $\tau$ rounds away and wins its duel against arm 2 but looses against arm 3, only arm 3 will be pulled at round $r$. Consequently, at round $r$, arm 2 will have a strictly larger number of pulls than arm 1 without having actually defeated the leader. This situation, illustrated on Figure 3.1, is not desirable. We fix this by imposing that any arm has to defeat the current leader to become the leader itself. Let us define

$$\mathcal{B}_r = \{k \in \mathcal{A}_{r+1} \cap \{N_k^\tau(r+1) \geq \min(r, \tau)/K\}\} \ .$$

Then for any $r \in \mathbb{N}$, we propose a new definition of the leader as round $r+1$ as

$$\ell^\tau(r+1) = \begin{cases} \text{argmax}_{k \in \{1,\dots,K\}} N_k^\tau(r+1) \text{ , if } N_{\ell^\tau(r)}^\tau(r+1) < \min(r, \tau)/(2K) \ . \\ \text{argmax}_{k \in \mathcal{B}_r \cup \{\ell^\tau(r)\}} N_k^\tau(r+1) \text{ otherwise.} \end{cases}$$

This modified definition of the leader ensures that an arm can become the leader only after earning at least $\tau/K$ samples and winning a duel against the current leader, or if the leader loses so many duels that its number of samples falls under a fixed threshold. Thanks to this definition it always holds that $N_{\ell^\tau(r)}^\tau(r) \geq \min(r, \tau)/(2K)$.

**Additional diversity flags** As in the vanilla LB-SDA, we use a sampling obligation to ensure that each arm has a minimal number of samples. However, in contrast to the stationary case, this very limited number of forced samples may not be sufficient to guarantee an adequate variety of duels, due to the forgetting window. To this end, the sampling obligation is coupled with a *diversity flag*. We define it as a binary random variable $D_k^\tau(r)$, satisfying $D_k^\tau(r) = 1$ only when, for the last $\lceil (K-1)(\log \tau)^2 \rceil$ rounds the three following conditions are satisfied: 1) some arm $k' \neq k$ has been leader during all these rounds, 2) $k'$ has not been pulled, and 3) $k$ has not been pulled and satisfy $N_k^\tau(r) \leq (\log \tau)^2$. In practice, there is a very low probability that these conditions are met simultaneously but this additional mechanism is required for the theoretical analysis. Note that the diversity flags have no impact on the computational cost

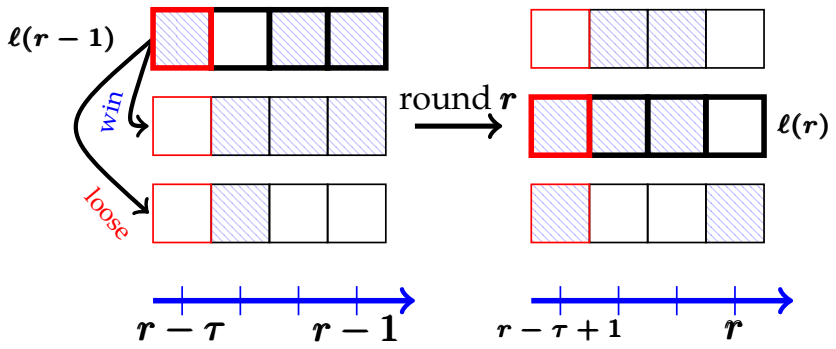


**Figure 3.1** – Illustration of a *passive leadership takeover* with a sliding window $\tau = 4$ when the standard definition of leader is used. The bold rectangle correspond to the leader. A blue square is added when an arm has an observation for the corresponding round and the red square correspond to the information that will be lost at the end of the round due to the sliding window.

of the algorithm as they require only to store the number of rounds since the last draw of the different arms (which can be updated recursively) as well as the last leader takeover. Arms that raise their diversity flag are automatically added to the set of pulled arms.

Bringing these parts together and keeping a forced exploration $f_\tau = \sqrt{\log(\tau)}$ gives the pseudo-code of SW-LB-SDA in Algorithm 3.2.

### 3.4.2 Regret Analysis in Abruptly Changing Environments

In this section we aim at upper bounding the dynamic regret in abruptly changing environments, as defined in Section 3.2. Our main result is the proof that the regret of SW-LB-SDA matches the asymptotic lower bound of Garivier and Moulines (2011).

**Theorem 3.4** (Upper bound on the dynamic regret of SW-LB-SDA). *If the time horizon $T$ and number of breakpoints $\Gamma_T$ are known, choosing $\tau = \mathcal{O}(\sqrt{T\log(T)/\Gamma_T})$ ensures that the dynamic regret of SW-LB-SDA satisfies*

$$\mathcal{R}_T = \mathcal{O}\left(\sqrt{T\Gamma_T \log T}\right) ,$$

*if Assumptions 2.4 and 2.20 are satisfied during each stationary phase.*

To prove this result we need to assume that, during each stationary period, the rewards satisfy Assumptions 2.4 (concentration of the means with exponential decay) and 2.20 (dominant left tail for the best arm). In contrast, current state-of-the-art algorithms for non-stationary bandits typically require the assumption that the rewards are *bounded* to obtain similar guarantees.

Hence, this result is of particular interest for tasks involving unbounded reward distributions. SW-LB-SDA can also be used for general bounded rewards with the same guarantees by using the *binarization trick* (Agrawal and Goyal, 2013b). However, the knowledge of the horizon $T$ and the estimated number of change point $\Gamma_T$ is still required to obtain optimal rates, and removing this assumption is an interesting direction for future works, for instance inspired by (Auer et al., 2019; Besson et al., 2022). In the following we provide a high-level outline of the analysis of Theorem 3.4. Our objective is to provide an intuition on the dominant terms of the regret, and on the interest of the additional mechanisms introduced compared to LB-SDA.

**Proof sketch** For the $\Gamma_T + 1$ stationary phases $[t_\phi, t_{\phi+1} - 1]$ with $\phi \in \{1, \dots, \Gamma_T\}$, we define $r_\phi$ as the first round where an observation from the phase $\phi$ was pulled. Introducing the gaps $\Delta_k^\phi = \mu_{t_\phi}^* - \mu_{t_\phi,k}$ and denoting the optimal arm $k_\phi^\star$, we can rewrite the regret as

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{\phi=1}^{\Gamma_T} \sum_{r=r_\phi-1}^{r_{\phi+1}-2} \sum_{k \neq k_\phi^\star} \mathbb{1}\left(k \in \mathcal{A}_{r+1}\right) \Delta_k^\phi\right] = \sum_{\phi=1}^{\Gamma_T} \sum_{k \neq k_\phi^*} \mathbb{E}[N_k^\phi] \Delta_k^\phi \,,$$

where we define $N_k^\phi = \sum_{r=r_\phi-1}^{r_{\phi+1}-2} \mathbb{1}(k \in \mathcal{A}_{r+1})$ as the number of pulls of arm $k$ during a phase $\phi$ when it is suboptimal. We highlight that the sequence $(r_\phi)_{\phi \geq 1}$ is a random variable that depends on the trajectory of the algorithm, but this causes no additional difficulty for upper bounding the regret. We introduce $\delta_\phi = t_{\phi+1} - t_\phi$ the length of a phase $\phi$. Using Lemma 25 from Garivier and Moulines (2011), we obtain the following result.

**Proposition 3.5.** *Consider a phase $\phi$ and define $A_k^{\phi,\tau} = b_k^\phi \log(\tau)$ for some constant $b_k^\phi > 0$. The expected number of pulls of a sub-optimal arm $k$ can be upper bounded as*

$$\mathbb{E}[N_k^\phi] \leq 2\tau + \frac{\delta_\phi A_k^{\phi,\tau}}{\tau} + c_{k,1}^{\phi,\tau} + c_{k,2}^{\phi,\tau} + c_{k,3}^{\phi,\tau} \,,$$

*where denoting by $D_k^\tau(r) \in \{0, 1\}$ the variable stating if the diversity flag is raised or not we defined*

$$c_{k,1}^{\phi,\tau} = \mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(k \in \mathcal{A}_{r+1}, \ell^\tau(r) = k_\phi^*, N_k^\tau(r) \geq A_k^{\phi,\tau}, D_k^\tau(r) = 0\right)\right] \,,$$

$$c_{k,2}^{\phi,\tau} = \mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(\ell^\tau(r) = k_\phi^*, D_k^\tau(r) = 1\right)\right] \,,$$

$$c_{k,3}^{\phi,\tau} = \mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(\ell^\tau(r) \neq k_\phi^*\right)\right] \,.$$

Though the notation is more complicated these three terms are actually quite similar to the decomposition in the stationary case. Furthermore, our objective will be to prove that they are actually not dominant in the regret bound. Re-defining the notion of *saturated arm* of previous section as each arm $k$ that have been sampled more than the threshold $A_k^{\phi,\tau}$ it holds that

- $c_{k,1}^{\phi,\tau}$ is an upper bound on the expected number of times a *saturated sub-optimal arm* can defeat $k_\phi^*$ while $k_\phi^*$ is the current leader.
- $c_{k,2}^{\phi,\tau}$ is an upper bound on the regret that can be caused by the diversity flag.
- $c_{k,3}^{\phi,\tau}$ is, as $\mathbb{E}\left[\sum_{r=1}^{T-1}\mathbb{1}(\mathcal{Z}_r)\right]$ in the previous section an upper bound of the regret that can be caused by the leader being a sub-optimal arm.

The three terms have hence intuitive interpretation and summarize well the technical contributions behind Theorem 3.4. We now introduce the novel concentration result for SW-LB-SDA that will allow us to upper bound these terms.

**Lemma 3.6.** *We consider a stationary phase $\phi$ and a MAB $(\nu_1^\phi,\dots,\nu_K^\phi)$. Let $k_\phi^*$ denote the arm with the largest mean. We assume that each arm $\nu_k^\phi$ satisfies Assumption 2.4 for some rate function $I_k^\phi$. Then, for any constant $n\in\mathbb{N}$ satisfying $n\geq f_\tau=\sqrt{\log\tau}$, by letting $\tilde{n}=\min(n,\lfloor\tau/(2K)\rfloor)$ under SW-LB-SDA it holds that*

$$\mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2}\mathbb{1}\left(k\in\mathcal{A}_{r+1},\ell^\tau(r)=k_\phi^*,N_k^\tau(r)\geq n,D_k^\tau(r)=0\right)\right]\leq\delta_\phi(\tau+1)\frac{e^{-\tilde{n}\omega_k^\phi}}{1-e^{-\omega_k^\phi}}\,, \tag{3.7}$$

*where we defined $\omega_k^\phi=\min\left(I_k^\phi\left(\frac{1}{2}(\mu_k^\phi+\mu_{k_\phi^*}^\phi)\right),I_{k_\phi^*}^\phi\left(\frac{1}{2}(\mu_k^\phi+\mu_{k_\phi^*}^\phi)\right)\right)$, and $\delta_\phi$ is the length of the phase and $\tau$ the size of the sliding window. Similarly,*

$$\mathbb{E}\left[\sum_{r=r_\phi+\tau-2}^{r_{\phi+1}-2}\mathbb{1}\left(k_\phi^*\notin\mathcal{A}_{r+1},\ell^\tau(r)=k,N_{k_\phi^*}^\tau(r)\geq n\right)\right]\leq\delta_\phi(\tau+1)\frac{e^{-\tilde{n}\omega_k^\phi}}{1-e^{-\omega_k^\phi}}\,. \tag{3.8}$$

*Proof.* We start with the first claim. Under the considered event, ($n\geq f(\tau)$ and $D_k^\tau(r)=0$) arm $k$ can be drawn only if it has won its duel against $k_\phi^*$. The duel itself is a comparison between the mean of two blocks containing at least $\tilde{n}=\min(n,\tau/(2K))$ observations because of the definition of the leader in this part. For any constant $\xi_k$, we have either $\widehat{\mu}_k^\tau(r)\geq\xi_k$ or $\widehat{\mu}_{\ell,k}^\tau(r)\leq\xi_k$, where $\widehat{\mu}_k^\tau(r)$ and $\widehat{\mu}_{\ell,k}^\tau(r)$ denote respectively the mean used by arm $k$ and arm $\ell$ in

their duel at round $r$. For the sake of simplicity we choose $\xi_k = \frac{1}{2}(\mu_k^\phi + \mu_{k_\phi^*}^\phi)$. We then write

$$\begin{aligned} A &\coloneqq \mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(k \in \mathcal{A}_{r+1}, \ell^\tau(r) = k_\phi^*, N_k^\tau(r) \geq n, D_k^\tau(r) = 0\right)\right] \\ &\leq \mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(k \in \mathcal{A}_{r+1}, \widehat{\mu}_{k_\phi^*,k}^\tau(r) \leq \xi_k, N_{k_\phi^*}^\tau(r) \geq \tau/(2K), N_k^\tau(r) \geq n\right)\right] \\ &\quad + \mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(k \in \mathcal{A}_{r+1}, \widehat{\mu}_k^\tau(r) \geq \xi_k, N_{k_\phi^*}^\tau(r) \geq \tau/(2K), N_k^\tau(r) \geq n\right)\right] . \end{aligned}$$

Assuming rewards are sequentially collected in a stream $Y_{k,1}^\phi, \ldots, Y_{k,\widehat{\Delta}_\phi}^\phi$ for a given arm $k$, all possible blocks of observations are uniquely described by two quantities: $N_k^\phi(r)$ (number of data collected since the beginning of phase $\phi$) and $N_k^\tau(r)$ (number of data collected in the last $\tau$ rounds). We will use this property to bound the two sums.

We start by the term featuring the arm $k$, and introduce

$$S_k^{n,m}(r) = \{k \in \mathcal{A}_{r+1},\ \widehat{\mu}_k^\tau(r) \geq \xi_k,\ N_k^\phi(r) = m + n - 1, N_k^\tau(r) = n\} ,$$

that allows to use that

$$\{k \in \mathcal{A}_{r+1}, \widehat{\mu}_k^\tau(r) \geq \xi_k, N_k^\tau(r) \geq n\} \subset \bigcup_{n_k=n}^{\hat{\delta}_\phi} \bigcup_{m_k=1}^{\hat{\delta}_\phi} S_k^{n_k,m_k}(r) . \tag{3.9}$$

Furthermore, the same block (same value for both $n$ and $m$) can not be used for upcoming rounds because the total count will be incremented. More specifically, for the arm $k$ for any possible block there is at most one round for which the indicator function can be 1., i.e.

$$\sum_{n_k=n}^{\hat{\delta}_\phi} \sum_{m_k=1}^{\hat{\delta}_\phi} \sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(S_k^{n_k,m_k}(r)\right) \leq \sum_{n_k=n}^{\hat{\delta}_\phi} \sum_{m_k=1}^{\hat{\delta}_\phi} \mathbb{1}\left(\bar{Y}_{k,m_k:m_k+n_k-1} \geq \xi_k\right) .$$

Similarly, we denote $Y_{k_\phi^\star,1}, ..., Y_{k_\phi^\star,\hat{\delta}_\phi}$ the set of possible rewards for the arm $k_\phi^\star$ and let

$$S_{k_\phi^\star}^{n,m}(r) = \{k \in \mathcal{A}_{r+1},\ \widehat{\mu}_{k_\phi^\star,k}^\tau(r) \leq \xi_k,\ N_{k_\phi^*}^\phi(r) = m + n - 1, N_{k_\phi^*}^\tau(r) = n\} .$$

We also have

$$\{k \in \mathcal{A}_{r+1}, \widehat{\mu}_{k_\phi^\star,k}^\tau(r) \leq \xi_k, N_{k_\phi^\star}^\tau(r) \geq n'\} \subset \bigcup_{n^\star=n'}^{\hat{\delta}_\phi} \bigcup_{m^\star=1}^{\hat{\delta}_\phi} S_{k_\phi^\star}^{n^\star,m^\star}(r) . \tag{3.10}$$

The main difference here is that several rounds can use the same block of observations of $k_\phi^\star$. This can be explained because when the indicator function equals 1 the arm $k$ is drawn instead of $k_\phi^\star$ and the previous argument do not hold anymore. Yet, $N_{k_\phi^*}^\tau(r)$ can not remain unchanged for more than $\tau$ steps because of the sliding window. This implies in particular,

$$\sum_{n^\star=n'}^{\hat{\delta}_\phi}\sum_{m^\star=1}^{\hat{\delta}_\phi}\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2}\mathbb{1}(S_{k_\phi^\star}^{n^\star,m^\star}(r)) \leq \tau\sum_{n^\star=n'}^{\hat{\delta}_\phi}\sum_{m^\star=1}^{\hat{\delta}_\phi}\mathbb{1}\left(\bar{Y}_{k_\phi^*,m^\star:m^\star+n^\star-1}\leq\xi_k\right) .$$

Bringing things together and applying the previous inequality with $n' = \lfloor\tau/(2K)\rfloor$ we obtain

$$A \leq \mathbb{E}\left[\sum_{m^\star=1}^{\hat{\delta}_\phi}\sum_{n^\star=n'}^{\hat{\delta}_\phi}\tau\mathbb{1}\left(\bar{Y}_{k_\phi^*,m^\star:m^\star+n^\star-1}\leq\xi_k\right)+\sum_{m_k=1}^{\hat{\delta}_\phi}\sum_{n_k=n}^{\hat{\delta}_\phi}\mathbb{1}\left(\bar{Y}_{k,m_k:m_k+n_k-1}\geq\xi_k\right)\right] .$$

We then have to handle carefully the fact that $\widehat{\delta}_\phi$ is a random variable depending on the bandit algorithm. Indeed, as several arms can be pulled at each round we don't know what will be the length of a phase in terms of rounds. However, this quantity is upper bounded by the length of the phase in terms of arms pulled $\delta_\phi$.

Thus, using the concentration inequality corresponding to the family of distributions for an appropriate rate function we can write

$$\begin{aligned}
A &\leq \sum_{m^\star=n}^{\delta_\phi}\sum_{n^\star=n'}^{\delta_\phi}\tau\mathbb{P}\left(\bar{Y}_{k_\phi^*,m^\star:m^\star+n^\star-1}\leq\xi_k\right)+\sum_{m_k=1}^{\delta_\phi}\sum_{n_k=n}^{\delta_\phi}\mathbb{P}\left(\bar{Y}_{k,m_k:m_k+n_k-1}\geq\xi_k\right)\\
&\leq \sum_{m^\star=1}^{\delta_\phi}\sum_{n^\star=n'}^{\delta_\phi}\tau e^{-n^\star I_{k_\phi^*}(\xi_k)}+\sum_{m_k=n}^{\delta_\phi}\sum_{n_k=n}^{\delta_\phi}e^{-n_k I_k(\xi_k)}\\
&\leq \delta_\phi\left(\tau\frac{e^{-n' I_{k_\phi^*}(\xi_k)}}{1-e^{-I_{k_\phi^*}(\xi_k)}}+\frac{e^{-nI_k(\xi_k)}}{1-e^{-I_k(\xi_k)}}\right)\\
&\leq \delta_\phi(\tau+1)\frac{e^{-\widetilde{n}\omega_k}}{1-e^{-\omega_k}} ,
\end{aligned}$$

where in the last inequality we have introduced $\tilde{n} = \min(n, n') = \min(n, \lfloor\tau/(2K)\rfloor)$.

Finally, the proof of the second statement is a direct adaptation of this proof by inverting $k$ and $k_\phi^*$. We don't need the event $D_k^\phi(r) = 0$ because if $k_\phi^*$ is not drawn it has necessarily lost its duel against the leader $k$. □

**Upper bounding $c_{k,1}^{\phi,\tau}$** Equation (3.7) in Lemma 3.6 is enough to upper bound $c_{k,1}^{\phi,\tau}$, by replacing $n$ by $A_k^{\phi,\tau}$. Assuming that $A_k^{\phi,\tau} \leq \tau/(2K)$ it holds that

$$\mathbb{E}[c_{k,1}^{\phi,\tau}] \leq \delta_\phi(\tau+1)\frac{e^{-A_k^{\phi,\tau}\omega_k}}{1-e^{-\omega_k}} \,. \tag{3.11}$$

**Upper bounding $c_{k,2}^{\phi,\tau}$** If the *diversity flag* is activated while $k_\phi^*$ is leader, then $k_\phi^*$ has lost at least $\lceil (K-1)(\log\tau)^2 \rceil$ successive duels while being leader. Hence, for at least one of them the sub-optimal arm has at least $(\log\tau)^2$ observations, and by definition the diversity flag was not activated during this round. Hence, we can apply again Lemma 3.6 with $n = (\log\tau)^2$, obtaining

$$\mathbb{E}[c_{k,2}^{\phi,\tau}] \leq \sum_{k'\neq k_\phi^*} \delta_\phi(\tau+1)\frac{e^{-(\log\tau)^2\omega_{k'}}}{1-e^{-\omega_{k'}}} \,. \tag{3.12}$$

**Upper bounding $c_{k,3}^{\phi,\tau}$** As in the stationary case, this term is the most difficult to handle. The main challenge is to upper bound the probability that the *optimal arm* is *not saturated* after a large number of rounds.

The remaining parts of the proof are similar to the stationary case. We first consider the case when the optimal arm has already been leader during the last $\tau$ rounds. The additional mechanisms in the non stationary case (new definition of leader, diversity flag) adds new possible scenario for a leadership takeover, and details can be found in (Baudry et al., 2021b). We only report that the contribution of this term to the upper bound of $c_{k,3}^{\phi,\tau}$ is of the form

$$D^\phi := 3(K-1)\delta_\phi(\tau+1)^3 \sum_{k\neq k_\phi^*} \frac{e^{-\left\lfloor \frac{\tau}{2K(K-1)} \right\rfloor \omega_k}}{1-e^{-\omega_k}} \,,$$

which comes from identifying the relevant events for leadership takeovers and using Lemma 3.6. For large enough $\tau$ this term is actually smaller than the previous ones we derived due as it decays exponentially with $\tau$ (up to polynomial terms).

We then consider the case when $k_\phi^*$ *has never been leader* during the last $\tau$ rounds, and use Markov inequality to relate this event to the number of duels lost by arm $k_\phi^*$ and obtain an upper bound

$$\bar{D}^\phi := \mathbb{E}\left[ \sum_{r=r_\phi+2\tau-1}^{r_{\phi+1}-2} \frac{2}{\tau} \sum_{s=r-\tau}^{r-1} \mathbb{1}\left(k_\phi^* \notin \mathcal{A}_{s+1}, \ell^\tau(s) \neq k_\phi^*\right) \right] \,.$$

We then consider whether $N_{k_\phi^*}^\tau(r) \geq A_{k_\phi^*}^{\phi,\tau}$ or not. If this is the case we can use Lemma 3.6 to derive a first upper bound

$$\bar{D}_1^\phi \leq 2\mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(k_\phi^* \notin \mathcal{A}_{r+1}, \ell^\tau(r) \neq k_\phi^*, N_{k_\phi^*}^\tau(r) \geq A_{k_\phi^*}^{\phi,\tau}\right)\right] \leq 2\delta_\phi(\tau+1)\sum_{k\neq k_\phi^*} \frac{e^{-A_{k_\phi^*}^{\phi,\tau}\omega_k}}{1-e^{-\omega_k}} .$$

The rest of the proof requires additional work compared with the stationary case. Indeed, if $k_\phi^*$ has been pulled a lot in the previous windows its index may change a lot. To avoid this we further consider whether $N_{k_\phi^*}^\tau(r-\tau) \geq A_{k_\phi^*}^{\phi,\tau}$ or not.

In the first case, $k_\phi^*$ necessarily lost a duel with exactly $A_{k_\phi^*}^{\phi,\tau}$ observations at some point. Using a union bound and Lemma 3.6 this event contributes to our upper bound up to the following factor

$$\bar{D}_2^\phi := \delta_\phi \tau(\tau+1)e^{-A_{k_\phi^*}^{\phi,\tau}\omega_k} .$$

In the last remaining case both $N_{k_\phi^*}^\tau(r-\tau)$ and $N_{k_\phi^*}^\tau(r)$ are smaller than $A_{k_\phi^*}^{\phi,\tau}$. In that case, we can finally use the arguments introduced in Chapter 2 to conclude, evaluating the *diversity of sub-samples* and using the *balance condition*.

Under these events $k_\phi^*$ competes with at most $2A_{k_\phi^*}^{\phi,\tau}$ *different sub-sample means* in the entire window $[r-\tau, r-1]$. This is due to the fact that the sub-sample changes only if $k_\phi^*$ is pulled (can happen at most $A_{k_\phi^*}^{\phi,\tau}$ times) or if $k_\phi^*$ loses one observation from the window $[r-2\tau, r-\tau-1]$ due to the sliding window (which can also happen at most $A_{k_\phi^*}^{\phi,\tau}$ times).

Thanks to these properties we know that during the interval $[r-\tau, r-1]$, $k_\phi^*$ lost at least $\tau - A_{k_\phi^*}^{\phi,\tau}$ duels and that a fraction $1/(2A_{k_\phi^*}^{\phi,\tau})$ of them occurred while the index of $k_\phi^*$ remained the same. Applying the same methodology as in the stationary case we can identify that there exists some $\beta \in (0,1)$ such that for any value of $\tau$ large enough $k_\phi^*$ lost at least a number of duels $M^\tau$ against non-overlapping blocks of some challenger $k$, with a fixed sub-sample of size larger than $f_\tau$ (the forced exploration), with

$$M^\tau = \left\lfloor \frac{\beta\tau}{2(K-1)^2(\log\tau)^2(A_{k_\phi^*}^{\phi,\tau})^2} \right\rfloor .$$

These observations allow to obtain a final contribution to our upper bound with the term

$$\bar{D}_3^\phi := 2\delta_\phi A_{k_\phi^*}^{\phi,\tau} \sum_{k\neq k_\phi^*} \sum_{j=\sqrt{\log\tau}}^{A_{k_\phi^*}^{\phi,\tau}} \alpha_k^\phi(M^\tau, j) ,$$

where we assumed for simplicity that $A_{k^*_\phi}^{\phi,\tau}$ is integer. Here $\alpha_k^\phi$ are *balance functions*, as defined in Definition 2.9, for arm $k$ and phase $\phi$.

We recall (Proposition 2.23) that for SPEF the balance function satisfies

$$\alpha_k(M^\tau, j) \leq e^{-j\omega_k^\phi} u + (1-u)^{M^\tau} \ .$$

for some constant $\omega_k^\phi$. We choose the value $u = \frac{3\log\tau}{M^\tau}$, which leads to

$$\begin{aligned}(1-u)^{M^\tau} &= \exp\left(M^\tau \log(1-u)\right) \\ &= \exp\left(M^\tau \log\left(1 - \frac{3\log\tau}{M^\tau}\right)\right) \\ &\leq \exp\left(-3\log\tau\right) \\ &\leq \frac{1}{\tau^3} \ .\end{aligned}$$

If we plug this expression to upper bound the sums we obtain

$$\bar{D}_3^\phi \leq 2\delta_\phi A_{k^*_\phi}^{\phi,\tau}(K-1)\left[\frac{e^{-\sqrt{\log\tau}\omega^\phi}}{1-e^{-\omega^\phi}}\frac{3\log\tau}{M^\tau} + \frac{A_{k^*_\phi}^{\phi,\tau}}{\tau^3}\right] \ ,$$

where $\omega^\phi = \min_{k\neq k^*_\phi} \omega_k^\phi$. Even if these terms look impressive we explain in the next section that they are not first order terms in the regret analysis.

**Summary** Due to the large number of terms introduced in the analysis we provide in this section a clarification of the final upper bound we obtained for the regret. We proved that for any given phase $\phi$,

$$\mathbb{E}[N_k^\phi] \leq 2\tau + \frac{\delta_\phi A_k^{\phi,\tau}}{\tau} + \mathbb{E}[c_{k,1}^{\phi,\tau}] + \mathbb{E}[c_{k,2}^{\phi,\tau}] + D^\phi + \bar{D}_1^\phi + \bar{D}_2^\phi + \bar{D}_3^\phi \ ,$$

where we provided explicit upper bounds for all the terms, summing as

$$\begin{aligned}\mathbb{E}[N_k^\phi] \leq 2\tau &+ \frac{\delta_\phi A_k^{\phi,\tau}}{\tau} + \delta_\phi(\tau+1)\frac{e^{-A_k^{\phi,\tau}\omega_k}}{1-e^{-\omega_k}} + \delta_\phi(\tau+1)\sum_{k'\neq k^*_\phi}\frac{e^{-(\log\tau)^2\omega_{k'}}}{1-e^{-\omega_{k'}}} \\ &+ 3(K-1)\delta_\phi(\tau+1)^3\sum_{k\neq k^*_\phi}\frac{e^{-\left\lfloor\frac{\tau}{2K(K-1)}\right\rfloor\omega_k}}{1-e^{-\omega_k}} + 2\delta_\phi(\tau+1)\sum_{k\neq k^*_\phi}\frac{e^{-A_{k^*_\phi}^{\phi,\tau}\omega_k}}{1-e^{-\omega_k}}\end{aligned}$$

$$+ \delta_\phi \tau(\tau+1) e^{-A^{\phi,\tau}_{k^*_\phi}\omega_k} + 2\delta_\phi A^{\phi,\tau}_{k^*_\phi}(K-1)\left[\frac{e^{-\sqrt{\log\tau}\omega^\phi}}{1-e^{-\omega^\phi}}\frac{3\log\tau}{M^\tau} + \frac{A^{\phi,\tau}_{k^*_\phi}}{\tau^3}\right] .$$

While this bound is relatively scary, we can now tune $A^{\phi,\tau}_k$ and $\tau$ to obtain the desired order of magnitude. First, we can tune $A^{\phi,\tau}_k = \mathcal{O}(\log(\tau))$ large enough in order to make all the terms with $A^{\phi,\tau}_k$ in their exponent as $o\left(\frac{\delta_\phi}{\tau}\right)$. We then remark that this condition is also satisfied by the terms with exponents in $\Omega((\log\tau)^2)$ and $\Omega(\tau)$. The most challenging term is the first term of the upper bound of $\bar{D}^\phi_3$. Thankfully, the exponent in $\sqrt{\log(\tau)}$ ensures an upper bound in e.g $\mathcal{O}\left(\frac{\delta_\phi A^{\phi,\tau}_{k^*_\phi}}{\tau\log(\tau)}\right)$. Putting these results together we conclude that

$$\mathbb{E}[N^\phi_k] \leq 2\tau + \mathcal{O}\left(\frac{\delta_\phi \log(\tau)}{\tau}\right) + o\left(\frac{\delta_\phi \log(\tau)}{\tau}\right) .$$

We can finally tune $\tau$ by considering the first two terms of the upper bound and summing on the phases, which provides the optimal tuning and guarantees of Theorem 3.4.

## 3.5 Experiments

In this section we test empirically the algorithms presented in this chapter. We first check that the performance of LB-SDA-LM is indeed close to the one of LB-SDA in stationary environments, and then implement some experiments with non-stationary arms.

**Limiting the storage in stationary environments.** In our first experiment[1] reported on Figure 3.2, we compare LB-SDA and LB-SDA-LM on a stationary instance with $K = 2$ arms with Bernoulli distributions for a horizon $T = 10000$. We add natural competitors (Thompson Sampling (Thompson, 1933), kl-UCB (Cappé et al., 2013)), that know ahead of the experiment that the reward distributions are Bernoulli and are tuned accordingly. The arms satisfy $(\mu_1, \mu_2) = (0.05, 0.15)$ with a gap $\Delta = 0.1$. We run LB-SDA-LM with a memory limit $m_r = \log(r)^2 + 50$, which gives a storage ranging from 50 to 150 samples for each arm (much smaller than the horizon T = 10000). The regret are averaged on 2000 independent replications and the upper and lower quartiles are reported. In this setup LB-SDA-LM performs similarly to KL-UCB, and the impact of limiting the memory is mild, when compared to LB-SDA. This illustrates that even with relatively small gaps (here 0.1), a substantial reduction of the storage can be done with only minor loss of performance with LB-SDA-LM.

[1]The code for obtaining the different figures reported in the chapter is available at https://github.com/YRussac/LB-SDA.

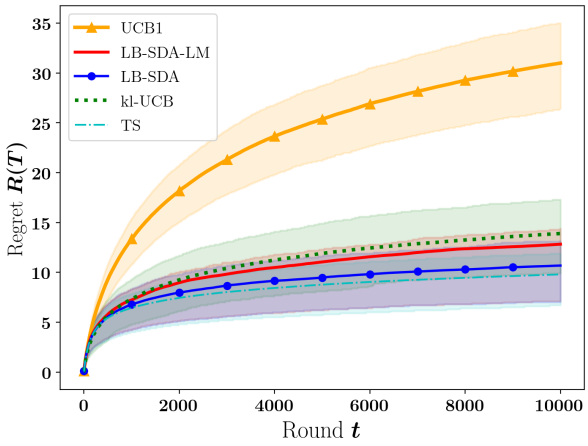


**Figure 3.2** – Cost of storage limitation on a Bernoulli instance. The reported regret are averaged over 2000 independent replications.

**Empirical performance in abruptly changing environments.** In the second experiment, we compare different state-of-the-art algorithms on a problem with $K = 3$ Bernoulli-distributed arms. The means of the distributions are represented on the left hand side of Figure 3.3 and the performance averaged on 2000 independent replications are reported on Figure 3.4. Two changepoint detection algorithms, CUSUM (Liu et al., 2017) and M-UCB (Cao et al., 2019) are compared with progressively forgetting policies based on upper confidence bound, SW-klUCB and D-klUCB adapted from (Garivier and Moulines, 2011), or Thompson sampling, DTS (Raj and Kalyani, 2017) and SW-TS (Trovo et al., 2020). We also add EXP3S (Auer et al., 2002a) designed for adversarial bandits and our SW-LB-SDA algorithm for the comparison. The different algorithms make use of the knowledge of $T$ and $\Gamma_T$.

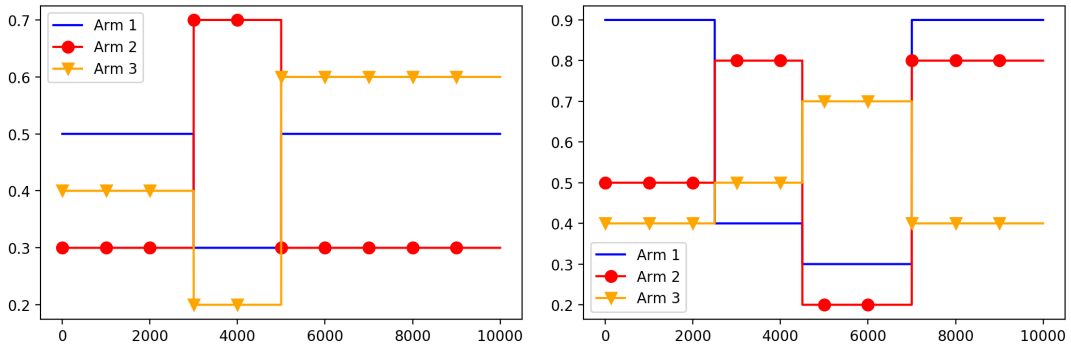


**Figure 3.3** – Evolution of the means: Left, Bernoulli arms (Fig. 3.4); Right, Gaussian arms (Figs. 3.5 and 3.6).

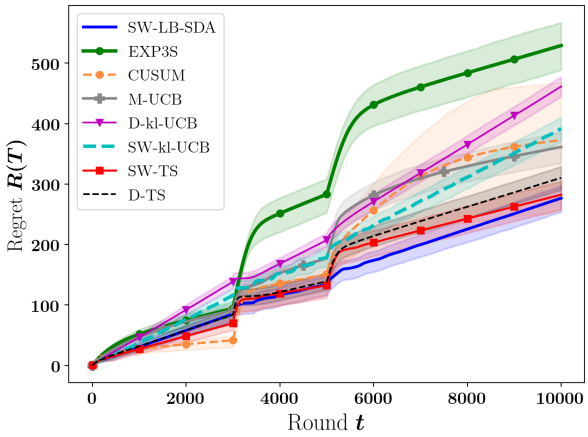


**Figure 3.4** – Performance on the Bernoulli instance of Figure 3.3, on 2000 independent replications.

To allow for fair comparison, we use for SW-LB-SDA, the same value of $\tau = 2\sqrt{T\log(T)/\Gamma_T}$ that is recommended for SW-UCB Garivier and Moulines (2011). D-UCB uses the discount factor suggested by Garivier and Moulines (2011), $1/(1-\gamma) = 4\sqrt{T/\Gamma_T}$. The changepoint detection algorithms need extra information such has the minimal gap for a breakpoint and the minimum length of a stationary phase. For M-UCB, we set $w = 800$ and $b = \sqrt{w/2\log(2KT^2)}$ as recommended by Cao et al. (2019) but set the amount of exploration to $\gamma = \sqrt{K\Gamma_T\log(T)/T}$ following (Besson et al., 2022). In practice, using this value rather than the theoretical suggestion from (Cao et al., 2019) improved significantly the empirical performance of M-UCB for the horizon considered here. For CUSUM, $\alpha$ and $h$ are tuned using suggestions from (Liu et al., 2017), namely $\alpha = \sqrt{\Gamma_T/T\log(T/\Gamma_T)}$ and $h = \log(T/\Gamma_T)$. On this specific instance, using $\varepsilon = 0.05$ (to satisfy Assumption 2 of (Liu et al., 2017)) and $M = 50$ gives good performance. For the EXP3S algorithm, following (Auer et al., 2002a) the parameters $\alpha$ and $\gamma$ are tuned as follows: $\alpha = 1/T$ and $\gamma = \min(1, \sqrt{K(e+\Gamma_T\log(KT)/((e-1)T)})$.

This problem is challenging because a policy that focuses on arm 1 to minimize the regret in the first stationary phase also has to explore sufficiently to detect that the second arm is the best in the second phase. SW-LB-SDA has performance comparable to the forgetting TS algorithms and is the best performing algorithm in this scenario. Note that both TS algorithms use the assumption that the arms are Bernoulli whereas SW-LB-SDA does not. SW-klUCB performs better than D-klUCB and its regret closely matches the one from the changepoint detection algorithms. By observing the lower and the upper quartiles, one sees that the performance of CUSUM vary much more than the other algorithms depending on its ability to detect the breakpoints. Finally, EXP3S, which can adapt to more general adversarial settings, lags behind the other algorithms in this abruptly changing stochastic environment.

In the third experiment with $\Gamma_T = 3$ breakpoints, the $K = 3$ arms comes from Gaussian distributions with a fixed standard deviation of $\sigma = 0.5$ but time dependent means. The

evolution of the arm's means is pictured on the right of Figure 3.3 and Figure 3.5 displays the performance of the algorithms. CUSUM and M-UCB can not be applied in this setting because they both consider bounded distributions. Even if no theoretical guarantees have been proved for Thompson sampling with a sliding window or discount factors when the distribution are Gaussian, we add them as competitors. The analysis of SW-UCB and D-UCB was done under the bounded reward assumption but the algorithms can be adapted to the Gaussian case. Yet, the tuning of the discount factor and the sliding window had to be adapted to obtain reasonable performance, using $\tau = 2(1+2\sigma)\sqrt{T\log(T)/\Gamma_T}$ for D-UCB and $\gamma = 1 - 1/(4(1+2\sigma))\sqrt{\Gamma_T/T}$ for SW-UCB (considering that, practically, most of the rewards lie under $1+2\sigma$). For reference, Figure 3.5 also displays the performance of the UCB1 algorithm that ignores the non-stationary structure. Clearly, SW-LB-SDA, in addition of being the only algorithm analyzed in this setting with unbounded rewards, also has the best empirical performance.

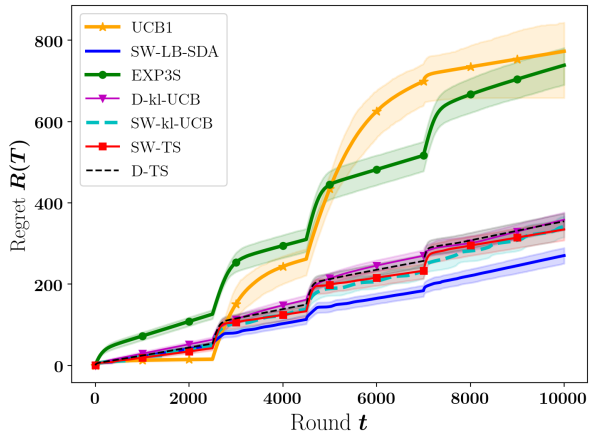


**Figure 3.5** – Performance on a Gaussian instance with a constant standard deviation of $\sigma = 0.5$ averaged on 2000 independent runs.

**Changes affecting the variance.** The last experiment features the same Gaussian means but with different standard errors. The standard error takes the values $0.5, 0.25, 1$ and $0.25$, respectively, in the four stationary phases. The algorithms based on upper confidence bound are given the maximum standard error $\sigma = 1$, whereas SW-LB-SDA is not provided with any information of this sort. Figure 3.6 shows that the non-parametric nature of SW-LB-SDA is effective, with a significant improvement over state-of-the-art methods in such settings.

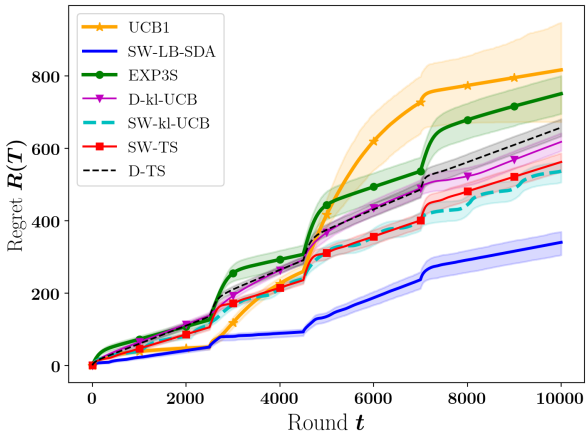


**Figure 3.6** – Performance on a Gaussian instance with time dependent standard deviations averaged on 2000 independent replications.

**Input:** $K$ arms, horizon $T$, $\tau$ length of sliding window
**Initialization:** $t = 1, r = 1, \forall k \in \{1, ..., K\} : N_k = 0, N_k^\tau = 0$
**while** $t < T$ **do**
  $\mathcal{A} = \{\}, \ell = \text{leader}(N, Y, \tau)$
  **if** $r = 1$ **then**
    $\mathcal{A} = \{1, \dots, K\}$ ; ▷ Draw each arm once
  **end**
  **else**
    **for** $k \neq \ell \in \{1, ..., K\}$ **do**
      Compute $D_k^\tau(r)$ ; ▷ Compute the diversity flag
      **if** $N_k^\tau \leq \sqrt{\log(\tau)}$ *or* $D_k^\tau(r) = 1$ **then**
        $\mathcal{A} = \mathcal{A} \cup \{k\}$ ; ▷ $k$ pulled because of diversity flag or forced exploration
      **end**
      **else**
        Run LB-SDA duel between $\ell$ and $r$ with their history collected during the $\tau$ last round. ; ▷ $k$ pulled by winning the duel
      **end**
    **end**
    **if** $|\mathcal{A}| = 0$ **then**
      $\mathcal{A} = \{\ell\}$ ; ▷ If no winning challenger $\ell$ is pulled
    **end**
  **end**
  **for** $k \in \mathcal{A}$ **do**
    Pull arm $k$, observe reward $Y_{k,t}$
    Update $N_k = N_k + 1, N_k^\tau = N_k^\tau + 1, t = t + 1$
  **end**
  **for** $k \in \{1, ..., K\}$ **do**
    **if** $k \in \mathcal{A}_{r-\tau+1}$ **then**
      $N_k^\tau = N_k^\tau - 1$
    **end**
  **end**
  $r = r + 1$
**end**

**Algorithm 3.2:** SW-LB-SDA

# Chapter 4

# Sub-sampling for Extreme Bandits

In Chapters 2 and 3 we proved that an algorithm based on a simple sub-sampling scheme, LB-SDA, performs very well for some bandit problems. In this chapter we consider a variant of bandits where the learner seeks to collect the largest possible reward, known as *Extreme Bandits*. This problem is difficult because the algorithm has to compare the heaviness of the distributions' tails, with potentially little prior information. In that case the non-parametric nature of LB-SDA is appealing, but requires to carefully choose a criterion to compare two tails. To that extent, we introduce *Quantile of Maxima* (QoMax) after studying properties of the maximum of i.i.d random variables under mild assumptions. We show that QoMax allows to build a simple Explore-Then-Commit (ETC) strategy, QoMax-ETC, achieving strong asymptotic guarantees despite its simplicity. We then propose and analyze a more adaptive algorithm, QoMax-SDA, performing pairwise comparison of QoMax estimates inside LB-SDA. Strikingly, QoMax-ETC and QoMax-SDA are more efficient than existing approaches under several aspects: (1) their non-parametric nature allows to derive strong theoretical guarantees under very mild assumptions on the tails, (2) in the experiments that we performed they lead to overall significantly better empirical performance, and (3) they enjoy a significant reduction of the memory and time complexities. This chapter if adapted from (Baudry et al., 2022).

## Contents

## 4.1 Introduction

In this thesis we consider the question of adapting Multi-Armed Bandits algorithms to work with alternative performance metric to the expected reward, that can be of practical relevance. In Section 1.2.2 of Chapter 1 we introduced *Extreme Bandit*, that is a variant of MAB where the learner seeks to collect the largest possible reward. We discussed that the algorithms that have been developed in this setting could also be used for an alternative problem where the learner would like to maximize the expected minimum, or equivalently sample most often the arm with the lightest left tail (asymptotically, if it exists). While this variant is maybe more relevant for the case-study in agriculture that we introduced by being a way to model extreme risk-aversion of the farmers, we consider in this chapter following the version of this problem with the maximum.

Letting $X_{k,t}$ be the reward that would be obtained from sampling arm $k$ at time $t$, a bandit algorithm (or policy) selects an arm $I_t$ using past observations and receives the reward $X_{I_t,t}$. The rewards stream $(X_{k,t})$ is drawn i.i.d. from $\nu_k$ and independently from other rewards streams. In this work, we assume that all arms have an unbounded support (the finite support case is studied by (Nishihara et al., 2016)), and define the *extreme regret* of a policy In this context, (Carpentier and Valko, 2014) define the extreme regret of a policy as

$$\mathcal{R}_T^{\pi} = \max_{k \leq K} \mathbb{E}[\max_{t \leq T} X_{k,t}] - \mathbb{E}_{\pi}[\max_{t \leq T} X_{I_t,t}] \, . \tag{4.1}$$

We recall from Section 1.2.2 that we expect an extreme bandit algorithm to achieve a *vanishing regret*, in the *weak sense* ($\mathcal{R}_T^{\pi} = o\left(\max_{k \leq K} \mathbb{E}[\max_{t \leq T} X_{k,t}]\right)$) or in the *strong sense* (if $\lim_{T \to \infty} \mathcal{R}_T^{\pi} = 0$). To obtain these guarantees, algorithms need to use available information on the tails of distributions. However, precise knowledge (e.g a parametric or semi-parametric model) may not be accessible to the learner in many realistic cases.

For this reason, we revisit the extreme bandit problem with the idea of designing algorithms based on *pairwise comparisons of tails* with provable guarantees under minimal assumptions on the arms. The motivation clearly stems from the study of bandit algorithms based on sub-sampling in Chapter 2 and 3, that perform "fair" pairwise comparisons of empirical means based on an equal sample size, and attain good performance for several types of distributions without changing the algorithm.

In Section 4.2, we highlight the limitation of comparing directly the maxima of $n$ i.i.d. samples and introduce the *Quantile of Maxima* (QoMax) estimator. Instead of computing the maximum of $n$ samples, the learner separates the collected data into *batches* of equal size and compute the quantile of order $q$ of the maxima over the different batches. QoMax is inspired by the Median of Means estimator (Alon et al., 1999) that was used for heavy-tail bandits (Bubeck et al., 2013). We derive upper bounds on the probability that one QoMax exceeds

another, that are instrumental to design our algorithms. In Section 4.3, we first propose an Explore-Then-Commit algorithm using QoMax, for which we establish vanishing regret in the strong sense under the mild assumption that the bandit model has a dominant arm. Albeit simple, this approach requires some tuning which depends on the horizon $T$. To overcome this limitation, we propose in Section 4.4 the QoMax-SDA algorithm which combines QoMax with the LB-SDA strategy. We prove that it achieves vanishing regret for arms with exponential or polynomial tails and also provide some elements of analysis under the weaker dominant arm assumption. In Section 6.6, we highlight the efficiency of our algorithms which allow for a significant reduction of the storage and computational cost while outperforming existing approaches empirically.

## 4.2 Comparing Tails of Distributions with Quantiles of Maxima

In this section, we motivate our new Quantile of Maxima (QoMax) estimator used for comparing the tails of two distributions based on $n$ i.i.d. samples of each. We first present the assumptions under which we are able to exhibit some properties of QoMax.

The first way to judge the heaviness of a distribution is to evaluate what probability is allocated into extreme values. To do that, we introduce the *survival function* of a distribution.

**Definition 4.1** (Survival Function). *We define the survival function $G$ of the distribution $\nu$ as*

$$G(x) : x \in \mathbb{R} \mapsto \mathbb{P}_{X\sim\nu}(X > x) \ .$$

This notion is central in the assumptions we will make on the arms' distributions. In the following we will consider two different assumptions: a first one where tails have a parametric asymptotic equivalent, and a mild non-parametric assumption.

**Definition 4.2** (Exponential or polynomial tails). *Let $\nu$ be a distribution of survival function $G$. If when $x \to +\infty$,*

1. *$G(x) \sim Cx^{-\lambda}$ for some $C > 0, \lambda > 1$ then $\nu$ has a **polynomial tail**.*
2. *$G(x) \sim C\exp(-\lambda x)$ for some $C > 0, \lambda \in \mathbb{R}^+$ then $\nu$ has an **exponential tail**.*

These *semi-parametric* assumptions (which say nothing about the lower part of the distribution) have been introduced by Bhatt et al. (2021). We remark that a polynomial tail is a weaker condition than the second-order Pareto assumption from (Carpentier and Valko, 2014) that

we introduced in Chapter 1 (Definiton 1.12). We now define a general notion that allows to compare two (arbitrary) tails.

**Definition 4.3** (Dominating tail). *Let $G_1$ and $G_2$ be the survival functions of two distributions $\nu_1$ and $\nu_2$. We say that the tail of $\nu_1$ **dominates** the tail of $\nu_2$ (we write $\nu_1 \succ \nu_2$) if the ratio of their survival functions is larger than a fixed constant for large enough values, that is*

$$\nu_1 \succ \nu_2 \iff \exists C > 1,\ \exists x \in \mathbb{R} : \forall y > x\ ,\ G_1(y) > C G_2(y)\ .$$

In the rest of the chapter we will consider a bandit model that has a dominating arm, denoted by 1 without loss of generality: we assume that $\nu_1 \succ \nu_k$ for all $k \neq 1$. Under this assumption, arm 1 is optimal in the sense that for $T$ large enough an oracle strategy would select this arm only. To the best of our knowledge, this is the weakest assumption introduced so far for extreme bandits.

### 4.2.1 Direct Comparison of Maxima

Let $\nu_X$ and $\nu_Y$ be two distributions from which we observe $n$ i.i.d. samples denoted by $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ respectively. A natural idea to compare their tails is to use the samples' maxima, that we denote by $X_n^+$ and $Y_n^+$ respectively. For these estimators to serve as a proxy for comparing the tails, we need the probability $\mathbb{P}(X_n^+ < Y_n^+)$ to decay fast enough when $\nu_X \succ \nu_Y$, ideally exponentially with the sample size. Unfortunately, the following result shows that this is not possible even under semi-parametric assumptions.

**Lemma 4.4** (Lower bound). *Assume that both $\nu_X$ and $\nu_Y$ have either polynomial or exponential tails, with respective parameters $(C_X, \lambda_X)$ and $(C_Y, \lambda_Y)$, with $\lambda_X < \lambda_Y$ (so that $\nu_X \succ \nu_Y$). Then,*

$$\mathbb{P}\left(X_n^+ \leq Y_n^+\right) = \Omega\left(n^{-\frac{\lambda_Y}{\lambda_X}}\right)\ .$$

*Proof.* Let $F_X$ and $F_Y$ be the respective cdf of the two distributions. For any sequence $(m_n)_{n\in\mathbb{N}}$ we lower bound the probability of interest as follows:

$$\begin{aligned}\mathbb{P}(X_n^+ \leq Y_n^+) &\geq \mathbb{P}(Y_1 \geq \max_{1\leq i\leq n} X_i) \\ &= \mathbb{E}_{Y_1}[F_X(Y_1)^n] = \int_{\mathbb{R}} F_X(x)^n \mathrm{d}F_Y(x)\end{aligned}$$

$$\geq \int_{m_n}^{+\infty} F_X(x)^n \mathrm{d}F_Y(x) \geq F_X(m_n)^n (1 - F_Y(m_n)) \ .$$

For exponential tails we can choose $m_n = \frac{1}{\lambda_X}\log(n)$ and obtain an asymptotic equivalent of this term of $e^{-C_X} \frac{C_Y}{n^{\frac{\lambda_Y}{\lambda_X}}}$. Furthermore, choosing $m_n = n^{1/\lambda_X}$ for polynomial tails provides the same result. This concludes the proof. □

Lemma 4.4 proves that direct comparison of maxima is not satisfying to obtain an exponential decay (in $n$) of $\mathbb{P}(X_n^+ < Y_n^+)$. However, it is also interesting to determine if there is a decay at all. We hence propose an upper bound of this probability. To obtain it, we use a trick that we already used in previous chapters: for any sequence $(x_n)$, it holds that

$$\mathbb{P}(X_n^+ < Y_n^+) \leq \mathbb{P}(X_n^+ \leq x_n) + \mathbb{P}(Y_n^+ > x_n) \ .$$

Using first that $\mathbb{P}(X_n^+ \leq x) = (1 - G_X(x))^n \leq \exp(-nG_X(x))$ and then that $\mathbb{P}(Y_n^+ > x) \leq \sum_{i=1}^n \mathbb{P}(Y_i > x) = nG_Y(x)$, and optimizing for $x_n$ yields the following result, that will be useful in the next section.

**Lemma 4.5** (Comparison of Maxima under semi-parametric assumptions)**.** *Assume that both $\nu_X$ and $\nu_Y$ have either polynomial or exponential tails, with respective second parameter $\lambda_X$ and $\lambda_Y$, with $\lambda_X < \lambda_Y$ (so that $\nu_X \succ \nu_Y$). Define $\delta = \frac{\lambda_X}{\lambda_Y} - 1 > 0$, then there exists a sequence $(x_n)$ and an integer $n_{X,Y}$ such that for all $n \geq n_{X,Y}$,*

$$\max\{\mathbb{P}(X_n^+ \leq x_n), \mathbb{P}(Y_n^+ \geq x_n)\} = \mathcal{O}\left(\frac{(\log n)^{\delta+1}}{n^\delta}\right).$$

*Proof.* The key of the proof is to consider $x_n$ "slightly" below $G_X^{-1}(1/n)$. Consider the exponential tails first, for which $G_X(x) \sim C_X \exp(-\lambda_1 x)$ and $G_Y(x) \sim C_Y \exp(-\lambda_Y x)$. We prove without loss of generality the result by continuing the proof as if the survival functions were exactly equal to their asymptotic equivalents. We choose

$$x_n = \frac{1}{\lambda_X}\left(\log n + \log(C_X) - \log(\delta \log n)\right) \ ,$$

and then compute $G_X(x_n)$ and $G_Y(x_n)$. First,

$$G_X(x_n) = C_X \exp(-(\log n + \log C_X - \log(\delta \log n)))$$

$$= \frac{\delta(\log n)}{n} \ .$$

Then,

$$\begin{aligned} G_Y(x_n) &= C_Y \exp\left(-(\delta+1)\left(\log n + \log C_X - \log(\delta \log n)\right)\right) \\ &= \frac{1}{n^{\delta+1}} \times (\delta \log n)^{\delta+1} \times \frac{C_Y}{C_X^{\delta+1}} \end{aligned}$$

So finally, we obtain

$$\mathbb{P}\left(Y_n^+ \geq x_n\right) = \mathcal{O}\left(\frac{(\log n)^{\delta+1}}{n^\delta}\right) \quad \text{and} \quad \mathbb{P}\left(X_n^+ \leq x_n\right) \leq \frac{1}{n^\delta} \ ,$$

which gives the first part of the result.

For polynomial tails we define the sequence

$$x_n = (C_X n)^{\frac{1}{\lambda_X}} \times (\delta \log n)^{-\frac{1}{\lambda_X}} \ .$$

We obtain $\exp(-nG_X(x_n)) = n^{-\delta}$, and $nG_2(x_n) = \mathcal{O}\left(\frac{(\log n)^{\delta+1}}{n^\delta}\right)$, completing the proof. □

Lemma 4.5 shows that we can upper bound the decay rate of the probability that one maximum exceeds another. However, this rate $\delta$ is problem-dependent and can be arbitrarily small. As pointed out by Carpentier and Valko (2014) it can actually be seen as the Extreme Bandits equivalent of the *gap* in bandits, we therefore call $\delta$ the ***tail gap***. With this notation the lower bound of Lemma 4.4 is of order $\Omega(n^{-(1+\delta)})$.

Interestingly, we can also obtain a result when the two distributions only satisfy $\nu_X \succ \nu_Y$ without further assumption.

**Lemma 4.6.** *Assume that $\nu_X \succ \nu_Y$. Then, for any $q \in (0,1)$ there exists $n_{\nu_X,\nu_Y,q} \in \mathbb{N}$, a sequence $(x_n)_{n\in\mathbb{N}}$ and some $\varepsilon > 0$ such that for all $n \geq n_{\nu_1,\nu_2,q}$,*

$$\mathbb{P}(X_n^+ \leq x_n) \leq q - \varepsilon \ , \ \textit{and} \quad \mathbb{P}(Y_n^+ \leq x_n) \geq q + \varepsilon \ .$$

*Proof.* Let $q \in (0,1)$ and $\varepsilon \in (0,q)$. We define the sequence $(x_n)$ by

$$G_X(x_n) = 1 - (q-\varepsilon)^{\frac{1}{n}},$$

so that $\mathbb{P}(X_n^+ \leq x_n) = q - \varepsilon$. As $\nu_X \succ \nu_Y$, there exists a constant $C > 1$ such that $G_X(x) \geq CG_Y(x)$ for $x$ large enough. Hence, as $x_n \to +\infty$ it holds that $G_1(x_n) > CG_2(x_n)$ for $n$ large

enough. For such $n$ we have

$$\begin{aligned}\mathbb{P}(Y_n^+ \leq x_n) &= (1 - G_Y(x_n))^n \\ &\geq (1 - \frac{1}{C} G_X(x_n))^n \\ &= \left(1 - \frac{1}{C}\left(1 - (q-\varepsilon)^{\frac{1}{n}}\right)\right)^n.\end{aligned}$$

We then use that for $0 \leq x \leq 1$, $\log(1-x) \geq \frac{-x}{\sqrt{1-x}}$ to get

$$\begin{aligned}\mathbb{P}(Y_n^+ \leq x_n) &\geq \exp\left(\frac{n}{C}\left((q-\varepsilon)^{\frac{1}{n}} - 1\right) \times \frac{1}{\sqrt{1 - \frac{1}{C}(1 - (q-\varepsilon)^{\frac{1}{n}})}}\right) \\ &\geq \exp\left(\frac{\log(q-\varepsilon)}{C} \frac{1}{\sqrt{1 - \frac{1}{C}(1 - (q-\varepsilon)^{\frac{1}{n}})}}\right)\end{aligned}$$

and then state that for $n$ large enough this lower bound can be arbitrarily close to $(q-\varepsilon)^{\frac{1}{C}}$, which can be made strictly larger than $q + \varepsilon$ if $\varepsilon$ is small enough, as $C$ is fixed. Hence, for an appropriate choice of $\varepsilon$ we found a sequence $x_n$ satisfying the statement of the lemma. □

### 4.2.2 Quantile of Maxima (QoMax)

Results similar to those of Section 4.2.1 have been previously encountered in the bandit literature. In (Bubeck et al., 2013), the authors study the problem of bandit with heavy tails, prove a concentration inequality in $n^{-\delta}$ for some $\delta > 0$ and use this result to build several estimators with faster convergence. Among them, they consider the Median-of-Means (MoM) introduced by Alon et al. (1999). Building on the results of previous section, we consider a natural variant of MoM, that we call Quantile of Maxima (QoMax). The principle of QoMax is simple: the learner chooses a quantile $q$, and has access to $N = b \times n$ data $\mathcal{Y} = (Y_{m,i})_{m \leq n, i \leq b}$. It then allocates the data in $b$ batches of size $n$ and: (1) finds the maximum of each batch, (2) computes the empirical quantile of order $q$ of the $b$ maxima. We summarize QoMax in Algorithm 4.1.

**Input:** quantile $q$, $b$ batches of size $n$, table of observations $(Y_{m,i})_{m \leq n, i \leq b}$
**for** $i = 1, \ldots, b$ **do**
  Compute $(Y_n^+)^{(i)} = \max\{Y_{1,i}, \ldots, Y_{n,i}\}$
**end**
**Return:** quantile of order $q$ of $\{(Y_n^+)^{(1)}, \ldots, (Y_n^+)^{(b)}\}$

**Algorithm 4.1:** Quantile of Maxima (QoMax)

For a finite set of size $b$, the quantile $q$ is simply the observation of rank $\lceil bq \rceil$ in the list of sorted data (in increasing order). In the sequel we denote by $\bar{X}^q_{n,b}$ and $\bar{Y}^q_{n,b}$ the QoMax of order $q$ computed from two datasets $(X_{m,i})_{m\leq n, i\leq b}$ and $(Y_{m,i})_{m\leq n, i\leq b}$.

We are now ready to state the crucial property of QoMax estimators that will be used in the analyses of our algorithms for Extreme Bandits.

**Theorem 4.7** (Comparison of QoMax)**.** *Let $\nu_X$ and $\nu_Y$ be two distributions satisfying $\nu_X \succ \nu_Y$ and $q \in (0,1)$. Then, if $b$ is large enough* ***there exists*** *a sequence $x_n$, a constant $c > 0$, and an integer $n_{\nu_1,\nu_2,q}$ such that for $n \geq n_{\nu_1,\nu_2,q}$,*

$$\max\left\{\mathbb{P}(\bar{X}^q_{n,b} \leq x_n), \mathbb{P}(\bar{Y}^q_{n,b} \geq x_n)\right\} \leq \exp(-cb) \ .$$

*If the tails are furthermore either polynomial or exponential with a* ***positive tail gap****, then the result holds* ***for any*** *$c > 0$ and $n$ larger than some $n_{c,\nu_1,\nu_2,q}$.*

*Proof.* We let $\mathrm{kl}(x,y) = x\log(x/y) + (1-x)\log((1-x)/(1-y))$ denote the binary relative entropy. Just like for the analysis of Median-of-Means, the starting point is to relate deviations inequalities for a QoMax to deviation inequalities for binomial distributions. Letting $(X^+_n)^{(i)}$ (resp. $(Y^+_n)^{(i)}$) denote the maximum over the $i$-th batch of observations from $\nu_X$ (resp. $\nu_Y$),

$$\begin{aligned}\mathbb{P}(\bar{X}^q_{n,b} \leq x) &\leq \mathbb{P}\left(\sum_{i=1}^{b} \mathbb{1}\Big((X^+_n)^{(i)} \leq x\Big) \geq bq\right) \\ &\leq \exp(-b \times \mathrm{kl}(q, \mathbb{P}(X^+_n \leq x))) \ .\end{aligned}$$

The last step applies the Chernoff inequality to a binomial distribution with parameters $b$ and $p = \mathbb{P}(X^+_n \leq x)$, and holds whenever $\mathbb{P}(X^+_n \leq x) \leq q$. Similarly, if $\mathbb{P}(Y^+_n \geq x) \leq 1-q-1/b$, we have

$$\begin{aligned}\mathbb{P}(\bar{Y}^q_{n,b} \geq x) &\leq \mathbb{P}\left(\sum_{i=1}^{b} \mathbb{1}\Big((Y^+_n)^{(i)} \geq x\Big) \geq b - bq - 1\right) \\ &\leq \exp(-b\mathrm{kl}(1-q-1/b, \mathbb{P}(Y^+_n \geq x)))\end{aligned}$$

For exponential and polynomial tails, thanks to Lemma 4.5 there exists a sequence $(x_n)$ such that both $\mathbb{P}(X^+_n \leq x_n)$ and $\mathbb{P}(Y^+_n \geq x_n)$ satisfy the desired property, and the result follows easily. With the notation of Lemma 4.6, Theorem 4.7 then holds for

$$c = \min\left(\mathrm{kl}(q, q-\varepsilon), \mathrm{kl}(1-q-\varepsilon/2, 1-q-\varepsilon)\right) \ ,$$

provided that the batch size is larger than $2/\varepsilon$. □

It follows from Theorem 4.7 that $\mathbb{P}(\bar{X}^q_{n,b} \leq \bar{Y}^q_{n,b}) \leq 2\exp(-cb)$ for $n$ large enough. Strikingly, this result tells us that, under the simple assumption that one tail dominates, the comparison of QoMax computed with the same parameters will not be in favor of the dominating arm with a probability that **decreases exponentially with the batch size**.

**Remark 4.8.** *In general QoMax is **not** an estimate of the expectation of the maximum. We will use it only for the purpose of **comparing two tails**, in order to find the heaviest.*

**Remark 4.9** (Choice of quantile level $q$)**.** *Note that Theorem 4.7 holds for any value of $q \in (0, 1)$, but the impact of $q$ is materialized in the (problem-dependent) sample size $n_{\nu_1,\nu_2,q}$ needed for the inequality to hold. For the practitioner, we think that in most cases choosing $q = 1/2$ is appropriate. Still, in Section 6.6 we exhibit a difficult setting where a choice of $q$ close to 1 is helpful.*

## 4.3 QoMax-ETC

In this section, we propose QoMax-ETC, a simple Explore-Then-Commit algorithm using QoMax estimators. The algorithm is reported in Algorithm 4.2 and works as follows. First, the learner selects a quantile $q$, and given the time horizon $T$ picks a batch size $b_T$ and a sample size $n_T$. Then, the exploration phase starts where every arm is pulled $N_T = b_T \times n_T$ times allocated in $b_T$ batches of size $n_T$. At the end of this step, the learner computes a $q$-QoMax estimator from the history of each arm using the different batches. Next comes the exploitation phase where the algorithm pulls the arm $I_T$ with the largest QoMax until time $T$.

**Input:** $K$ arms, horizon $T$, quantile $q$, number of batches $b_T$, number of samples per batch $n_T$
**for** $k = 1, \dots, K$ **do**
  Pull arm $k$, $b_T \times n_T$ times
  Allocate the data in $b_T$ batches of size $n_T$
  Compute their QoMax, $\bar{Y}^q_{k,n_T,b_T}$ (Algorithm 4.1)
**end**
**for** $t = K \times n_T \times b_T + 1, \dots, T$ **do**
  Pull arm $I_T = \text{argmax}_k \bar{Y}^q_{k,n_T,b_T}$
**end**

**Algorithm 4.2:** QoMax-ETC

We remark that an ETC algorithm has already been proposed by (Achab et al., 2017) for extreme bandits. Their algorithm differs from ours by the choice of the arm $I_T$ drawn in

the exploitation phase: they build an upper confidence bound on the maximum under the assumption that the distributions are second-order Pareto (Definition 1.12) and select $I_T$ as the arm with largest upper confidence bound. In contrast, QoMax-ETC does not assume anything about the arms distributions. In our case, Theorem 4.7 is the main motivation for building an ETC strategy: with a large enough batch size $b_T$ and sample size $n_T$ strong concentration of QoMax estimates can be obtained. We now analyze QoMax-ETC under a bandit model $\nu = (\nu_1, \dots, \nu_K)$ such that $\nu_1 \succ \nu_k$ for all $k \neq 1$.

**Proposition 4.10** (Regret of QoMax-ETC). *Let $\pi$ be an ETC policy sampling $N_T = n_T \times b_T$ times each arm during the exploration phase. If $T \geq KN_T$,*

$$\mathcal{R}_T^\pi \leq \underbrace{\mathbb{E}\left[\max_{t\leq T} Y_{1,t}\right] - \mathbb{E}\left[\max_{t\leq T-KN_T} Y_{1,t}\right]}_{\textit{Exploration cost}} + \underbrace{\mathbb{P}(I_T \neq 1)\mathbb{E}\left[\max_{t\leq T} Y_{1,t}\right]}_{\textit{Cost of picking a wrong arm}} .$$

*Proof.* We recall that $N_T = b_T \times n_T$ is the number of pulls of each arm during the exploration phase of the ETC algorithm (see Algorithm 4.2) and that $X_{k,t}$ corresponds to the observation of arm $k$ at time $t$ (if any). We also denote by $Y_{k,n}$ the $n$-th observation collected from arm $k$. The ETC simplifies a lot the study of the extremal regret, as we can separate the explore and commit phase in the analysis. First, an exact decomposition of the expected value of the policy is

$$\mathbb{E}\left[\max_{t\leq T} X_{I_t,t}\right] = \mathbb{E}\left[\max\left\{\max_k \max_{t\leq KN_T} X_{k,t}, \max_{t=[KN_T+1,T]} X_{I_T,t}\right\}\right] .$$

We obtain the lower bound by ignoring the exploration phase and using that the rewards collected during the exploitation phase are conditionally independent of the outcome of the exploration phase

$$\begin{aligned}
\mathbb{E}\left[\max_{t\leq T} X_{I_t,t}\right] &\geq \mathbb{E}\left[\max_{t=[KN_T+1,T]} X_{I_t,t}\right] = \mathbb{E}\left[\max_{t=[KN_T+1,T]} X_{I_T,t}\right] \\
&= \mathbb{E}\left[\max_{t=[KN_T+1,T]} X_{I_T,t} \sum_{k=1}^K \mathbb{1}(I_T = k)\right] = \sum_{k=1}^K \mathbb{E}\left[\max_{t=[KN_T+1,T]} X_{I_T,t}\mathbb{1}(I_T = k)\right] \\
&= \sum_{k=1}^K \mathbb{P}(I_T = k)\mathbb{E}\left[\max_{t=[KN_T+1,T]} Y_{k,t}\right] \geq \mathbb{P}(I_T = 1)\mathbb{E}\left[\max_{t=[1,T-KN_T]} Y_{1,t}\right] \\
&= (1 - \mathbb{P}(I_T \neq 1))\mathbb{E}\left[\max_{t=[1,T-KN_T]} Y_{1,t}\right] \\
&\geq \mathbb{E}\left[\max_{t\leq T-KN_T} Y_{1,t}\right] - \mathbb{P}(I_T \neq 1)\mathbb{E}\left[\max_{t\leq T} Y_{1,t}\right] .
\end{aligned}$$

We also used that if the distributions are supported on $\mathbb{R}$ the expectation of their maximum is positive for $T$ large enough. This concludes the proof.

□

This proposition shows that the regret of the ETC algorithm can be properly controlled by two factors:

1. the probability of picking a wrong arm for the exploitation phase.
2. the gap between the growth rate of the maximum over $T$ or $T - KN_T$ observations of the dominant arm, that we call "exploration cost" as it is fully determined by the length of the exploration phase and the arms' distributions.

In the rest of the chapter we will assume that the distribution of the dominant arm satisfies the following assumption.

**Assumption 4.11.** $\mathbb{E}\left[Y_T^+\right] = o(T)$, *and for any* $\gamma < 1$ *if* $N_T = o(T^\gamma)$ *then*

$$\mathbb{E}\left[Y_T^+\right] - \mathbb{E}\left[Y_{T-N_T}^+\right] \xrightarrow[T\to+\infty]{} 0 \,.$$

This condition is satisfied for nearly all distributions encountered in practice (e.g polynomial, exponential or gaussian tails). The following results support this claim by providing a generic way to upper bound the exploration cost and an application to semi-parametric tails.

**Proposition 4.12** (Universal upper bound on the exploration cost)**.** *For any distribution of survival function* $G$*, for any constant* $B > 0$ *it holds that*

$$\mathbb{E}[Y_T^+] - \mathbb{E}[Y_{T-N_T}^+] \leq N_T \left(\frac{B}{T} + \int_B^\infty G(x)dx\right)$$

*Proof.* We write

$$\begin{aligned}
\mathbb{E}[Y_T^+] - \mathbb{E}[Y_{T-N_T}^+] &= \mathbb{E}\left[Y_{T-N_T+1:T}^+ \mathbb{1}\left(Y_T^+ \neq Y_{T-N_T+1:T}^+\right)\right] \\
&\leq \mathbb{E}\left[Y_{T-N_T+1:T}^+ \mathbb{1}\left(Y_T^+ \neq Y_{T-N_T+1:T}^+\right) \mathbb{1}\left(Y_{T-N_T+1:T}^+ \leq B\right)\right] \\
&\quad + \mathbb{E}\left[Y_{T-N_T+1:T}^+ \mathbb{1}\left(Y_{T-N_T+1:T}^+ > B\right)\right] \\
&\leq B\mathbb{P}\left(Y_T^+ \neq Y_{T-N_T+1:T}^+\right) + \int_B^\infty \mathbb{P}\left(Y_{T-N_T+1:T}^+ > x\right) dx
\end{aligned}$$

$$
\begin{aligned}
&= B\frac{N_T}{T} + \int_B^\infty \mathbb{P}\left(Y^+_{N_T} > x\right) dx \\
&\leq B\frac{N_T}{T} + N_T \int_B^\infty \mathbb{P}(Y_1 > x) dx \\
&\leq N_T \left(\frac{B}{T} + \int_B^\infty G(x) dx\right).
\end{aligned}
$$

where we have used the fact that that maximum has the same probability to be attained in each element and the union bound $\mathbb{P}\left(Y^+_{N_T} > x\right) \leq \sum_{i=1}^{N_T} \mathbb{P}(Y_i > x)$. □

To prove that Assumption 4.11 is satisfied for exponential and polynomial tails, we then exhibit a value of $B$ such that the resulting upper bound in Proposition 4.12 tends to 0.

**Lemma 4.13.** *Assumption 4.11 is satisfied for exponential and polynomial tails.*

*Proof.* If the tail is exponential there exists $\lambda > 0$ and for any choice of $B$ there exists a constant $C_B$ such that $G(x) \leq C_B e^{-\lambda x}$ if $x \geq B$, so

$$
\begin{aligned}
\mathbb{E}[Y^+_T] - \mathbb{E}[Y^+_{T-N_T}] &\leq N_T \left(\frac{B}{T} + C_B \int_B^\infty e^{-\lambda x} \mathrm{d}x\right) \\
&= N_T \left(\frac{B}{T} + \frac{C_B}{\lambda} e^{-\lambda B}\right).
\end{aligned}
$$

If there exists $\gamma \in (0,1)$ such that $N_T = o(T^\gamma)$, choosing $B = \frac{\log(T)}{\lambda}$ yields the result.

If the tail is polynomial, there exists similarly a constant $\lambda$ and for any $B$ a constant $C_B$ such that $G(x) \leq C_B x^{-\lambda}$ for $x \geq B$, so

$$
\begin{aligned}
\mathbb{E}[Y^+_T] - \mathbb{E}[Y^+_{T-N_T}] &\leq N_T \left(\frac{B}{T} + C_B \int_B^\infty \frac{1}{x^\lambda} \mathrm{d}x\right) \\
&= N_T \left(\frac{B}{T} + \frac{C_B}{\lambda} B^{1-\lambda}\right)
\end{aligned}
$$

Choosing $B = T^{1/\lambda}$ yields

$$
\mathbb{E}[Y^+_T] - \mathbb{E}[Y^+_{T-N_T}] \leq \left(1 + \frac{C_B}{\lambda}\right) \frac{N_T}{T^{1-\frac{1}{\lambda}}}
$$

If *for all* $\gamma \in (0,1)$, $N_T = o(T^\gamma)$ then in particular $N_T = o(T^{1-\frac{1}{\lambda}})$ and $\lim_{T\to\infty} \mathbb{E}[Y^+_T] - \mathbb{E}[Y^+_{T-N_T}] = 0$. □

**Remark 4.14.** *In both cases we chose $B = G^{-1}(1/T)$, which should work for other cases as well.*

We now state our main theoretical claim for QoMax-ETC.

**Theorem 4.15** (Vanishing regret of QoMax-ETC)**.** *Consider a bandit $\nu = (\nu_1, \dots, \nu_K)$ with $\nu_1 \succ \nu_k$ for $k \neq 1$. Under Assumption 4.11, for any quantile $q \in (0,1)$ and any sequence $(b_T, n_T)$ satisfying*

$$\frac{b_T}{\log(T)} \to +\infty \text{ and } n_T \to +\infty \,,$$

*the regret of QoMax-ETC with parameters $(q, b_T, n_T)$ is **vanishing in the strong sense**. Furthermore, for polynomial/exponential tails with positive tail gaps this result also holds for $b_T = \Omega(\log T)$.*

*Proof.* From Theorem 4.7, there exists constants $c_k$ for $k \geq 2$ such that for $T$ large enough (such that $n_T$ becomes larger than $n_{\nu_1,\nu_k,q}$), it holds that

$$\mathbb{P}(I_T \neq 1) \leq \sum_{k=2}^{K} \mathbb{P}(\bar{X}^q_{k,n_T,b_T} > \bar{X}^q_{1,n_T,b_T}) \leq \sum_{k=2}^{K} e^{-c_k b_T}$$

It follows that $\mathbb{P}(I_T \neq 1) = o(T^{-1})$ if $b_T/\log(T) \to \infty$ and we conclude with Proposition 4.10 and Assumption 4.11. For polynomial or exponential tails, as the above inequality holds for any value of $c_k$, $b_T = \Omega(\log T)$ is sufficient to obtain $\mathbb{P}(I_T \neq 1) = o(T^{-1})$. □

Even if Theorem 4.15 is stated in an asymptotic way, we emphasize that its proof provides a finite-time upper bound on the probability of picking a wrong arm, $\mathbb{P}(I_T \neq 1)$, that is valid provided that $T$ is larger than some (problem-dependent) constant. In particular, $T$ needs to be large enough so that $n_T \geq \max_{k\neq 1} n_{\nu_1,\nu_k,q}$ where $n_{\nu_1,\nu_k,q}$ is the number of samples needed in Theorem 4.7 for the concentration of QoMax. This number is not always large. For example if we have two Pareto distributions with parameters $\lambda_1 = 1.5$ and $\lambda_2 = 2$, $n_T = 3$ is enough. Using our regret decomposition, this result would lead to a finite-time upper bound on the extremal regret for distributions for which a finite-time bound on the exploration cost is available.

To satisfy the theoretical requirements while obtaining good empirical performance, we recommend using $b_T = (\log(T))^2$ and $n_T = \log(T)$ when running the algorithm. All the experiments reported in Section 4.5 use these values. QoMax-ETC is computationally appealing and has strong asymptotic guarantees. However in practice we found that its performance can vary significantly depending on the choices of $b_T$ and $n_T$, which should in particular use a reasonable guess for the horizon $T$. For this reason, in the next section we propose QoMax-SDA, which is still based on QoMax comparisons but is anytime (i.e. independent on $T$) and requires less parameter tuning.

## 4.4 QoMax-SDA

In this section we present QoMax-SDA, an algorithm using a sub-sampling mechanism based on LB-SDA, that we studied extensively in Chapters 2 and 3. We detail the key principles of the algorithm and propose a theoretical analysis.

### 4.4.1 Algorithm and Implementation

From a high level QoMax-SDA follows the structure of the Sub-sampling Dueling Algorithms (SDA) introduced in Chapter 2. The algorithm operates in successive rounds composed of (1) the selection of a leader, (2) the different duels between the leader and the challengers and (3) a data collection phase. We develop each of those steps in the sequel.

**Data and leader selection** At the beginning of a round $r$, the learner has access to the history of the different arms denoted by $\mathcal{Y}_k^r$. For the needs of the QoMax, the collected rewards for arm $k$ are gathered within $b_k(r)$ batches of equal size $n_k(r)$ such that $|\mathcal{Y}_k^r| = b_k(r)n_k(r)$. $n_k(r)$ is called the *number of queries* and corresponds to the number of times the arm $k$ has been selected by the learner at the end of round $r$. The leader at round $r$, denoted by $\ell(r)$, is the arm that has been queried the most up to round $r$, that is $\ell(r) \in \text{argmax}_{k \leq K}\ n_k(r)$. The $K-1$ remaining arms are called *challengers*. In case of equality, ties are broken first in favor of the arm with the largest QoMax, then at random.

**Duels** Once the leader is selected, it plays a duel against each challenger. As in previous chapters we denote by $\mathcal{A}_{r+1}$ the set of arms that will be pulled at the end of round $r$. An arm $k$ is added to $\mathcal{A}_{r+1}$ in two cases: (1) if it wins its duel or (2) if its number of queries is too small: $n_k(r) \leq f(r)$ for a fixed function $f(r)$ representing the *forced exploration*. As for standard SDA the leader is pulled only if no challenger is added to $\mathcal{A}_{r+1}$. We now detail the duel procedure that is reported in Algorithm 4.3. We assume that an infinite stream of rewards is available for each arm, in the form of an array with an infinite number of rows and columns, so that we denote the rewards of arm $k$ by $(Y_{k,n,b})_{n\in\mathbb{N},b\in\mathbb{N}}$, where $Y_{k,n,b}$ corresponds to the $n$-th sample of $b$-th batch from arm $k$. We further assume that the number of batches available for an arm $k$ depends only on its number of queries $n_k(r)$ so that $b_k(r) = \lceil B(n_k(r))\rceil$ for some function $B$. Following the principle of SDA, the duel is a comparison of the QoMax of the challenger using its entire history and the QoMax of the leader on a sub-sample of its history.

Our sub-sampling mechanism is inspired by LB-SDA, but has to consider two dimensions: we keep the rewards collected from $n_k(r)$ the **last queries** of the $b_k(r)$ **first batches** from $\ell(r)$. This way the QoMax from the leader and the challenger are computed using the same amount

**Input:** $q$, arm $k$, leader $\ell$, current history, batch count and batch size: $(\mathcal{Y}_m, b_m, n_m)$ for $m \in \{k, \ell\}$
Compute $I_k = \text{QoMax}(q, b_k, n_k, \mathcal{Y}_k)$ (Alg. 4.1) ; ▷ QoMax of the challenger
Collect $\mathcal{Y}_\ell = (Y_{\ell,i,j})_{i\in[\mathbf{n_\ell - n_k + 1}, \mathbf{n_\ell}], j\in[\mathbf{1}, \mathbf{b_k}]}$ ; ▷ Leader's sub-sample
Compute $I_\ell = \text{QoMax}(q, b_k, n_k, \mathcal{Y}_\ell)$ (Alg. 4.1) ; ▷ QoMax used for the leader
**Return:** $\text{argmax}_{m\in\{k,\ell\}} I_m$

**Algorithm 4.3:** Duel ($q$-QoMax comparison)

of data, but the diversity in the sub-samples (see Chapter 2 for intuitions on this topic) comes from the "query" dimension. We explain why in next paragraph.

**Data Collection** We now detail the data collection procedure that is used by QoMax-SDA and illustrated on Figure 4.1. If we query an arm $k$ at round $r$, we will (1) add 1 observation to each existing batch, and (2) collect enough data to build a new batch. We formalize this for one arm in Algorithm 6.

**Input:** queried arm $k$, history $\mathcal{Y}_k$ of size $n_k \times b_k$, target batch size $S$
Add $b_k$ new data drawn from $\nu_k$ in a new row of $\mathcal{Y}_k$ ; ▷ Update existing batches
**while** $b_k < S$ **do**
    Add $n_k + 1$ data drawn from $\nu_k$ in a new column of $\mathcal{Y}_k$ ; ▷ Create new batch
    $b_k = b_k + 1$ ; ▷ Update batch count
**end**

**Algorithm 4.4:** CollectData procedure (without storage reduction trick)

**Implementation of QoMax-SDA** The combination of the leader selection, the duel step, and the data collection gives QoMax-SDA, reported in Algorithm 4.5. While we wrote the algorithm with this function for simplicity, we actually do not recommend to use Algorithm 6 for the data collection step. Indeed, our algorithm can enjoy a significant reduction of storage with two different tricks:

1. The maxima can be stored efficiently : when a new value $x$ is added inside a given batch, all stored values smaller than $x$ are deleted as the algorithm will never need them again. However we need to store the round at which the observation was collected in case we have to sub-sample this arm.

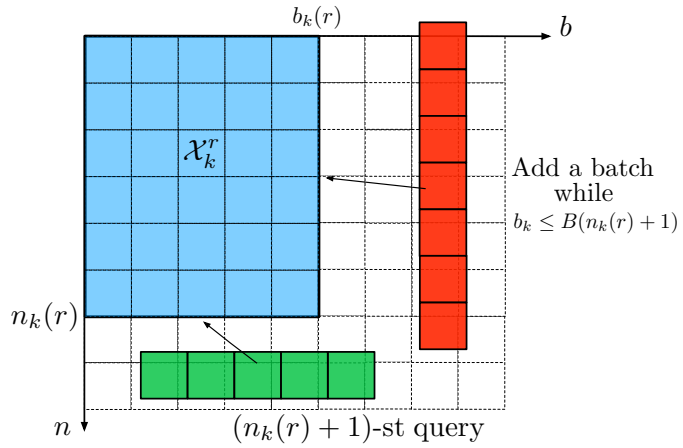


**Figure 4.1** – Illustration of the CollectData procedure at round $r$ for a challenger $k \in \mathcal{A}_{r+1}$ with data $\mathcal{X}_k^r$.

2. Create new batches for the leader only when it has to match the number of batches of the second most pulled arm, to avoid creating (and storing) unused batches. Indeed, if the algorithm ends up pulling an arm most of the time (which is expected), this will create new batches for the leader that are never used in the duels because only the first $b_k(r)$ batches are used when the leader competes with arm $k$. In the implementation this requires to (1) never add new batches when the leader is pulled, and (2) create a new batch for the leader (i.e pull it $n_\ell(r)$ times) when a challenger $k$ is pulled such that $b_k(r) > b_\ell(r)$

The second point is furthermore interesting for exploration as it allows to play more duels for a fixed number of draws. We provide some results on the memory saved thanks to those tricks in Appendix 4.7.

Finally, we can note that a forced exploration, through the function $f$ (independent on $T$), is necessary under general assumptions as in all existing algorithms.

### 4.4.2 Extreme Regret Analysis

We now provide an analysis of QoMax-SDA under the same assumption as before: $\nu_1 \succ \nu_k$ for all $k \neq 1$. Let $N_k(t)$ denote the number of pulls of arm $k$ at time $t$. We start with a generic regret decomposition.

**Proposition 4.16** (Regret decomposition with a low probability event)**.** *Define the event*

$$\xi_T := \{N_1(T) \leq T - KM_T\} ,$$

**Input:** $K$ arms, quantile level $q$
exploration function $f$, batch function $B$
**Initialization:** $r = 1$
$\forall k \in \{1, ..., K\}$: $n_k = 1$, $\mathcal{Y}_k = \{X_1^k\}$ ; ▷ Draw each arm once
**for** $r \geq 2$ **do**
  $\mathcal{A} = \{\}$, $\ell = \text{leader}((n_k)_{k\in\{1,...,K\}}, (\mathcal{Y}_k)_{k\in\{1,...,K\}})$ ; ▷ Define the leader
**end**
**for** $k \neq \ell \in \{1, ..., K\}$ **do**
  **if** $n_k < f(r)$ *or* Duel$(k, \ell) = k$ (Alg. 4.3) **then**
    $\mathcal{A} = \mathcal{A} \cup \{k\}$ ; ▷ $k$ pulled if it is not enough explored or wins the duel
  **end**
**end**
**if** $|\mathcal{A}| = 0$ **then**
  $\mathcal{A} = \{l\}$ ; ▷ Draw the leader if no winning challenger
**end**
**for** $k \in \mathcal{A}$ **do**
  CollectData$(\mathcal{Y}_k, B(r))$ (Alg. 6) ; ▷ Data collection
  $n_k = n_k + 1$
**end**

**Algorithm 4.5:** QoMax-SDA (simplified data collection procedure)

*where $(M_T)_{T\in\mathbb{N}}$ is a fixed sequence. Then, for $T \geq KM_T$, for any constant $x_T \in \mathbb{R}$, it holds that,*

$$\mathcal{R}_T^\pi \leq \underbrace{\mathbb{E}\left[\max_{t\leq T} Y_{1,t}\right] - \mathbb{E}\left[\max_{t\leq T-KM_T} Y_{1,t}\right]}_{\textit{Exploration cost}} + \underbrace{x_T\mathbb{P}(\xi_T) + \mathbb{E}\left[\max_{t\leq T} Y_{1,t}\mathbb{1}\left(\max_{t\leq T} Y_{1,t} \geq x_T\right)\right]}_{\textit{Cost incurred by } \xi_T} .$$

*Proof.* Performing the exact same steps as for the proof of Proposition 4.10 one can first obtain

$$\mathbb{E}\left[\max_{t\leq T} X_{I_t,t}\right] \geq \mathbb{E}\left[\max_{t\leq T-KM_T} Y_{1,t}\right] - \mathbb{E}\left[\max_{t\leq T} Y_{1,t}\mathbb{1}(\xi_T)\right] .$$

However, contrarily to the ETC strategy for SDA the variable $\mathbb{1}(\xi_T)$ and the rewards are not conditionally independent. Additional steps are required to analyze the two quantities separately. Using the notation $Y_T^+ = \max_{t\leq T} Y_{1,t}$ for simplicity, and then considering a constant $x_T \in \mathbb{R}$ we can write

$$\begin{aligned}\mathbb{E}\left[\max_{t\leq T} Y_{1,t}\mathbb{1}(\xi_T)\right] &= \mathbb{E}[Y_T^+\mathbb{1}(\xi_T)] \leq \mathbb{E}[Y_T^+\mathbb{1}(\xi_T)\mathbb{1}(Y_T^+ \leq x_T)] + \mathbb{E}[Y_T^+\mathbb{1}(\xi_T)\mathbb{1}(Y_T^+ \geq x_T)] \\ &\leq x_T\mathbb{P}(\xi_T) + \mathbb{E}[Y_T^+\mathbb{1}(Y_T^+ \geq x_T)] .\end{aligned}$$

This concludes the proof. □

**Remark 4.17** (Comparison with Proposition 4.10)**.** *The expression we obtained can be compared with the result for ETC strategies. The first part (exploration cost) is the same. However, the second term is more complicated as we could simply write* $\mathbb{P}(\xi_T)\mathbb{E}[\max_{t\leq T} X_{1,t}]$ *for the ETC strategy. If we did this we would now obtain* $\mathbb{P}(\xi_T) \times \mathbb{E}\left[Y_T^+ | \xi_T^c\right]$*, where the second term is unfortunately not equal to* $\mathbb{E}\left[X_T^+\right]$*. We cannot proceed further with this upper bound as the conditional expectation may be very intricate to compute. However, we conjecture that the upper bound* $\mathbb{E}\left[Y_T^+ | \xi_T^c\right] \leq \mathbb{E}\left[Y_T^+\right]$ *should hold. Indeed,* $\xi_T^c$ *and the maximum should be positively correlated, as this event corresponds to arm* 1 *being pulled a lot, and hence performing quite well. This seems however difficult to prove.*

Another interesting observation is that, even if the "cost incurred by $\xi_T$" features two terms, interestingly only the first term depends on the algorithm. In fact, just as with Proposition 4.10 our analysis again depends mostly on upper bounding the probability of a "bad" event. We propose the following result for QoMax-SDA.

**Lemma 4.18** (Upper bound on $\mathbb{P}(\xi_T)$ for QoMax-SDA)**.** *Consider a bandit* $\nu = (\nu_1, \dots, \nu_K)$ *satisfying* $\nu_1 \succ \nu_k, \forall k \neq 1$*. For any* $q \in (0,1)$*, any* $M_T$ *and any* $\gamma > 0$*, under QoMax-SDA with parameters* $B(n) = n^\gamma$ *and* $f(r) = (\log r)^{\frac{1}{\gamma}}$*,*

$$\mathbb{P}(\xi_T) = \mathcal{O}\left((\log T)^{\frac{1}{\gamma}} M_T^{-\frac{1}{1+\gamma}}\right) .$$

*Moreover, for all* $k \neq 1$*,* $\mathbb{E}[n_k(T)] = \mathcal{O}((\log T)^{1/\gamma})$*.*

*Sketch of proof.* We first use that $\mathbb{P}(\xi_T) \leq \sum_{k=2}^K \mathbb{P}(N_k(T) \geq M_T)$. Using that $N_k(T) = b_k(T) \times n_k(T) = n_k(T)^{1+\gamma}$ and Markov inequality we obtain

$$\mathbb{P}(\xi_T) \leq M_T^{-\frac{1}{1+\gamma}} \sum_{k=2}^K \mathbb{E}[n_k(T)] .$$

It remains to study the expected number of queries of sub-optimal arms $k \geq 2$. This can be done following the outline of the proof of Theorem 2.12 in Chapter 2 for SDA in the standard setting while using the deviation inequalities from Theorem 4.7 instead of the ones from Assumption 2.4. □

The second term in the "cost incurred by $\xi_T$" only depends on the distribution of the optimal arm and can be further upper bounded assuming exponential and polynomial tails, leading to the following result.

**Theorem 4.19** (Upper bound on the regret of QoMax-SDA). *Under the assumptions of Lemma 4.18 it further holds that*

1. *the regret of QoMax-SDA is vanishing in the strong sense for exponential tails*
2. *the regret of QoMax-SDA is vanishing in the weak sense for polynomial tails.*

*Sketch of proof.* For parametric tails, we can compute the growth rate of $\mathbb{E}[\max_{t\leq T} X_{1,t}]$ with respect to $T$. This permits to tune the values of $M_T$ and $x_T$ to properly balance the terms in the regret decomposition. The difference in the convergence for **(1)** and **(2)** comes from the fact that the exploration cost scales logarithmically with the time horizon when using exponential tails, whereas the dependency is polynomial with polynomial tails. □

We note that there is no hope to upper bound the last term in the regret decomposition of Proposition 4.10 assuming only that arm 1 dominates the others, so we could not establish vanishing regret for QoMax-SDA under this assumption. However, at least weakly vanishing regret could be established using the conjecture we make in Remark 4.17. Even if we were not able to prove this, we note that QoMax-SDA achieves state-of-the-art performance for exponential and polynomial tails, and that Lemma 4.18 provides a strong indicator of the good performance of QoMax-SDA under more general assumptions, as it shows that the algorithm queries each sub-optimal arm $\mathcal{O}((\log T)^{\frac{1}{\gamma}})$ times. With this number of queries the total number of data collected from sub-optimal arms would be $\mathcal{O}((\log T)^{1+\frac{1}{\gamma}})$. Knowing if it is possible to design an extreme bandit algorithm that would use only $\mathcal{O}(\log(T))$ data for sub-optimal arms as in the standard setting under mild assumptions on the tails is an interesting open question. We now turn our attention to the practical benefits of using QoMax-based algorithms.

## 4.5 Practical performance

In all of our experiments, we compare QoMax-SDA and QoMax-ETC with ThresholdAscent (Streeter and Smith, 2006b), ExtremeHunter (Carpentier and Valko, 2014), ExtremeETC (Achab et al., 2017) and MaxMedian (Bhatt et al., 2021). We use the parameters suggested in the original papers (see details in Baudry et al. (2022)). Namely, $b = 1$ for ExtremeHunter/ETC, $s = 100, \delta = 0.1$ for ThresholdAscent, $\varepsilon_t = (t+1)^{-1}$ for MaxMedian. For QoMax-ETC, we use $b_T = (\log T)^2$ batches of $n_T = \log T$ samples. This matches the size of the exploration phase of ExtremeETC and allows for a fair comparison. For QoMax-SDA, we choose $\gamma = 2/3$, which seems to work well across all examples. All the results presented in this section are obtained with these values.

### 4.5.1 Time and Memory Complexity

We summarize in Table 4.1 the storage and computational time required by the different adaptive and ETC algorithms that we consider, with the aforementioned parameters. The smallest values in each category are colored in blue. We do not include ThresholdAscent in the table because the comparison is unfair, as it uses a fixed number of data but is not theoretically grounded. We refer the reader to (Bhatt et al., 2021) for the complexities of the baselines, and we refer to Appendix D.2 of (Baudry et al., 2022) for details of all computation, and the expression of the complexities according to the algorithms' parameters.

For QoMax-ETC, the memory needed is $Kb_T$ as we only store the current maximum of each batch during the exploration phase. The time complexity is $\mathcal{O}(\max(Kb_T n_T, b_T \log b_T))$ by comparing the duration of the experiment and the time needed to compute the quantiles before exploiting. We just assumed that finding a quantile of a list of size $n$ costs $\Omega(\log(n))$, plugging the values of $b_T$ and $n_T$ gives the result. The time complexity of QoMax-SDA is in $\mathcal{O}(KT \log T)$ as its main cost consists in sorting data online, just like MaxMedian. The storage of QoMax-SDA is obtained thanks to the two tricks: one allows to keep $\mathcal{O}(\log T)$ batches, the other $\mathcal{O}(\log T)$ samples per batch *for the leader*. On the contrary, the complexity for the challengers remains in $\mathcal{O}(\log T \log\log T)$, therefore the dependency in $K$ only appears as a second order term.

**Table 4.1** – Average time and storage complexities of Extreme Bandit algorithms for a time horizon $T$.

| Algorithm | Memory | Time |
|---|---|---|
| Extreme Hunter | $T$ | $\mathcal{O}(T^2)$ |
| MaxMedian | $T$ | $\mathcal{O}(KT \log T)$ |
| **QoMax-SDA** | $\mathcal{O}((\log T)^2)$ | $\mathcal{O}(KT \log T)$ |
| Extreme ETC | $\mathcal{O}\left(K(\log T)^3\right)$ | $\mathcal{O}\left(K(\log T)^6\right)$ |
| **QoMax-ETC** | $\mathcal{O}(K(\log T)^2)$ | $\mathcal{O}(K(\log T)^3)$ |

QoMax-SDA offers an exponential reduction of the storage cost compared to ExtremeHunter and MaxMedian, while being as computationally efficient as MaxMedian. On the other hand, choosing the same length for the exploration phase of the two ETC leads to a significantly smaller time complexity for QoMax-ETC. Hence, both QoMax-SDA and QoMax-ETC present a substantial improvement over their counterparts.

### 4.5.2 Empirical Performance

We compare the empirical performance of the QoMax algorithms with the different competitors on synthetic data. We reproduced 6 experiments from previous works: all experiments from (Bhatt et al., 2021) (Experiments 1-4 for us), and the experiments 1 and 2 from (Carpentier and Valko, 2014) (5-6 in this work). We also implement new experiments with other families of distributions to highlight the generality of our approach. We choose to present in this section our methodology for evaluating Extreme Bandits algorithms, and the results for a selection of experiments, that illustrate well our findings across all the settings we tested. The results for the other experiments can be found in (Baudry et al., 2022).

**Empirical evaluation** We consider 4 performance criteria:

I Empirical evaluation of the extreme regret, by averaging the maxima collected on each trajectory.

II Fraction of pulls of the optimal arm.

III Empirical distribution of the number of pulls of optimal arm, to observe potential failures for some trajectories or how greedy an algorithm can be.

IV Empirical distribution of the maximal reward and in particular some quantiles, that will be more robust that the empirical expected maximum.

In our experiments each criterion is estimated over $N = 10^4$ independent trajectories for different values of the horizon $T$. Most works report only criteria **(I)**, and **(II)** was first proposed by (Bhatt et al., 2021). Our analysis shows that the extreme regret of a strategy is closely related to its capacity to sample the optimal arm $T - o(T)$ times, so we think that **(II)** is indeed a good performance indicator. Criterion **(III)** gives a broader picture of what can happen in the experiments, and in our results we display the following quantiles of the empirical distribution of best arm pulls: $[1\%, 10\%, 25\%, 50\%, 75\%, 90\%, 99\%]$. In particular, the lower quantiles are interesting to see how much an algorithm can under-sample the best arm.

Regarding **(I)**, we note that estimating the expectation $\mathbb{E}[\max_{t\leq T} X_{I_t,t}]$ featured in the extreme regret is very hard: we highlight the fact that this expectation is taken on the *distribution of the maximum over $T$ samples*, which is heavy-tailed in many cases. For that reason, standard Monte-Carlo estimators will have a very large variance. Hence, we propose the following estimation strategy *when a tight approximation of $\mathbb{E}[Y^+_{1,T}]$ is known*. We first find the value $q^+_T$ such that $\mathbb{E}[X^+_{1,T}]$ is equal to the quantile of order $q^+_T$ of the distribution of $Y^+_{1,T}$. We then compute the empirical quantile of order $q^+_T$ of the set of maxima collected in each trajectory, denoted by

$X_{\lceil q_T^+ \rceil}$, as an estimator of their expected maximum. This allows to compute what we call Proxy Empirical Regret (PER), that we define below.

**Definition 4.20** (Proxy Empirical Regret (PER))**.** *Assume that we perform $N$ runs of the algorithm up to horizon $T$. Further assume that $\mathbb{E}[Y_{1,T}^+]$ is known, and that the distribution of $Y_{1,T}^+$ is also known and is denoted by $\nu_T^+$. Let $q_T^+$ be the quantile of $\nu^+$ satisfying $VaR_{q_T^+}(\nu_T^+) = \mathbb{E}[Y_{1,T}^+]$. Then, the* Proxy Empirical Regret *is computed by replacing the empirical estimate of $\mathbb{E}[X_T^+]$ by the empirical estimate of $q_T^+$ in the regret definition, namely*

$$\mathcal{R}_{T,\pi}^{proxy} = \frac{\mathbb{E}[Y_{1,T}^+] - X_{\lceil q_T^+ \rceil}^+}{\mathbb{E}[Y_{1,T}^+]} ,$$

*where $X_1^+, \dots, X_N^+$ are the maxima obtained in each of the $N$ runs. The normalization aims at facilitating the check of a weakly vanishing regret.*

More precisely, by definition $q_T^+$ satisfies $q_T^+ = \mathbb{P}(X_T^+ \leq \mathbb{E}[X_T^+]) \approx \exp(-TG(\mathbb{E}[X_T^+]))$. In the experiments we plug the equivalents of $\mathbb{E}[X_T^+]$ in each setting: for Pareto distribution we obtain $q_T^+ \approx \exp\left(-\frac{1}{\Gamma(1-1/\lambda)^\lambda}\right)$, while for exponential distributions we obtain $q_T^+ \approx e^{-1}$.

We are able to compute the PER to approximate **(I)** for experiments 1-6. When **(I)** is not available we recommend looking at **(IV)** with the same quantiles as for **(III)**.

**Experiments** We describe the settings of a selection of experiments from (Baudry et al., 2022) that we present in this section. We will then refer to them by their number (e.g exp.1), keeping the same numbering as in the paper.

- Experiment 1 (exp.1 in (Bhatt et al., 2021)): $K = 5$ Pareto distributions with tail parameters $\lambda_k \in [2.1, 2.3, 1.3, 1.1, 1.9]$.
- Experiment 3 (exp.3 in (Bhatt et al., 2021)) $K = 10$ Exponential arms with a survival function $G_k(x) = e^{-\lambda_k x}$ with parameters $\lambda_k = [2.1, 2.4, 1.9, 1.3, 1.1, 2.9, 1.5, 2.2, 2.6, 1.4]$.
- Experiment 6 (exp.2 in (Carpentier and Valko, 2014)) $K = 3$ arms, including 2 Pareto distributions with $\lambda_k \in [1.5, 3]$, and arm 3 is a mixture Dirac/Pareto: pulls 0 with $80\%$ probability, reward from a Pareto distribution with $\lambda = 1.1$ with $20\%$ probability. Hence, the last arm dominates asymptotically.
- Experiment 7: we consider $K = 5$ log-normal arms with parameters $\mu_k \in [1, 1.5, 2, 3, 3.5]$ and $\sigma_k \in [4, 3, 2, 1, 0.5]$. When $T$ is large enough the parameter $\sigma$ determines which arm dominates (arm 1 in our case). We recall that if $X$ follows a log-normal distribution with parameters $(\mu, \sigma)$ then $\log(X) \sim \mathcal{N}(\mu, \sigma)$

The code to reproduce the experiments is available on Github.

**Objective of each experiment** Before reporting the results, we explain why each experiment is interesting for the empirical evaluation of Extreme Bandits algorithms. Experiment 1 is quite difficult because the tail gap between arm 3 and arm 4 is relatively small. All algorithms are supposed to have guarantees in this setting so their comparison is fair. Experiments 3 allows to test the different algorithms with exponential tails, showing the performance of the algorithms when the tails are not polynomial. Experiment 6 will be interesting for discussing the limits of parameter-free approaches, as the dominant tail provides low rewards with relatively high probability. Finally, experiment 7 allows to try the algorithms on heavy-tail distributions that do not have a polynomial tail.

**Parameters** We recall the parameters used for the different experiments. For each experiment, we run $N = 10^4$ independent trajectories for 10 time horizons $T \in [10^3, 2.5 \times 10^3, 5 \times 10^3, 7.5 \times 10^3, 9 \times 10^3, 10^4, 1.5 \times 10^4, 2 \times 10^4, 3 \times 10^4, 5 \times 10^4]$. This methodology allows for a fair comparison between ETC and more adaptive strategies, and can stabilize the results as an extreme trajectory can introduce bias for only one time horizon. The parameters we used for each algorithm are the following:

- ThresholdAscent: $s = 100, \delta = 0.1$, as suggested in (Streeter and Smith, 2006b).
- ExtremeETC/ExtremeHunter: $b = 1$, as in (Carpentier and Valko, 2014). As the authors, we use $\delta = 0.1$ for the experiments instead of the theoretical value that is too large for the time horizons considered, and $D = E = 10^{-3}$ for the UCB. Other theoretically-motivated parameters are $r = T^{-1/(2b+1)}$ (fraction of samples used for the tail estimation), $N = (\log T)^{\frac{2b+1}{b}}$ (length of the initial exploration phase). $\delta = \exp(-\log^2(T))/(2TK)$ in the paper but set to $0.1$ here.
- MaxMedian: The exploration probability is set to $\varepsilon_t = 1/(1+t)$ as suggested in (Bhatt et al., 2021).
- QoMax-ETC: We test $q = 1/2$ and $q = 0.9$, $b_T = (\log T)^2$ and $n_T = \log T$ to match both the theoretical requirements of Section 4.3 and the length of the exploration phase of ExtremeETC for a fair comparison.
- QoMax-SDA: $f(r) = (\log r)^{\frac{1}{\gamma}}$ and $B(n) = n^\gamma$ for $\gamma = 2/3$, which works well across all the experiments. The quantile is either equal to $q = 1/2$ or $q = 0.9$.

**Results** For each experiment, we report the results according to the criteria **(I)-(IV)** that are defined above. The criteria **(I)-(II)** are reported side by side for each experiment in Figures 4.2-4.8, except for Exp.7 for which (I) cannot be computed. Tables 4.3-4.9 associated with **(III)**

report the result for the statistics on the number of pulls of the best arm on all trajectories at $T = 5 \times 10^4$ . Finally, Tables 4.4-4.10 related to **(IV)** report the results for the statistics on the empirical distribution of the maxima on all trajectories at $T = 5 \times 10^4$.

We summarize our key observations on the results with the following points:

- **Non-robustness of reporting the average maximum collected.** Several examples can serve to illustrate this point. For experiment 1 (Table 4.4) if we look at the average maximum only, we would conclude that QoMax-SDA with $q = 1/2$ is by far the best algorithm with an average of $1.8 \times 10^5$ ($1.1 \times 10^5$ for the second). However, we see that the quantiles of the maxima distributions are almost identical to those of other QoMax algorithms. Hence, even if $99\%$ of their distribution matches, QoMax-SDA with $q = 1/2$ has a nearly $70\%$ better average caused by less than $1\%$ of the trajectories. The same thing seems to happen on different problems: the $10^4$ and $8.5 \times 10^3$ of $1/2$-QoMax-ETC and ExtremeETC are clearly over-estimated means in experiment 2 considering that they both have the same quantiles as $1/2$-QoMax-SDA (even a bit worse), which has an average of $7.5 \times 10^3$, and MaxMedian with $7.9 \times 10^3$. Without surprise, this phenomenon is more present when the tails are heavier. Hence looking at the average maxima is meaningful only with the statistics from Experiments 3 and 4 with lighter tails.

- **QoMax Performance.** QoMax algorithms clearly outperform their competitors in Experiments **1, 3, 7** according to all criteria. As those experiments include polynomial, exponential and log-normal tails with different number of arms, this shows the generality and efficiency of the QoMax approach. QoMax-SDA seems to work better than QoMax-ETC, in particular it is competitive even for small time horizons ($T < 5 \times 10^3$) in most experiments. However, we see that QoMax-ETC almost matches the performance of QoMax-SDA for $T = 5 \times 10^4$. For a practitioner who would be interested in larger time horizons QoMax-ETC seems to be a perfectly suitable choice.

- On the contrary, **ExtremeHunter** performs significantly better than **ExtremeETC** for larger horizons: the probability of mistake of the latter is still quite large, and the ability of ExtremeHunter to recover from a mistake is valuable, but we recall that the time complexity of ExtremeHunter is detrimental for the practitioner. Results from Experiment **3** show that the two algorithms are not able to handle exponential tails.

- **ThresholdAscent** is never the best algorithm but has the advantage of being consistently better than the uniform strategy (according to **(II)**), as it always pulls the best arm at a frequency larger that $1/K$. It is the most stable baseline in terms of **(III)** (it always has the narrowest range for the statistics we consider), but this is detrimental to its capacity to collect large values.

- We tested **MaxMedian** on larger time horizons than in the original paper, which explains the difference in some results. Indeed, we observe that in Experiments **1, 3, 5**, MaxMedian is quite competitive for shorter time horizons ($T \leq 10^4$), but almost stops improving at this step. This suggests that the algorithm does not explore enough, which is confirmed by a closer look at **(III)**: the number of pulls of the best arm are either very close to $0\%$ or to $100\%$ in most of the cases, which is a behavior specific to this algorithm and that we would like to avoid in practice. This behavior also has an impact on the statistics on the maxima distributions **(IV)**. The exploration function may be responsible for this, and results may be better with a larger forced exploration.

- **Experiment 6** shows that in some examples parametric algorithms can perform much better than non-parametric approaches. Indeed, the distribution of arm $3$ enters in the second-order Pareto family, and the parameter $b = 1$ makes ExtremeHunter calibrate its parameters with the $\approx 5\%$ best samples of each arm. This is enough for the algorithm to "detect" the Pareto tail of the mixture and sample it most often. Most of the other algorithms fail, including QoMax. However, two important remarks on QoMax can be made based on this experiment: the $0.9$-QoMax-SDA performs much better than the others, showing that when the tails are harder to detect choosing a larger quantile can be valuable. Furthermore, we tested another experiment imposing at least $100$ samples in each batch. This time, $0.9$-QoMax-SDA was able to pull the best arm $60\%$ of the time. Hence, the practitioner has the ability to increase the exploration and the quantile $q$ if very difficult tails are expected (typically where the value $x$ of Definition 4.3 is large), which depends on the characteristics of the problem at hand.

**Conclusion of the experiments** Overall, QoMax-based algorithms seem to be solid choices for the practitioner, as demonstrated in a variety of examples. Their strong theoretical guarantees and implementation tricks reducing the time and space complexities make them an efficient solution for the Extreme Bandits problem. They work well on most examples that we tried **with the same parameters** (avoiding painful tuning), including settings with different kind of tails (polynomial, exponential, log-normal) with different number of arms, and both easy and hard instances. We explained however with experiment 6 the limits of a distribution-free approach if we consider a hard problem. It also showed that in this case augmenting the quantile $q$ (and/or the forced exploration function $f$ for QoMax-SDA) used in QoMax algorithms can be beneficial. Furthermore, we can recommend to use QoMax-ETC when the time horizon is expected to be very large (larger than $5 \times 10^4$ for instance) and QoMax-SDA for smaller time horizons, as it seems to learn faster on all examples but it is more computationally demanding.

## Experiment 1

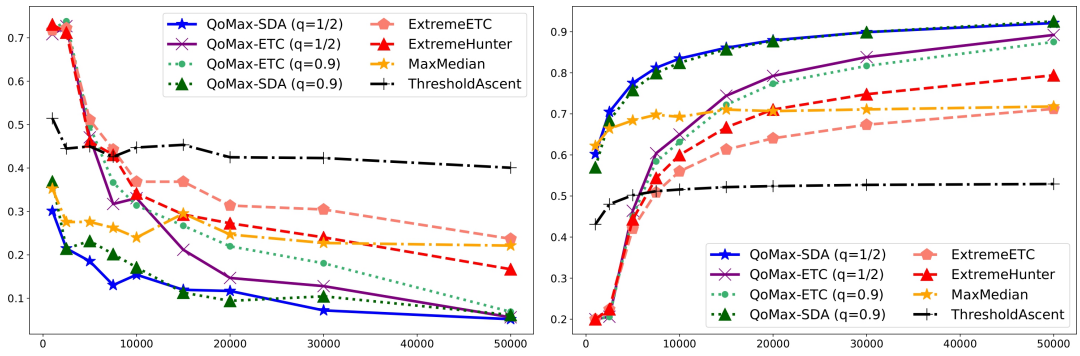


**Figure 4.2** – Experiment 1: Proxy Empirical Regret (left) and Number of pulls of the dominant arm (right), averaged over $10^4$ independent trajectories for $T \in \{10^3, 2.5 \times 10^3, 5 \times 10^3, 7.5 \times 10^3, 9 \times 10, 10^4, 1.5 \times 10^4, 2 \times 10^4, 3 \times 10^4, 5 \times 10^4\}$.

**Figure 4.3** – Statistics on the number of pulls of the best arm at $T = 5 \times 10^4$, Experiment 1.

| Algorithm | Average (%) | 1% | 10% | 25% | 50% | 75% | 90% | 99% |
|---|---|---|---|---|---|---|---|---|
| QoMax-SDA ($q = 1/2$) | 92 | 42 | 90 | 93 | 94 | 95 | 95 | 95 |
| QoMax-SDA ($q = 0.9$) | 93 | 14 | 87 | 93 | 96 | 97 | 98 | 98 |
| QoMax-ETC ($q = 1/2$) | 89 | 90 | 90 | 90 | 90 | 90 | 90 | 90 |
| QoMax-ETC ($q = 0.9$) | 88 | 3 | 90 | 90 | 90 | 90 | 90 | 90 |
| ExtremeETC | 71 | 3 | 3 | 90 | 90 | 90 | 90 | 90 |
| ExtremeHunter | 79 | 3 | 5 | 89 | 90 | 90 | 90 | 90 |
| MaxMedian | 72 | 0 | 0 | 0 | 100 | 100 | 100 | 100 |
| ThresholdAscent | 53 | 46 | 50 | 52 | 53 | 55 | 56 | 57 |

**Figure 4.4** – Statistics on the distributions of maxima at $T = 5 \times 10^4$, Experiment 1. Results divided by 100 to improve readability.

| Algorithm | Average | 1% | 10% | 25% | 50% | 75% | 90% | 99% |
|---|---|---|---|---|---|---|---|---|
| QoMax-SDA ($q = 1/2$) | 1852 | 41 | 81 | 130 | 245 | 547 | 1350 | 11371 |
| QoMax-SDA ($q = 0.9$) | 1042 | 39 | 78 | 128 | 239 | 529 | 1363 | 12539 |
| QoMax-ETC ($q = 1/2$) | 1058 | 40 | 79 | 126 | 232 | 530 | 1324 | 11054 |
| QoMax-ETC ($q = 0.9$) | 919 | 34 | 75 | 122 | 230 | 511 | 1301 | 10080 |
| ExtremeETC | 882 | 16 | 44 | 86 | 183 | 426 | 1089 | 9515 |
| ExtremeHunter | 1092 | 21 | 61 | 104 | 208 | 477 | 1226 | 9799 |
| MaxMedian | 785 | 3 | 37 | 83 | 180 | 436 | 1126 | 9240 |
| ThresholdAscent | 748 | 27 | 51 | 82 | 156 | 351 | 853 | 7771 |

## Experiment 3

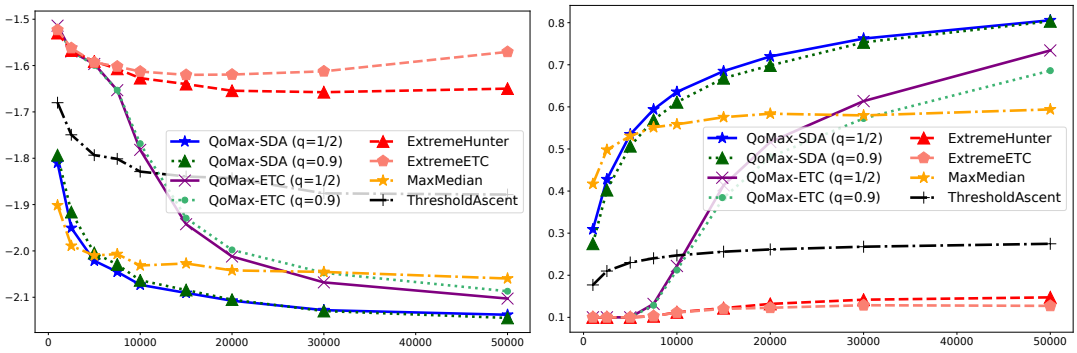


**Figure 4.5** – Experiment 3: Proxy Empirical Regret (left) and Number of pulls of the dominant arm (right), averaged over $10^4$ independent trajectories for $T \in \{10^3, 2.5 \times 10^3, 5 \times 10^3, 7.5 \times 10^3, 9 \times 10, 10^4, 1.5 \times 10^4, 2 \times 10^4, 3 \times 10^4, 5 \times 10^4\}$.

**Figure 4.6** – Statistics on the number of pulls of the best arm at $T = 5 \times 10^4$, Experiment 3.

| Algorithm | Average (%) | 1% | 10% | 25% | 50% | 75% | 90% | 99% |
|---|---|---|---|---|---|---|---|---|
| QoMax-SDA ($q = 1/2$) | 81 | 2 | 72 | 82 | 86 | 88 | 88 | 89 |
| QoMax-SDA ($q = 0.9$) | 80 | 2 | 59 | 80 | 87 | 91 | 93 | 95 |
| QoMax-ETC ($q = 1/2$) | 73 | 3 | 77 | 77 | 77 | 77 | 77 | 77 |
| QoMax-ETC ($q = 0.9$) | 69 | 3 | 3 | 77 | 77 | 77 | 77 | 77 |
| ExtremeETC | 13 | 3 | 3 | 3 | 3 | 3 | 77 | 77 |
| ExtremeHunter | 15 | 3 | 3 | 3 | 3 | 7 | 67 | 77 |
| MaxMedian | 59 | 0 | 0 | 0 | 98 | 100 | 100 | 100 |
| ThresholdAscent | 27 | 21 | 24 | 26 | 28 | 29 | 31 | 33 |

**Figure 4.7** – Statistics on the distributions of maxima at $T = 5 \times 10^4$, Experiment 3.

| Algorithm | Average | 1% | 10% | 25% | 50% | 75% | 90% | 99% |
|---|---|---|---|---|---|---|---|---|
| QoMax-SDA ($q = 1/2$) | 32 | 26 | 28 | 29 | 31 | 34 | 37 | 43 |
| QoMax-SDA ($q = 0.9$) | 32 | 25 | 28 | 29 | 31 | 34 | 37 | 44 |
| QoMax-ETC ($q = 1/2$) | 32 | 25 | 28 | 29 | 31 | 34 | 36 | 43 |
| QoMax-ETC ($q = 0.9$) | 31 | 24 | 27 | 29 | 31 | 33 | 36 | 43 |
| ExtremeETC | 26 | 18 | 21 | 23 | 25 | 29 | 32 | 39 |
| ExtremeHunter | 27 | 19 | 22 | 23 | 26 | 29 | 32 | 39 |
| MaxMedian | 31 | 21 | 25 | 28 | 31 | 33 | 36 | 43 |
| ThresholdAscent | 29 | 23 | 25 | 27 | 29 | 31 | 34 | 41 |

## Experiment 6

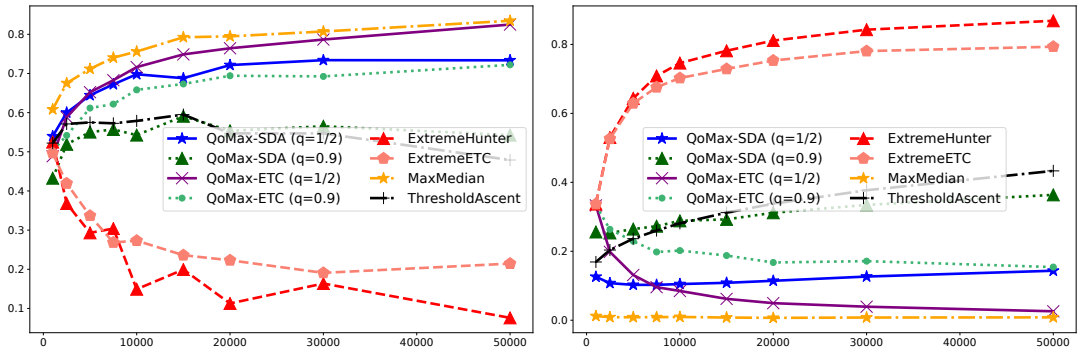


**Figure 4.8** – Experiment 6: Proxy Empirical Regret (left) and Number of pulls of the dominant arm (right), averaged over $10^4$ independent trajectories for $T \in \{10^3, 2.5 \times 10^3, 5 \times 10^3, 7.5 \times 10^3, 9 \times 10, 10^4, 1.5 \times 10^4, 2 \times 10^4, 3 \times 10^4, 5 \times 10^4\}$.

**Figure 4.9** – Statistics on the number of pulls of the best arm at $T = 5 \times 10^4$, Experiment 6.

| Algorithm | Average (%) | 1% | 10% | 25% | 50% | 75% | 90% | 99% |
|---|---|---|---|---|---|---|---|---|
| QoMax-SDA ($q = 1/2$) | 14 | 1 | 1 | 2 | 4 | 15 | 45 | 95 |
| QoMax-SDA ($q = 0.9$) | 36 | 0 | 1 | 3 | 22 | 75 | 90 | 98 |
| QoMax-ETC ($q = 1/2$) | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| QoMax-ETC ($q = 0.9$) | 15 | 3 | 3 | 3 | 3 | 3 | 95 | 95 |
| ExtremeETC | 79 | 3 | 3 | 95 | 95 | 95 | 95 | 95 |
| ExtremeHunter | 87 | 3 | 85 | 95 | 95 | 95 | 95 | 95 |
| MaxMedian | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| ThresholdAscent | 43 | 27 | 34 | 38 | 43 | 49 | 53 | 60 |

**Figure 4.10** – Statistics on the distributions of maxima at $T = 5 \times 10^4$, Experiment 6. Results divided by 100 to improve readability.

| Algorithm | Average | 1% | 10% | 25% | 50% | 75% | 90% | 99% |
|---|---|---|---|---|---|---|---|---|
| QoMax-SDA ($q = 1/2$) | 60 | 5 | 9 | 12 | 21 | 41 | 91 | 635 |
| QoMax-SDA ($q = 0.9$) | 120 | 6 | 10 | 15 | 28 | 64 | 155 | 1144 |
| QoMax-ETC ($q = 1/2$) | 40 | 5 | 8 | 11 | 18 | 33 | 64 | 306 |
| QoMax-ETC ($q = 0.9$) | 59 | 5 | 8 | 12 | 20 | 40 | 93 | 702 |
| ExtremeETC | 267 | 6 | 14 | 24 | 47 | 108 | 266 | 2687 |
| ExtremeHunter | 232 | 8 | 17 | 28 | 53 | 116 | 305 | 2620 |
| MaxMedian | 35 | 0 | 7 | 10 | 17 | 30 | 60 | 306 |
| ThresholdAscent | 136 | 7 | 12 | 18 | 33 | 70 | 170 | 1299 |

## Experiment 7

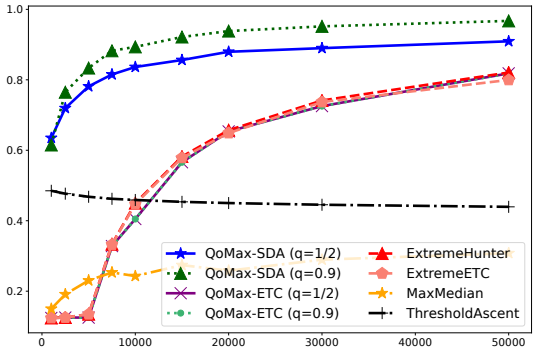


**Figure 4.11** – Experiment 7 (Log-normal arms): Number of pulls of the dominant arm, averaged over $10^4$ independent trajectories for $T \in \{10^3, 2.5 \times 10^3, 5 \times 10^3, 7.5 \times 10^3, 9 \times 10, 10^4, 1.5 \times 10^4, 2 \times 10^4, 3 \times 10^4, 5 \times 10^4\}$.

**Table 4.2** – Statistics on the distributions of number of pulls of the best arm at $T = 5 \times 10^4$, Exp. 7

| Algorithm | Average | 1% | 10% | 25% | 50% | 75% | 90% | 99% |
|---|---|---|---|---|---|---|---|---|
| QoMax-SDA ($q = 1/2$) | 94 | 85 | 94 | 95 | 95 | 95 | 95 | 95 |
| QoMax-SDA ($q = 0.9$) | 97 | 89 | 96 | 97 | 98 | 98 | 98 | 98 |
| QoMax-ETC ($q = 1/2$) | 90 | 90 | 90 | 90 | 90 | 90 | 90 | 90 |
| QoMax-ETC ($q = 0.9$) | 90 | 90 | 90 | 90 | 90 | 90 | 90 | 90 |
| ExtremeETC | 55 | 3 | 3 | 3 | 90 | 90 | 90 | 90 |
| ExtremeHunter | 63 | 13 | 40 | 45 | 53 | 90 | 90 | 90 |
| MaxMedian | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| ThresholdAscent | 57 | 55 | 56 | 57 | 58 | 58 | 58 | 58 |

**Table 4.3** – Statistics on the distributions of maxima at $T = 5 \times 10^4$, Experiment 7. Results divided by 1000 to improve readability.

| Algorithm | Average | 1% | 10% | 25% | 50% | 75% | 90% | 99% |
|---|---|---|---|---|---|---|---|---|
| QoMax-SDA ($q = 1/2$) | 1393 | 73 | 151 | 257 | 488 | 1090 | 2259 | 13764 |
| QoMax-SDA ($q = 0.9$) | 1401 | 79 | 163 | 260 | 524 | 1171 | 2830 | 13839 |
| QoMax-ETC ($q = 1/2$) | 1337 | 77 | 154 | 245 | 430 | 1007 | 2664 | 13651 |
| QoMax-ETC ($q = 0.9$) | 1459 | 84 | 150 | 251 | 461 | 987 | 2419 | 12654 |
| ExtremeETC | 957 | 6 | 12 | 30 | 214 | 581 | 1511 | 7422 |
| ExtremeHunter | 867 | 32 | 85 | 156 | 297 | 666 | 1569 | 10855 |
| MaxMedian | 76 | 0 | 0 | 0 | 0 | 0 | 15 | 1678 |
| ThresholdAscent | 1043 | 43 | 94 | 160 | 311 | 667 | 1648 | 10715 |

## 4.6 Appendix A: proofs of section 4.4 (SDA)

### 4.6.1 Proof of Lemma 4.18

We recall $\xi_T := \{N_1(T) \leq T - KM_T\}$. First, using $\sum_{k=1}^K N_k(T) = T$, we remark that

$$\mathbb{P}(N_1(T) \leq T - KM_T) \leq \mathbb{P}(\exists k \geq 2, N_k(T) \geq M_T) \leq \sum_{k=2}^K \mathbb{P}(N_k(T) \geq M_T) \ ,$$

We denote by $r_T$ the index of the round for which the number of observations equals or exceeds $T$. As at least one observation is collected at the end of the round it holds that $r_T \leq T$. Hence, we can obtain

$$\mathbb{P}(N_1(T) \leq T - KM_T) \leq \sum_{k=2}^K \mathbb{P}(n_k(r_T)b_k(r_T) \geq M_T) \leq \sum_{k=2}^K \mathbb{P}(n_k(T)b_k(T) \geq M_T) \ .$$

Using $b_k(T) = n_k(T)^\gamma$ and Markov inequality gives

$$\mathbb{P}(N_1(T) \leq T - KM_T) \leq \sum_{k=2}^K \mathbb{P}(n_k(T)^{1+\gamma} \geq M_T) \leq \sum_{k=2}^K \mathbb{P}(n_k(T) \geq M_T^{\frac{1}{1+\gamma}}) \leq \sum_{k=2}^K \frac{\mathbb{E}[n_k(T)]}{M_T^{\frac{1}{1+\gamma}}} \ .$$

For all $k \geq 2$, Lemma 4.21 shows that $\mathbb{E}[n_k(T)] = \mathcal{O}\left((\log T)^{\frac{1}{\gamma}}\right)$ with the tuning we choose for the algorithm. This is sufficient to conclude on the result.

**Lemma 4.21.** *Under assumption 4.11 and if $\nu_1 \succ \nu_k$ for all $k \in \{1, \dots, K\}$, if we consider QoMax-SDA with parameters $B(n) = n^\gamma$ and $f(r) = (\log r)^{\frac{1}{\gamma}}$, then for all $k \geq 2$ there exists a constant $C_k$ such that the number of pulls of arm $k$ at time $T$ satisfies*

$$\mathbb{E}[n_k(T)] \leq C_k \left(\log(T)\right)^{\frac{1}{\gamma}} + \mathcal{O}(1) \ .$$

We omit the proof of this result in this manuscript, as it is redundant with the proofs of Chapters 2 and 3. The interested reader can find the complete version in Baudry et al. (2022). The main ingredient is to replace the deviation inequalities of Assumption 2.4 by the ones we obtained for the QoMax estimates in Theorem 4.7. Furthermore, in this setting the forced exploration of $f(r) = (\log r)^{\frac{1}{\gamma}}$ queries (and hence $\log(r)$ batches due to the rule we set for the number of batches) ensures enough forced exploration so that we do not need to consider balance conditions, which provides the values we set for $f(r)$ and $b(r)$.

### 4.6.2 Proof of Theorem 4.19

**Theorem 4.19** (Upper bound on the regret of QoMax-SDA). *Under the assumptions of Lemma 4.18 it further holds that*

1. *the regret of QoMax-SDA is vanishing in the strong sense for exponential tails*
2. *the regret of QoMax-SDA is vanishing in the weak sense for polynomial tails.*

*Proof.* We instantiate the decomposition of Proposition 4.16 using the value of $\mathbb{P}(\xi_T)$ obtained in Lemma 4.18. Plugging all of these values and using Proposition 4.12, we write for $\pi$ being any instance of QoMax-SDA with parameter $\gamma$,

$$\begin{aligned}\mathcal{R}_T^\pi &\leq \underbrace{\mathbb{E}\left[\max_{t\leq T} Y_{1,t}\right] - \mathbb{E}\left[\max_{t\leq T-KM_T} Y_{1,t}\right]}_{\text{Exploration cost}} + \underbrace{x_T\mathbb{P}(\xi_T) + \mathbb{E}\left[\max_{t\leq T} Y_{1,t}\mathbb{1}\left(\max_{t\leq T} Y_{1,t}\geq x_T\right)\right]}_{\text{Cost incurred by } \xi_T} \\ &\leq \underbrace{KM_T\left[\frac{B_T}{T} + \int_{B_T}^{+\infty} G_1(x)dx\right]}_{A_1} + \underbrace{x_T\frac{C(\log T)^{\frac{1}{\gamma}}}{M_T^{\frac{1}{1+\gamma}}}}_{A_2} + \underbrace{T\int_{x_T}^{+\infty} G_1(x)dx}_{A_3},\end{aligned}$$

for any values of $x_T, B_T, M_T$, that we now specify for each of the two families considered.

**Exponential tails** We recall that if $G_1(x) = \mathcal{O}(\exp(-\lambda x))$, then for any $y \in \mathbb{R}$ we have

$$\int_y^{+\infty} G_1(x)dx = \mathcal{O}\left(\exp(-\lambda y)\right) .$$

First, if we choose $B_T = \frac{1}{\lambda}\log(T)$ then $A_1$ vanishes for any choice of $M_T = T^\alpha$ with $0 < \alpha < 1$. Similarly, choosing $x_T = \frac{2}{\lambda}\log T$ ensures that $A_3 = \mathcal{O}(1/T)$. Then, $A_2$ is $\mathcal{O}\left(\frac{(\log T)^{1+\frac{1}{\gamma}}}{M_T^{\frac{1}{1+\gamma}}}\right)$, which is vanishing for any choice of $M_T = T^\alpha$, $\alpha \in (0,1)$. We conclude that for exponential tails, $\lim_{T\to\infty}\mathcal{R}_T^\pi = 0$.

**Polynomial tails** Consider again $M_T = T^\alpha$, for some $\alpha \in (0,1)$. This time,

$$\int_y^{+\infty} G_1(x)dx = \mathcal{O}\left(\frac{1}{y^{\lambda-1}}\right) .$$

Plugging into $A_3$, we get a term of order $\mathcal{O}(T \times x_T^{1-\lambda})$. Let's take $x_T = T^\beta$ for some $\beta \in (0,1)$, we then have

$$A_3 = \mathcal{O}(T^{1+\beta(1-\lambda)}) \ .$$

Now consider $A_2$, omitting the polylog terms we obtain

$$A_2 = \mathcal{O}(T^{\beta - \frac{\alpha}{1+\gamma}}) \ .$$

Consider finally $A_1$. Choosing $B_T = T^{\frac{1}{\lambda}}$ (as in Appendix **??**) we obtain the tightest upper bound on the exploration cost:

$$A_1 = \mathcal{O}\left(\frac{M_T}{T^{1-\frac{1}{\lambda}}}\right) = \mathcal{O}(T^{\alpha - 1 + \frac{1}{\lambda}}) \ .$$

To get the smallest order with this proof technique we want to equalize all these three exponents, which gives

$$\alpha - 1 + \frac{1}{\lambda} = \beta - \frac{\alpha}{1+\gamma} = 1 + \beta(1-\lambda) \ .$$

For simplicity we write $\beta = \frac{1}{\lambda} + \eta$ and try to find $\eta$ instead. Re-writing the the three equalities yields

$$\alpha + \frac{1}{\lambda} - 1 = \frac{1}{\lambda} + \eta - \frac{\alpha}{1+\gamma} = \frac{1}{\lambda} - (\lambda - 1)\eta \ .$$

This can be further simplified in

$$\alpha - 1 = \eta - \frac{\alpha}{1+\gamma} = -(\lambda - 1)\eta \ .$$

This gives in particular a system of two equations with two unknowns $\eta$ and $\alpha$. By substituing $\alpha$ we get

$$\begin{aligned} \eta - \frac{1 - (\lambda - 1)\eta}{1+\gamma} &= -(\lambda - 1)\eta \\ \Leftrightarrow \eta\left[1 + \gamma + \lambda - 1 + (\lambda - 1)(1+\gamma)\right] &= 1 \ , \end{aligned}$$

which gives $\eta = \frac{1}{\lambda(2+\gamma)-1}$ and $\alpha = \frac{\lambda(1+\gamma)}{\lambda(2+\gamma)-1}$.

Plugging in these values, we obtain that $A_1$, $A_2$ and $A_3$ are all in $\mathcal{O}\left(T^{\frac{1}{\lambda}-\frac{\lambda-1}{\lambda(2+\gamma)-1}}\right) = o(T^{1/\lambda})$. Plugging the rate of growth of the maximum for polynomial tails we get that

$$\mathcal{R}_T^{\pi} = \underset{T\to\infty}{o}\left(\mathbb{E}\left[\max_{t\leq T} Y_{1,t}\right]\right).$$

□

**Remark 4.22.** *With this last result we can further say that the regret of QoMax-SDA is strongly vanishing if $\lambda_1$ is larger than $2+\sqrt{3}$, but we omitted this result in our statement since it is impossible to know in advance.*

## 4.7 Appendix B: Implementation Tricks for QoMax-SDA

In this section we provide both empirical and theoretical evidences on the memory gains obtained with the implementation trick we introduced for the storage of maxima in QoMax-SDA. We first recall this procedure in Algorithm 4.6

**Input:** List of indices $\mathcal{I} = \{i_1, \dots, i_L\}$, sorted list $\mathcal{X} = \{X_1, \dots, X_L\}$, $X_1 > X_2 > \cdots > X_L$, new index $i$, new data $X$
Find the largest $j \in \{1, \dots, L\}$ satisfying $X_j > X$ ; ▷ Binary search
Set $\mathcal{X} = \{X_1, X_2, \dots, X_j, X\}$ ; ▷ Remove $X_{j+1}, \dots, X_L$ and add $X$
Set $\mathcal{I} = \{i_1, \dots, i_j, i\}$ ; ▷ Remove $i_{j+1}, \dots, i_L$ and add $i$
**Return:** List of indices $\mathcal{I}$, list of data $\mathcal{X}$.

**Algorithm 4.6:** Efficient Update of a list of maxima for QoMax-SDA

**Empirical evidences of the efficiency of the storage trick** We propose simulations to verify that the solution we propose to store the data used by QoMAx-SDA is indeed efficient. We performed 1000 simulations for each sample size $N \in [10^2, 5\times10^2, 10^3, 2\times10^3, 5\times10^3, 10^4, 2\times10^4, 3\times10^4, 5\times10^4]$), and for 4 distributions: (1) a Pareto distribution with tail parameter 1.1, (2) a Pareto distribution with tail parameter 3, (3) an exponential distribution with parameter 1, (4) a standard normal distribution. We report in Figure 4.12 the average number of data stored by the algorithm for each sample sizes, along with the empirical 5% and 95% quantiles on the 1000 simulations and the curve $y = \log(N)$ for comparison. We observe that: (1) The results do not depend on the distribution. (2) All 4 curves are very close to exactly $\sum_{n=1}^{N} \frac{1}{n}$, which is as small as $\approx 10$ for a sample size of $5\times10^4$. (3) 90% of the simulations have no more than 17

data stored, and the maximum we observe on all 4 experiments is actually 23 which is very small compared to $N = 5 \times 10^4$.

Therefore, we conclude that the trick we introduced is indeed efficient and our experiments corroborate the intuition that it allows to store $\mathcal{O}(\log N)$ data out of $N$ on average. We now prove it formally in Lemma 4.23

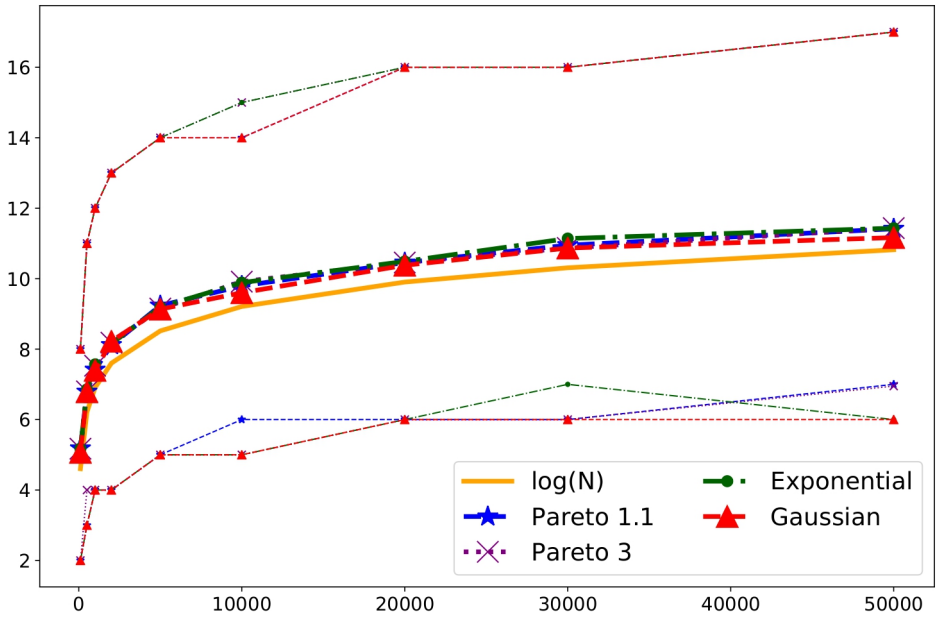


**Figure 4.12** – Average number of data kept in memory with the efficient storage of maxima, for 1000 simulations with sample size $N \in [10^2, 5 \times 10^2, 10^3, 2 \times 10^3, 5 \times 10^3, 10^4, 2 \times 10^4, 3 \times 10^4, 5 \times 10^4]$ and the empirical 5% and 95% quantiles.

**Lemma 4.23** (Expected memory with the efficient storing of maxima). *Denote by $C_N$ the random variable denoting the memory usage of a random i.i.d sample of size $N$ drawn from any distribution with the implementation trick from Alg. 4.6. For any $\nu$, it holds that*

$$\mathbb{E}[C_N] = \sum_{n=1}^{N} \frac{1}{n} \approx \log(N) \ .$$

*Proof.* Denote the sorted random samples by $X_1 > \cdots > X_N$. As the observations are i.i.d, all of them are equally likely to be in the last position. We consider $I_N$ the random variable denoting the index of the last observation, it holds that

$$\mathbb{E}[C_N] = \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}[C_N | I_N = j] \ .$$

Then, we remark that if $I_N = j$, all observations of higher order $X_{j+1}, \ldots, X_N$ are removed from the history. Hence, it only remains to count the average amount of data considering only $X_1, \ldots, X_{j-1}$, which is equal to $\mathbb{E}[C_{j-1}]$ and add 1 for the last observation. Using that $\mathbb{E}[C_1] = 1$,

$$
\begin{aligned}
&\mathbb{E}[C_N] = \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}[C_N|I_N = j] = \frac{1}{N}\sum_{j=1}^{N}(1+\mathbb{E}[C_{j-1}]) \\
&\Rightarrow (N+1)\mathbb{E}[C_{N+1}] - N\mathbb{E}[C_N] = \sum_{j=1}^{N+1}(1+\mathbb{E}[C_j]) - \sum_{j=1}^{N}(1+\mathbb{E}[C_j]) = 1+\mathbb{E}[C_N] \\
&\Rightarrow (N+1)(\mathbb{E}[C_{N+1}] - \mathbb{E}[C_N]) = 1 \\
&\Rightarrow \mathbb{E}[C_{N+1}] = \mathbb{E}[C_N] + \frac{1}{N+1} \\
&\Rightarrow \mathbb{E}[C_N] = \sum_{n=1}^{N}\frac{1}{n}\,.
\end{aligned}
$$

□

# Part II

# Dirichlet Sampling Strategies for Bounded Rewards and Beyond

# Chapter 5

# Non-Parametric Thompson Sampling for CVaR bandits

In this chapter we consider the CVaR bandit problems introduced in Chapter 1 (Section 1.2). While existing works in this setting mainly focus on *Upper Confidence Bound* algorithms, we present a new *Thompson Sampling* approach for CVaR bandits on bounded rewards that is flexible enough to solve a variety of problems with finite rewards. Building on a recent work by Riou and Honda (2020), we introduce B-CVTS for continuous bounded rewards and M-CVTS for multinomial distributions. On the theoretical side, we provide a non-trivial extension of their analysis that enables to upper bound their CVaR regret. Strikingly, our results show that these strategies are the first to provably achieve *asymptotic optimality* in CVaR bandits, matching the corresponding asymptotic lower bounds for this setting. We furthermore illustrate empirically the benefits of Thompson Sampling approaches, both in an experiment in agriculture using the DSSAT simulator and on various synthetic examples. The results presented in this Chapter were published in (Baudry et al., 2021a), and in Section 5.5 we present theoretical results that are not in the paper but are of interest for the practitioner.

## Contents

## 5.1 Introduction

In Section 1.2 of Chapter 1 we presented the literature on bandit algorithms with alternative performance metric, with a focus on *risk-aware* bandits, and more specifically on bandit algorithms evaluating the arms' distributions through their *Conditional Value at Risk*.

The *Conditional Value at Risk* (CVaR) at level $\alpha \in [0, 1]$ (see (Mandelbrot, 1997; Artzner et al., 1999)) is easily interpretable as the expected reward in the worst $\alpha$-fraction of the outcomes, and hence captures different preferences, from being neutral to the shape of the distribution ($\alpha = 1$, mean criterion) to trying to maximize the reward in the worst-case scenarios ($\alpha$ close to 0, typically in finance or insurance). Several definitions of the CVaR exist in the literature, depending on whether the samples are considered as *losses* or as *rewards*. (Brown, 2007; Thomas and Learned-Miller, 2019; Agrawal et al., 2021b) consider the *loss* version of CVaR. We here follow Galichet et al. (2013) and Tamkin et al. (2020) who use the *reward* version, defined for arm $k$ with distribution $\nu_k$ as

$$\text{CVaR}_\alpha(\nu_k) = \sup_{x \in \mathbb{R}} \left\{ x - \frac{1}{\alpha} \mathbb{E}_{X \sim \nu_k} \left[ (x - X)^+ \right] \right\} . \tag{5.1}$$

This implies that the best arm is the one with the *largest* CVaR. To simplify the notation we write $c_k^\alpha = \text{CVaR}_\alpha(\nu_k)$ in the sequel. Following e.g. (Tamkin et al., 2020), for unknown arm distributions $\boldsymbol{\nu} = (\nu_1, \dots, \nu_K)$ we measure the CVaR regret at time $T$ for some risk-level $\alpha$ of a sequential sampling strategy $\mathcal{A} = (A_t)_{t \in \mathbb{N}}$ as

$$\mathcal{R}_{\boldsymbol{\nu}}^\alpha(T) = \mathbb{E}_{\boldsymbol{\nu}} \left[ \sum_{t=1}^{T} \left( \max_k c_k^\alpha - c_{A_t}^\alpha \right) \right] = \sum_{k=1}^{K} \Delta_k^\alpha \mathbb{E}_{\boldsymbol{\nu}}[N_k(T)] , \tag{5.2}$$

where $\Delta_k^\alpha = \max_{k'} c_{k'}^\alpha - c_k^\alpha$ is the gap in CVaR between arm $k$ and the best arm, and $N_k(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = k)$ is the number of selections of arm $k$ up to round $t$. This corresponds to the *proxy regret* of Definition 1.8 using the $\text{CVaR}_\alpha$ as performance metric. However, for the problem we consider (CVaR bandits with bounded distributions) it will be easier to analyze than the regret of definition 1.6, furthermore (Cassel et al., 2018) proved that upper bounds on the proxy regret translates into an upper bound on the true regret. For this reason, we will actually call CVaR regret the proxy regret defined in Equation (5.2) for the rest of this chapter.

**Some basic CVaR properties** Before continuing this chapter we explain some well-known properties of the CVaR in order to improve the intuitions of the un-familiar reader. First, the definition of the CVaR as the solution of an optimization problem was first introduced by Rockafellar et al. (2000), to formalize previous heuristic definitions of the CVaR as an average over a certain part of the distribution. The definition (5.1) is indeed appealing as it applies to any distribution for which $\mathbb{E}[(x - X)^+]$ is defined, including both discrete and continuous

distributions. To understand the CVaR it is particularly useful to look at its expression in these two particular cases. First, for any continuous distribution $\nu$ of CDF $F_\nu$ it can be shown (see, e.g. Acerbi and Tasche (2002)) that

$$\mathrm{CVaR}_\alpha(\nu) = \mathbb{E}_{X\sim\nu}\left[X | X \leq F_\nu^{-1}(\alpha)\right].$$

This expression provides a good intuition on what the CVaR represents, as the expectation of the distribution including only the worst scenarios, representing the fraction $\alpha$ of the total mass. A similar definition exists for real-valued distributions $\nu$ with discrete support $\mathcal{X} = (x_1, x_2, \dots)$ (either finite or infinite). Assuming that the sequence $(x_i)$ is increasing and letting $p_i = \mathbb{P}_{X\sim\nu}(X = x_i)$, one has

$$\mathrm{CVaR}_\alpha(\nu) = \sup_{x_n \in \mathcal{X}} \left\{ x_n - \frac{1}{\alpha} \sum_{i=1}^{n-1} p_i (x_n - x_i) \right\}. \tag{5.3}$$

Indeed, the function to maximize in (5.1) is piece-wise linear, so the maximum is necessarily achieved in a point of discontinuity. In particular, we can easily prove that if $n_\alpha$ is the first index satisfying $\sum_{i=1}^{n_\alpha} p_i \geq \alpha$, then the supremum is achieved in $n_\alpha$ and

$$\begin{aligned}\mathrm{CVaR}_\alpha(\nu) &= x_{n_\alpha} - \frac{1}{\alpha} \sum_{i=1}^{n_\alpha - 1} p_i (x_{n_\alpha} - x_i) \\ &= \frac{1}{\alpha} \left( \sum_{i=1}^{n_\alpha - 1} p_i x_i + \left( \alpha - \sum_{i=1}^{n_\alpha - 1} p_i \right) x_{n_\alpha} \right).\end{aligned}$$

Hence in that case the CVaR can also be seen as an average when we consider the lower part of the distribution before reaching a total mass $\alpha$.

From the general definition (5.1), one can also observe that for $\alpha = 1$, $\mathrm{CVaR}_\alpha(\nu) = \mathbb{E}_{X\sim\nu}(X)$. Moreover, the mapping $\alpha \mapsto \mathrm{CVaR}_\alpha(\nu)$ is continuous on $(0, 1]$. Thus, considering CVaR bandits allows to smoothly interpolate between standard bandits (that correspond to $\alpha = 1$) and risk-averse problems. Finally, the CVaR is also conveniently Lipschitz w.r.t some well-chosen distances between distributions when they are bounded. This property is useful to prove the results presented in this chapter.

**Contributions** Our objective is to find algorithms minimizing the CVaR regret (Eq. 5.2) considering either distributions with discrete, finite support, or with continuous and bounded support, as we believe this is a relevant problem, motivated by the case-study in agriculture introduced at the beginning of this thesis. More precisely, we target first-order *asymptotic optimality* for these (sometimes called "non-parametric") families and first derive in Theorem 5.3 a lower-bound on the CVaR regret, adapting that of (Lai and Robbins, 1985; Burnetas and

Katehakis, 1996) to the CVaR criterion. This result highlights the right complexity term that should appear when deriving regret upper bounds.

On the algorithms side, two powerful variants of Thompson Sampling were introduced recently by Riou and Honda (2020) for the mean criterion, that enable to overcome the "parametric" limitation, in the sense that these approaches reach the minimal achievable regret given by the lower bound of Burnetas and Katehakis (1996) respectively for discrete and bounded distributions: *Multinomial Thompson Sampling (MTS)* and *Non-Parametric Thompson Sampling (NPTS)*. This timely contribution opens the room for a generalization to CVaR bandits, that we introduce in Section 5.2 with B-CVTS for bounded supports, and M-CVTS for multinomial arms. We then show that the two algorithms are *asymptotically optimal* (Theorem 5.4), and sketch the proof of the result for B-CVTS. Up to our knowledge, these are the first results showing asymptotic optimality of a CVaR regret minimization strategy. As expected, adapting the regret analysis from (Riou and Honda, 2020) is non-trivial; we highlight the main challenges of this adaption in Section 5.4. For instance, one of the key challenge was to handle boundary crossing probability for the CVaR, and another difficulty comes in the analysis of the non-parametric B-CVTS due to regularity properties of the Kullback-Leibler projection. In Section 5.6, we provide empirical results on a simplified version of the case-study in agriculture introduced in the foreword of this thesis, using the well-established DSSAT crop-yield simulator (Hoogenboom et al., 2019). These simulations highlight the benefits of using strategies based on Thompson Sampling in this CVaR bandit setting against state-of-the-art baselines: We compare to U-UCB and CVaR-UCB as they showcase two fundamentally different approaches to build a UCB strategy for a non-linear utility function. The first one is closely related to UCB, the second one exploits properties of the underlying CDF, which may generalize to different risk metrics. As claimed in (Tamkin et al., 2020), our experiments confirm that CVaR-UCB generally performs better than U-UCB. However, both TS strategies outperform UCB algorithms that tend to suffer from non-optimized confidence bounds. We complete this study with additional experiments on synthetic data that also confirm the benefits of TS.

## 5.2 Non-Parametric Thompson Sampling for CVaR Bandits

We present two novel algorithms based on Thompson Sampling and targeting the lower bound of Theorem 5.3 on the CVaR-regret, for any specified value of $\alpha \in (0, 1]$. These algorithms are inspired by the first algorithms based on Thompson Sampling matching the Burnetas and Katehakis lower bound for bounded distributions in the expectation setting, recently proposed by Riou and Honda (2020).

**Notation** We introduce the notation $C_\alpha(\mathcal{X}, p)$ for the CVaR of a discrete distribution of support $\mathcal{X}$ and probability $p \in \mathcal{P}^{|\mathcal{X}|}$, where $\mathcal{P}^n$ denotes the probability simplex of size $n$. For a multinomial arm $k$ we denote its known support by $\mathcal{X}_k = (x_k^1, \dots, x_k^{M_k})$ for some $M_k \in \mathbb{N}$, and its true probability vector by $p_k$. We also define $N_k^i(t)$ as the number of times the algorithm has observed $x_k^i$ for arm $k$ before the time $t$. For general bounded distributions we denote by $\nu_k$ the distribution of arm $k$ and introduce $\mathcal{X}_{k,t}$ the set of its observed rewards before time $t$, augmented with a known upper bound $B_k$ for the support of $\nu_k$. We further introduce $\mathcal{D}_n$ as the uniform distribution on the simplex $\mathcal{P}^n$, corresponding to the Dirichlet distribution $\mathrm{Dir}((1, ..., 1))$ (with $n$ ones). We provide some properties of the Dirichlet distribution in Appendix 5.7.

**M-CVTS** For multinomial distributions M-CVTS (Multinomial-CVaR-Thompson-Sampling), described in Algorithm 5.1, directly follows the Thompson Sampling principle introduced in Chapter 1. For each arm $k$, the probability $p_k$ is assumed to be drawn from $\mathcal{D}_{M_k}$, the uniform prior on $\mathcal{P}^{M_k}$. The posterior distribution at a time $t$ is $\mathrm{Dir}(\beta_{k,t})$, with $\beta_{k,t} = (N_k^i(t)+1)_{i\in\{1,\dots,M_k\}}$. At time $t$, M-CVTS draws a sample $w_{k,t} \sim \mathrm{Dir}(\beta_{k,t})$ for each arm $k$ and computes $c_{k,t}^\alpha = C_\alpha(\mathcal{X}_k, w_{k,t})$. Then, it selects $A_t = \mathrm{argmax}_k c_{k,t}^\alpha$. For $\alpha = 1$, this algorithm coincides with the Multinomial Thompson Sampling algorithm of (Riou and Honda, 2020).

**Input:** Level $\alpha$, horizon $T$, $K$, supports $\mathcal{X}_1, \dots, \mathcal{X}_K$
**Init.:** $t = 11$, $\forall k \in \{1, ..., K\}$: $\beta_k = \underbrace{(1, \dots, 1)}_{\text{length } M_k}$
**for** $t \in \{2, \dots, T\}$ **do**
  **for** $k \in \{1, \dots, K\}$ **do**
    Draw $w_k \sim Dir(\beta_k)$ ; ▷ Draw a probability from the posterior
    Compute $c_{k,t} = C_\alpha(\mathcal{X}_k, w_k)$ ; ▷ Compute the corresponding CVaR
  **end**
  Pull arm $A_t = \mathrm{argmax}_{k\in\{1,\dots,K\}} c_{k,t}$.
  Receive reward $r_{t,A_t}$.
  Update $\beta_{A_t}(j) = \beta_{A_t}(j) + 1$, for $j$ as $r_{t,A_t} = x_k^j$ ; ▷ Update the posterior
**end**

**Algorithm 5.1:** M-CVTS

**B-CVTS** We further introduce the B-CVTS algorithm (Bounded-CVaR-Thompson-Sampling) for general bounded distributions. B-CVTS, stated as Algorithm 5.2, bears some similarity with a Thompson Sampling algorithm, although it *does not* explicitly use a prior distribution. The algorithm retains the idea of using a noisy version of $\nu_k$, obtained by a *random re-weighting* of the previous observations. Hence, at a time $t$ the index used by the algorithm for an arm $k$ is $c_{k,t} = C_\alpha(\mathcal{X}_{k,t}, w_{k,t})$, where $w_{k,t} \sim \mathcal{D}_{N_k(t)}$ is drawn uniformly at random in the simplex $\mathcal{P}^{|\mathcal{X}_{k,t}|}$.

B-CVTS then selects the arm $A_t = \text{argmax}_k c_{k,t}$. For $\alpha = 1$, this algorithm coincides with the Non Parametric Thompson Sampling of (Riou and Honda, 2020) (NPTS). NPTS can be seen as an algorithm that computes for each arm a re-weighted average of the past observations. Our extension to CVAR bandits required to interpret this operation as the computation of the *expectation* of a *random perturbation of the empirical distribution*, which can be replaced by the computation of the CVaR of this perturbed distribution. Note that this idea generalizes beyond using the CVaR, that can be replaced with any criterion.

**Input:** Level $\alpha$, horizon $T$, $K$, upper bounds $B_1, \dots, B_K$
**Init.:** $t = 1$, $\forall k \in \{1, ..., K\}$, $\mathcal{Y}_k = \{B_k\}$, $N_k = 1$ ; ▷ Init. each history with $B_k$
**for** $t \in \{2, \dots, T\}$ **do**
  **for** $k \in \{1, \dots, K\}$ **do**
    Draw $w_k \sim \mathcal{D}_{N_k}$ ; ▷ Draw a weight vector uniformly at random
    Compute $c_{k,t} = C_\alpha(\mathcal{Y}_k, w_k)$ ; ▷ CVaR of the re-weighted emp. distrib.
  **end**
  Pull arm $A_t = \text{argmax}_{k \in \{1,\dots,K\}} c_{k,t}$, receive reward $r_{t,A_t}$.
  Update $\mathcal{Y}_{A_t} = \mathcal{Y}_{A_t} \cup \{r_{t,A_t}\}$, $N_{A_t} = N_{A_t} + 1$.
**end**

**Algorithm 5.2:** B-CVTS

**Remark 5.1.** *Interestingly, B-CVTS also applies to multinomial distributions (that are bounded). The resulting strategy differs from M-CVTS due to the initialization step using the knowledge of the support in M-CVTS. As we'll show, both actually have the same theoretical guarantees when distributions are multinomial.*

## 5.3 Asymptotic optimality of the CVTS algorithms

In this section we prove, after defining this notion, that M-CVTS and B-CVTS are *asymptotically optimal* in terms of the CVaR regret defined in Equation (5.2) for the distributions they cover.

### 5.3.1 Asymptotic Optimality in CVaR bandits

Lai and Robbins (1985) first gave an asymptotic lower bound on the regret for parameteric distribution, that was later extended by Burnetas and Katehakis (1996) to more general classes of distributions. We present below an intuitive generalization of this result for CVaR bandits, first introducing the function that measures the complexity of a CVaR bandit problem.

**Definition 5.2.** *Let $\mathcal{F}$ be a family of distributions, $\alpha \in (0,1]$, and $\mathrm{KL}(\nu,\nu')$ be the KL-divergence between $\nu \in \mathcal{F}$ and $\nu' \in \mathcal{F}$. For any $\nu \in \mathcal{F}$ and $c \in \mathbb{R}$, we define*

$$\mathcal{K}_{\inf}^{\alpha,\mathcal{F}}(\nu, c) := \inf_{\nu' \in \mathcal{F}} \{\mathrm{KL}(\nu,\nu') : \mathrm{CVaR}_\alpha(\nu') \geq c\} .$$

This definition generalizes the $\mathcal{K}_{\inf}^{\mathcal{F}}$ function considered in Chapter 1 for the expectation setting. We now state our result.

**Proposition 5.3** (Regret Lower Bound in CVaR bandits)**.** *Let $\alpha \in (0,1]$. Let $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_K$ be a set of bandit models $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)$ where each $\nu_k$ belongs to the class of distribution $\mathcal{F}_k$. Let $\mathcal{A}$ be a strategy satisfying $\mathcal{R}_{\boldsymbol{\nu}}^{\alpha}(\mathcal{A}, T) = o(T^\beta)$ for any $\beta > 0$ and $\nu \in \mathcal{F}$. Then for any $\nu \in \mathcal{D}$, for any sub-optimal arm $k$, under the strategy $\mathcal{A}$ it holds that*

$$\lim_{T \to +\infty} \frac{\mathbb{E}_{\boldsymbol{\nu}}[N_k(T)]}{\log T} \geq \frac{1}{\mathcal{K}_{\inf}^{\alpha,\mathcal{F}_k}(\nu_k, c^\star)},$$

*where $c^\star = \max_{i \in [K]} \mathrm{CVaR}_\alpha(\nu_i)$.*

Using Equation (5.2), this result directly yields an asymptotic lower bound on the regret. The proof of Proposition 5.3 follows from a classical change-of-distribution argument, as that of any lower bound proof in the bandit literature. It follows from the proof of Theorem 1 in (Garivier et al., 2019) originally stated for $\alpha = 1$, and the details can be found in Appendix D.1 of (Baudry et al., 2021a).

It is well known in the bandit literature that this lower bound can lead to the desired scaling of the regret in terms of the gaps. Indeed, for $\alpha = 1$, Pinsker inequality provides that for a distribution $\nu \in \mathcal{F}$ and $\mu \in \mathbb{R}$ satisfying $\mu - \mathbb{E}_\nu[X] = \Delta > 0$, $\mathcal{K}_{\inf}^{1,\mathcal{F}}(\nu, \mu) \geq \frac{\Delta^2}{2}$. This scaling in the gap is convenient to analyze algorithm that do not target asymptotic optimality, and we say that an algorithm achieving $\mathbb{E}[N_k(T)] = \mathcal{O}\left(\log(T)/\Delta_k^2\right)$ achieves *order optimality* (which is stronger than simply having a logarithmic regret).

We can now discuss how the lower bound in Proposition 5.3 yields a weaker regret bound expressed in terms of the CVaR gaps (by Pinsker inequality). Using Lemma A.2 of Tamkin et al. (2020), we can show that for any distributions $\nu_F$ and $\nu_G$ with respective CDFs $F$ and $G$ that are supported in $[0,1]$,

$$|\mathrm{CVaR}_\alpha(F) - \mathrm{CVaR}_\alpha(G)| \leq \frac{1}{\alpha} ||F - G||_\infty .$$

It follows from Pinsker's inequality that $\mathrm{KL}(\nu_F, \nu_G) \geq \alpha^2 \left(\mathrm{CVaR}_\alpha(F) - \mathrm{CVaR}_\alpha(G)\right)^2 / 2$. Therefore, in a bandit model in which all $\nu_k$ are supported in $[0,1]$ (that is, all $\mathcal{D}_k$ are equal to $\mathcal{P}([0,1])$, the set of probability measures on $[0,1]$), it follows that

$$\mathcal{K}_{\inf}^{\alpha,\mathcal{D}}(\nu_k, c^*) \geq (\alpha \Delta_k^\alpha)^2/2.$$

Combining this inequality together with the lower bound of Theorem 5.3, we obtain that the regret of an algorithm matching the lower bound is upper bounded by $\mathcal{O}\left(\sum_{k: c_k^\alpha < c^\star} \frac{\log(T)}{\alpha^2 \Delta_k^\alpha}\right)$, which is precisely the scaling of the CVaR regret bounds obtained for the U-UCB (Cassel et al., 2018) and CVaR-UCB (Tamkin et al., 2020). Assuming the above inequalities are tight for some distributions (which may not be the case), one may qualify these algorithms as "order-optimal", as their CVaR regret makes appear the good scaling in the gaps and in $\alpha$, just like the UCB1 algorithm (Auer et al., 2002a) for $\alpha = 1$. In this chapter we go beyond order-optimality, and we strive to design algorithms that are asymptotically optimal.

### 5.3.2 Regret Guarantees for M-CVTS and B-CVTS

In the next section we prove that both M-CVTS and B-CVTS match the lower bound of Theorem 5.3 for the families of distributions they respectively tackle, assuming that they are given the support of each distribution. Hence, under these hypotheses, the two algorithms are *asymptotically optimal*. Despite the recent development in CVaR bandits literature, to our knowledge no algorithm has been proved to match this lower bound yet.

**Theorem 5.4** (Asymptotic Optimality of M-CVTS and B-CVTS). *Consider $\boldsymbol{\nu} = (\nu_1, \dots, \nu_K)$ a bandit, and denote the best CVaR by $c_\star^\alpha = \max_{k \in \{1,\dots,K\}} \mathrm{CVaR}_\alpha(\nu_k)$.*

*(I) Assume that $\forall k$, $\nu_k$ is multinomial with known support $\mathcal{X}_k \subset \mathbb{R}^{M_k}$ for some $M_k \in \mathbb{N}$. Then, for any $\varepsilon > 0$ the regret of M-CVTS satisfies*

$$\mathcal{R}_{\boldsymbol{\nu}}^\alpha(T) \leq \sum_{k: \Delta_k^\alpha > 0} \Delta_k^\alpha \frac{1}{\mathcal{K}_{\inf}^{\alpha, \mathcal{X}_k}(\nu_k, c_\star^\alpha) - \varepsilon} \log T + o_\varepsilon(\log(T)) \ .$$

*(II) Assume that $\forall k \in \{1, \dots, K\}$, $\nu_k$ belongs the family $\mathcal{C}^{B_k}$ of continuous distributions supported in $[a_k, B_k]$ for some known $B_k > 0$ and $a_k \leq B_k$. Then, for any $\varepsilon > 0$ the regret of B-CVTS on $\boldsymbol{\nu}$ satisfies*

$$\mathcal{R}_{\boldsymbol{\nu}}^\alpha(T) \leq \sum_{k: \Delta_k^\alpha > 0} \Delta_k^\alpha \frac{1}{\mathcal{K}_{\inf}^{\alpha, \mathcal{C}^{B_k}}(\nu_k, c_\star^\alpha) - \varepsilon} \log T + o_\varepsilon(\log(T)) \ .$$

The two parts of this theorem are proved in (Baudry et al., 2021a). In this manuscript we choose to only focus on the B-CVTS algorithm as the proofs of the two statements share many similarities. In the following we highlight the key steps of the analysis. First, using Equation (5.2) it is sufficient to upper bound $\mathbb{E}[N_k(T)]$ for each sub-optimal arm $k$. As in previous chapters we assume that arm 1 is optimal to ease the notation. Then our analysis follows the general outline of that of Riou and Honda (2020), but some steps require a careful adaptation to CVaR bandits. First, the proof leverages some properties of the function $\mathcal{K}_{\inf}^{\alpha}$ for the sets of distributions we consider. Secondly, it requires novel boundary crossing bounds for Dirichlet distributions that we detail in Section 5.4.

**Proof Sketch** The first step of the analysis consists in upper bounding the number of selections of a sub-optimal arm by a *post-convergence* term (Post-CV) and a *pre-convergence* term (Pre-CV). The first term controls the probability that a sub-optimal arm *over-performs* when its empirical distribution is "close" to the true distribution of the arm, while the second term considers the possibility that the CVaR of arm 1 could be under-estimated. To measure how close two distributions are we use the Levy distance, that is defined as follow for two distributions of respective cdf $F$ and $G$,

$$D_L(F,G) = \inf\left\{\varepsilon > 0 : \forall x \in [0,B], F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon\right\} .$$

The following result summarizes this decomposition in two terms.

**Proposition 5.5** (Decomposition in Post-CV and Pre-CV terms)**.** *Under B-CVTS, the number of pulls of each sub-optimal arms satisfy*

$$\mathbb{E}[N_k(T)] \leq n_0(T) + \underbrace{\sum_{t=1}^{T} \mathbb{E}_{\nu_k,\pi}\left[\mathbb{1}\left(\mathcal{G}_k(t)\right) \times \mathbb{P}_{w\sim\mathcal{D}_{N_k(t)}}\left(C_\alpha(\mathcal{Y}_k(t),w) \geq c_1^\alpha - \varepsilon_1\right)\right]}_{\textit{post-convergence term (Post-CV)}}$$
$$+ \underbrace{\sum_{n=1}^{T} \mathbb{E}_{\mathcal{Y}_n\sim\nu_1^n}\left[\frac{\mathbb{P}_{w\sim\mathcal{D}_n}\left(C_\alpha(\mathcal{Y}_{1,n},w) \leq c_1^\alpha - \varepsilon_1\right)}{1 - \mathbb{P}_{w\sim\mathcal{D}_n}\left(C_\alpha(\mathcal{Y}_{1,n},w) \leq c_1^\alpha - \varepsilon_1\right)}\right]}_{\textit{pre-convergence term (Pre-CV)}} + \mathcal{O}(1) ,$$

*where $\mathcal{G}_k(t) = \{N_k(t-1) \geq n_0(T), D_L(F_k(t), F_k) \leq \varepsilon_2\}$ is the event corresponding to arm $k$ being sampled "enough" and "close" to its true distribution, $\mathcal{Y}_{1,n}$ is the set of $n$ first data collected from arm 1 and $\mathcal{Y}_k(t)$ is the set of data collected from arm $k$ up to time $t$. The empirical distribution $F_k(t)$ includes the upper bound $B$ added to the set of observations.*

The proof of proposition 5.5 uses standard arguments in bandits, and is a direct adaptation of the proof of NPTS from (Riou and Honda, 2020). It decomposes the number of pulls of each sub-optimal arm between a constant $n_0(T)$, a *post-convergence*, and a *pre-convergence* terms. Interestingly, the former depends only on arm $k$, and corresponds to this arm being sampled while having a sufficient number of samples and an empirical cdf that accurately reflect its true cdf. On the other hand, the later is the cost of under-estimating the best arm in the regret. Finally, the additional constant term corresponds to $\sum_{t=1}^{T} \mathbb{P}(\mathcal{G}_k(t)^c, A_t = k)$. Riou and Honda (2020) provide a method to derive a constant upper bound, that can be refined by using DKW inequality instead (Massart, 1990). We omit the development of this part, that can be found in the paper.

Further upper bounding (Post-CV) and (Pre-CV) requires to provide upper and lower bound on *boundary crossing probabilities* for Dirichlet random variables, that we define as follows.

**Definition 5.6** (Boundary Crossing Probabilities (BCP))**.** *A BCP is a probability of the form*

$$\mathbb{P}_{w\sim \mathrm{Dir}(\beta)} \left( C_\alpha(\mathcal{Y}, w) \geq c \right),$$

*for some known support $\mathcal{Y} = (y_1, \dots, y_n)$, parameter $\beta \in \mathbb{R}_+^n$ of the Dirichlet distribution, and some real value $c$ that are defined in context.*

Finally, proving the asymptotic optimality of B-CVTS consists in showing that both the pre-convergence and post-convergence terms can be further upper bounded by constants, while choosing

$$n_0(T) = \frac{1}{\mathcal{K}_{\inf}^{\alpha, \mathcal{C}^{B_k}}(\nu_k, c_\star^\alpha) - \varepsilon} \log(T) \,,$$

for any constant $\varepsilon > 0$. In the next section we present the technical tools that allow to draw these conclusions.

## 5.4 Further upper bounds on Pre-CV and Post-CV

In the analysis of NPTS, Riou and Honda (2020) provide several results to upper and lower bound the BCP for $\alpha = 1$. However, replacing the linear expectation by the CVaR, that is non-linear, causes several technical challenges that makes the adaptation not direct. In this section we provide the technical results that allow to complete the proof of Theorem 5.4 from Proposition 5.5: the dual form of the $\mathcal{K}_{\inf}^{\alpha, \mathcal{D}}$ function, and upper and lower bounds on the BCP.

### 5.4.1 Technical tools

**Dual of $\mathcal{K}_{\inf}^{\alpha,\mathcal{D}}$** Our first result is the derivation of the dual formulation of the $\mathcal{K}_{\inf}^{\alpha,\mathcal{D}}$ functional, providing an extension of the result of Honda and Takemura (2010) when $\alpha = 1$.

**Lemma 5.7.** *If a discrete distribution $F$ supported on $\mathcal{Y}$ satisfies $\mathbb{E}_F\left[\frac{(y-c)\alpha}{(y-X)^+}\right] < 1$, then for any $c > CVaR_\alpha(F)$ it holds that*

$$\mathcal{K}_{\inf}^{\alpha}(F,c) = \inf_{y\in\mathcal{Y}} \max_{\lambda\in\left[0,\frac{1}{\alpha(y-c)}\right)} g(y,\lambda,X) \ ,$$

*with $g(y,\lambda,X) = \mathbb{E}_F\left[\log(1-\lambda((y-c)\alpha) - (y-X)^+)\right]$.*

*If $\mathbb{E}_F\left[\frac{(y-c)\alpha}{(y-X)^+}\right] \geq 1$, then for any $c > CVaR_\alpha(F)$*

$$\mathcal{K}_{\inf}^{\alpha}(F,c) = \inf_{y\in\mathcal{Y}} \mathbb{E}_F\left(\frac{(y-X)^+}{(y-c)\alpha}\right) \ .$$

*Proof.* We let $\mathcal{P}^M$ denote he simplex of dimension $M$. We rewrite the optimization problem, defined for any $p \in \mathcal{P}^M$, $\alpha \in (0,1]$ and $c \in [0,1]$ as

$$\mathcal{K}_{\inf}^{\alpha,\mathcal{D}}(p,c) = \inf_{q\in\mathcal{P}^M} \{\mathrm{KL}(p,q) : C_\alpha(\mathcal{Y},q) \geq c\} \ .$$

First of all, we recall that $C_\alpha(\mathcal{Y},q) = \sup_{x\in\mathcal{D}}\left\{x - \frac{1}{\alpha}\mathbb{E}_{X\sim q}\left((x-X)^+\right)\right\}$. We then introduce the set

$$\begin{aligned}\mathcal{P}_{y,\alpha,c}^M &= \left\{q \in \mathcal{P}^M : y - \frac{1}{\alpha}\mathbb{E}_{X\sim q}((y-X)^+) \geq c\right\} \\ &= \left\{q \in \mathcal{P}^M : \mathbb{E}_{X\sim q}((y-X)^+) \leq (y-c)\alpha\right\} \ .\end{aligned}$$

Thanks to this definition we can rewrite the problem as

$$\mathcal{K}_{\inf}^{\alpha,\mathcal{D}}(p,c) = \min_{y\in\mathcal{D}}\left\{\inf_{q\in\mathcal{P}^M}\left\{\mathrm{KL}(p,q) : y - \frac{1}{\alpha}\mathbb{E}_{X\sim q}((y-X)^+) \geq c\right\}\right\} \ ,$$

where we used that $\{q : C_\alpha(\mathcal{Y},q) \geq c\} = \cup_{y\in\mathcal{D}}\left\{q \in \mathcal{P}^M : y - \frac{1}{\alpha}\mathbb{E}_{X\sim q}\left((X-y)^+\right) \geq c\right\}$.

Now, we can first solve the problem $\inf_{q\in\mathcal{P}^M_{y,\alpha,c}} \mathrm{KL}(p,q)$ for a fixed value of $y$, satisfying $y > c$ (else the feasible set is empty). We write the Lagrangian of this problem:

$$H(q,\lambda_1,\lambda_2) = \sum_{i=1}^{M} p_i \log\left(\frac{p_i}{q_i}\right) + \lambda_1\left(\sum_{i=1}^{M} q_i - 1\right) + \lambda_2\left(\sum_{i=1}^{M} q_i(y-y_i)^+ - \alpha(y-c)\right) ,$$

and want to solve $\max_{\lambda_1>0,\lambda_2>0} \min_q H(q,\lambda_1,\lambda_2)$. To this end, we write

$$\frac{\partial H}{\partial q_i} = -\frac{p_i}{q_i} + \lambda_1 + (y-y_i)^+ .$$

Setting the derivative to $0$ yields

$$q_i = \frac{p_i}{\lambda_1 + \lambda_2(y-y_i)^+} .$$

We can check that the inequality constraint is achieved. Moreover, exploiting the two constraints leads to $\lambda_1 + \lambda_2\alpha(y-c) = 1$. This finally gives

$$q_i = \frac{p_i}{1 - \lambda_2((y-c)\alpha - (y-y_i)^+)} .$$

Note that this solution is only valid if $\lambda_2 \leq \frac{1}{\alpha(y-c)}$. We have two possibilities: 1) the maximum is achieved in $[0, \frac{1}{\alpha(y-c)})$, in this case we have

$$\begin{aligned}\mathcal{K}^{\alpha,\mathcal{D}}_{\inf}(p,c) &= \inf_{y\in\mathcal{D}} \max_{\lambda\in\left[0,\frac{1}{\alpha(y-c)}\right)} \sum_{i=1}^{M} p_i \log\Big(1 - \lambda_2((y-c)\alpha - (y-y_i)^+)\Big) \\ &= \inf_{y\in\mathcal{D}} \max_{\lambda\in\left[0,\frac{1}{\alpha(y-c)}\right)} \mathbb{E}_{X\sim p}[\log\Big(1 - \lambda_2((y-c)\alpha - (y-X)^+)\Big)] .\end{aligned}$$

The other possibility is that the function is still increasing in $\lambda_2 = \frac{1}{\alpha(y-c)}$. For this case, we check the sign of $\frac{\partial \mathbb{E}_{X\sim F}[\log\left(1+\lambda_2((y-c)\alpha-(y-X)^+)\right)]}{\partial\lambda}$ at point $\lambda = \frac{1}{\alpha(y-c)}$, that is the same as the one of $(y-c)\alpha\left(1 - \mathbb{E}_F\left(\frac{(y-c)\alpha}{(y-X)^+}\right)\right)$. We see that the function can only be increasing if $\mathbb{E}_F\left(\frac{(y-c)\alpha}{(y-X)^+}\right) < 1$, and the solution is then $q_i = \frac{p_i(y-c)\alpha}{y-y_i}$, which provides $\mathcal{K}^{\alpha,\mathcal{D}}_{\inf}(p,c) = \inf_y \mathbb{E}_F\left(\frac{(y-X)^+}{(y-c)\alpha}\right)$. This concludes the proof. □

This result matches the one of Honda and Takemura (2010) for $\alpha = 1$ and $y = 1$, and is similar to the one obtained by Agrawal et al. (2021b)[Theorem 6] for a more complex set of distributions (which is hence less explicit). Furthermore, Agrawal et al. (2021b)[Lemma 4] prove the continuity of $\mathcal{K}^{\alpha,\mathcal{Y}}_{\inf}$ under this condition, which is required in several part of our proofs.

**Upper Bound on the BCP** Building on the dual form of the $\mathcal{K}_{\inf}^{\alpha,\mathcal{D}}$ for discrete distributions we just introduced, we can derive an upper bound of the BCP using similar techniques as Riou and Honda (2020). We consider a known support $\mathcal{Y} = (y_1, \dots, y_n)$ and the Dirichlet distribution $\mathcal{D}_n$ defined in Section 5.2. We further denote by $F_{\mathcal{Y}}$ the uniform distribution on $\mathcal{Y}$ (or empirical distribution of $\mathcal{Y}$), and $C_\alpha(\mathcal{Y})$ its CVaR.

**Lemma 5.8** (Upper Bound on the BCP). *Let $\mathcal{Y} = (y_1, \dots, y_n)$ for some known $B > 0$ and $n \in \mathbb{N}$, for any $c > C_\alpha(\mathcal{Y})$ it holds that*

$$\mathbb{P}_{w\sim\mathcal{D}_n}(C_\alpha(\mathcal{Y}, w) \geq c) \leq n \exp^{-n\mathcal{K}_{\inf}^{\alpha,\bar{\mathcal{Y}}}(F_{\mathcal{Y}},c)} ,$$

*where $\bar{\mathcal{Y}} = \max\{y_1, \dots, y_n\}$. Furthermore, the multiplicative factor $n$ can be removed if $\alpha = 1$.*

*Proof.* We first use the formulation of the CVaR of Equation (5.3) for discrete distributions, and a union bound to obtain

$$\begin{aligned}
\mathbb{P}_w\left(C_\alpha(\mathcal{Y}, w) \geq c\right) &= \mathbb{P}_w\left(\sup_{y\in\mathcal{Y}}\left\{y - \frac{1}{\alpha}\sum_{i=1}^n w_i(y-y_i)^+\right\} \geq c\right) \\
&\leq \sum_{y\in\mathcal{Y}} \mathbb{P}_w\left(y - \frac{1}{\alpha}\sum_{i=1}^n w_i(y-y_i)^+ \geq c\right) \\
&\leq n \max_{y\in\mathcal{Y}} \mathbb{P}_w\left(y - \frac{1}{\alpha}\sum_{i=1}^n w_i(y-y_i)^+ \geq c\right)
\end{aligned}$$

We see that the multiplicative $n$ comes from the union bound on the possible values of $y$ for the quantile of order $\alpha$. When $\alpha = 1$ it is always equal to the maximum value in $\mathcal{Y}$, hence the union bound is not necessary.

We then handle $\mathbb{P}\left(\alpha(y-c) - \sum_{i=1}^n w_i(y-y_i)^+ \geq 0\right)$ for a fixed value of $y$. We follow the path of Riou and Honda (2020), using that a Dirichlet random variable $w = (w_1, \dots, w_n)$ can be written in terms of $n$ independent random variables $R_1, \dots, R_n$ following an exponential distribution, as $w_i = \frac{R_i}{\sum_{j=1}^n R_j}$. Using this property and multiplying by $\sum_{j=1}^n R_j$ we obtain

$$\begin{aligned}
P_y := \mathbb{P}\left(\alpha(y-c) - \sum_{i=1}^n w_i(y-y_i)^+ \geq 0\right) \leq& \mathbb{P}\left(\sum_{i=1}^n R_i\left(\alpha(y-c) - (y-y_i)^+\right) \geq 0\right) \\
\leq& \mathbb{E}\left[\exp\left(t\sum_{i=1}^n R_i\left(\alpha(y-c) - (y-y_i)^+\right)\right)\right] ,
\end{aligned}$$

where we used Markov's inequality for some $t \in \left[0, \frac{1}{(y-c)\alpha}\right)$. We then conclude by deriving the MGF of the exponential variables,

$$\begin{aligned} P_y &\leq \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(R_i t\left(\alpha(y-c) - (y-y_i)^+\right)\right)\right] \\ &\leq \exp\left(-\sum_{i=1}^{n} \log\left(1 - t\left(\alpha(y-c) - (y-y_i)^+\right)\right)\right) \\ &\leq \sup_{y \in [c,B]} \left\{\exp\left(-n\mathbb{E}_{F_{\mathcal{Y}}}\left[\log\left(1 - t\left(\alpha(y-c) - (y-Y)^+\right)\right)\right]\right)\right\} . \end{aligned}$$

We then put the sup on $y$ inside the exponent, and recognize the dual form of $\mathcal{K}_{\inf}^{\alpha,\mathcal{Y}}$, which concludes the proof. □

**Lower Bounds on the BCP** We now establish two lower bounds on the BCP. The first one concerns the case where we only want the CVaR of the noisy distribution to exceed the CVaR of the empirical distribution generated by the dataset $\mathcal{Y}$.

**Lemma 5.9.** *Assume that $\mathcal{Y} = (y_1, \dots, y_n)$ and $y_1 < \cdots < y_n$, then $y_{\lceil n\alpha \rceil}$ is the empirical $\alpha$ quantile of the set and $y_1$ its minimum, and it holds that*

$$\mathbb{P}_{w \sim \mathcal{D}_n}\left(C_\alpha(\mathcal{Y}, w) \geq C_\alpha(\mathcal{Y})\right) \geq \frac{1}{25n^3}(y_{\lceil n\alpha \rceil} - y_1) .$$

*Proof.* We assume that $\mathcal{Y}$ is known and ordered, i.e $y_1 \leq y_2 \leq \cdots \leq y_n$. We then write

$$A = \mathbb{P}_{w \sim \mathcal{D}_n}\left(C_\alpha(\mathcal{Y}, w) \geq C_\alpha(\mathcal{Y})\right) .$$

Thanks to the definition of the CVaR provided by Equation (5.1) it holds that

$$A = \mathbb{P}_w\left(\sup_{y \in \mathcal{Y}}\{y - \frac{1}{\alpha}\sum_{i=1}^{n} w_i(y-y_i)^+\} \geq \sup_{z \in \mathcal{Y}}\{z - \frac{1}{\alpha n}\sum_{i=1}^{n}(z-y_i)^+\}\right) .$$

First, if we know $y_1, \dots, y_n$ then the second term is deterministic and the sup is actually achieved in $y_{\lceil n\alpha \rceil}$. Secondly, the inequality is true if at least one term in the left element satisfies it, so we can write

$$
\begin{aligned}
A =& \mathbb{P}\left(\sup_{z\in\mathcal{Y}}\left\{z - \frac{1}{\alpha}\sum_{i=1}^{n} w_i(z-y_i)^+\right\} \geq y_{\lceil n\alpha\rceil} - \frac{1}{\alpha n}\sum_{i=1}^{n}(y_{\lceil n\alpha\rceil}-y_i)^+\right)\\
\geq& \mathbb{P}\left(y_{\lceil n\alpha\rceil} - \frac{1}{\alpha}\sum_{i=1}^{n} w_i(y_{\lceil n\alpha\rceil}-y_i)^+ \geq y_{\lceil n\alpha\rceil} - \frac{1}{\alpha n}\sum_{i=1}^{n}(y_{\lceil n\alpha\rceil}-y_i)^+\right)\\
=& \mathbb{P}\left(\sum_{i=1}^{n} w_i(y_{\lceil n\alpha\rceil}-y_i)^+ \leq \frac{1}{n}\sum_{i=1}^{n}(y_{\lceil n\alpha\rceil}-y_i)^+\right)\\
=& \mathbb{P}\left(\sum_{i=1}^{n} w_i\frac{B-(y_{\lceil n\alpha\rceil}-y_i)^+}{B} \geq \frac{1}{n}\sum_{i=1}^{n}\frac{B-(y_{\lceil n\alpha\rceil}-y_i)^+}{B}\right) .
\end{aligned}
$$

As the variable $\frac{B-(y_{\lceil n\alpha\rceil}-y_i)^+}{B}$ belongs to $[0,1]$ we can apply the lemma 17 of Riou & Honda and get

$$
A \geq \frac{1}{25n^2B}\left(B - \frac{1}{n}\sum_{i=1}^{n}(B-(y_{\lceil n\alpha\rceil}-y_i)^+)\right) = \frac{1}{25n^3B}\sum_{i=1}^{n}(y_{\lceil n\alpha\rceil}-y_i)^+ .
$$

We conclude by omitting all the terms except $(y_{\lceil n\alpha\rceil} - y_1)$ in the sum. □

We finally study the case when the size of the support is $|\mathcal{Y}| = M$, for some known $M \in \mathbb{N}$ and when the considered distributions are the *frequencies* of each observation in $\mathcal{Y}$ out of $n \in \mathbb{N}$ many observations, which we represent by the set $\mathcal{Q}_n^M = \left\{(\beta, p) \in \mathbb{N}^{*n} \times \mathcal{P}^M : p = \frac{\beta}{n}\right\}$. Our last result is a lower bound on the BCP for distributions in $\mathcal{Q}_n^M$.

**Lemma 5.10** (Lower Bound on the BCP for multinomial distributions). *For any $(M,n) \in N^2$ and $(\beta,p) \in \mathcal{Q}_n^M$, let $p^\star \in \mathcal{P}^M$ be any vector satisfying $C_\alpha(\mathcal{X}, p^\star) \geq c$. It holds that*

$$
\begin{aligned}
\mathbb{P}_{w\sim\mathrm{Dir}(\beta)}\left(C_\alpha(\mathcal{X},w) \geq c\right) &\geq \frac{n!}{\prod_{i=1}^{M}\beta_i!}\frac{\beta_M}{np_M^\star}\prod_{j=1}^{M}(p_j^\star)^{\beta_j}\\
&\geq \frac{1}{n}\mathbb{P}_{\mathrm{Mult}(n,p)}(\beta) \times e^{-n\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}(p,c)} \geq C_M\frac{\exp\left(-n\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}(p,c)\right)}{n^{\frac{M+1}{2}}} ,
\end{aligned}
$$

*where $C_M = \sqrt{2\pi}\left(\frac{M}{2.13}\right)^{\frac{M}{2}}$, and $\mathbb{P}_{\mathrm{Mult}(n,q)}(\beta)$ denotes the probability that a vector drawn from a multinomial distribution with $n$ trials and probability $q$ is equal to $\beta$.*

*Proof.* We follow the sketch of the proof of Lemma 14 of Riou and Honda (2020) using Equation (5.3). We start by stating that there exists some $p^*$ such that $\mathcal{K}_{\inf}^{\alpha,\mathcal{Y}}\left(\frac{\beta}{n}, c\right) = \mathrm{KL}\left(\frac{\beta}{n}, p^*\right)$. The existence of $p^*$ is ensured by the fact that the function $\mathcal{K}_{\inf}^{\alpha,\mathcal{Y}}$ is the solution of the minimization

of a continuous function on a compact set. We consider the set

$$\mathcal{S}_2 = \{w \in \mathcal{P}^{M+1} : w_i \in [0, p_i^*], \forall j \leq M-1, w_M \geq p_M^*\}.$$

Let us remark that $\forall p \in \mathcal{S}_2, C_\alpha(\mathcal{Y}, p) \geq C_\alpha(\mathcal{Y}, p^*) \geq c$. Indeed, if we transfer some of the mass from some items of the support to largest items we can only increase the CVaR. It then holds that

$$\begin{aligned}
\mathbb{P}_{w\sim Dir(\beta)}\left(C_\alpha(\mathcal{Y}, w) \geq c\right) &\geq \mathbb{P}_{w\sim Dir(\beta)}\left(w \in \mathcal{S}_2\right) \\
&= \frac{\Gamma(n)}{\prod_{i=1}^M \Gamma(\beta_i)} \int_{x\in S_2} \prod_{i=1}^M y_i^{\beta_i - 1} dx \\
&\geq \frac{\Gamma(n)}{\prod_{i=1}^M \Gamma(\beta_i)} (p_M^*)^{\beta_M - 1} \prod_{j=1}^{M-1} \int_{y_j=0}^{p_j^*} y_j^{\beta_j - 1} dy_j \\
&= \frac{\Gamma(n)}{\prod_{i=1}^M \Gamma(\beta_i)} (p_M^*)^{\beta_M - 1} \prod_{j=1}^{M-1} \frac{(p_j^*)^{\beta_j}}{\beta_j} \\
&= \frac{\Gamma(n)}{\prod_{i=1}^M \Gamma(\beta_i)} \frac{\beta_M}{p_M^\star} \prod_{j=1}^{M} \frac{(p_j^*)^{\beta_j}}{\beta_j} \ ,
\end{aligned}$$

which proves the first inequality. We then exhibit the KL-divergence between two multinomial distributions using

$$\begin{aligned}
\mathbb{P}_{w\sim Dir(\beta)}\left(w \in \mathcal{S}_2\right) &\geq \frac{\Gamma(n)}{\prod_{i=1}^M \Gamma(\beta_i)} \prod_{j=1}^{M} \frac{(p_j^*)^{\beta_j}}{\beta_j} \\
&= \frac{\Gamma(n)}{\prod_{i=1}^M \Gamma(\beta_i)} \prod_{j=1}^{M} \left(\frac{p_j^*}{\beta_j}\right)^{\beta_j} \times \prod_{j=1}^{M} \beta_j^{\beta_j - 1} \\
&= \frac{\Gamma(n)}{\prod_{i=1}^M \Gamma(\beta_i + 1)} \prod_{j=1}^{M} \left(\frac{p_j^*}{\beta_j/n}\right)^{\beta_j} \times \prod_{j=1}^{M} \left(\frac{\beta_j}{n}\right)^{\beta_j} \\
&= \frac{\Gamma(n)}{\prod_{i=1}^M \Gamma(\beta_i + 1)} \prod_{j=1}^{M} \left(\frac{\beta_j}{n}\right)^{\beta_j} \exp\left(-n\mathrm{KL}\left(\frac{\beta}{n}, p^\star\right)\right) \ .
\end{aligned}$$

This corresponds to the second inequality in the lemma, as we can remark that the terms before the exponential correspond to the desired multinomial distributions, up to a factor $1/n$. We finally provide a lower bound of this quantity using Stirling formula (similarly to the proof scheme of Riou and Honda (2020)),

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leq n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n (1 + C(n)) \ ,$$

with $C(n) = \frac{1}{12n} + \frac{1}{288n^2}$. We then obtain that

$$\begin{aligned}
\frac{\Gamma(n)}{\prod_{i=1}^{M}\Gamma(\beta_i+1)}\prod_{j=1}^{M}\left(\frac{\beta_j}{n}\right)^{\beta_j} &= \frac{1}{n}\frac{n!}{n^n}\prod_{j=1}^{M}\frac{\beta_j^{\beta_j}}{\beta_j!} \\
&\geq \sqrt{\frac{2\pi}{n}}e^{-n} \times \prod_{j=1}^{M}\frac{1}{1+C(\beta_j)}\frac{e^{\beta_j}}{\sqrt{2\pi\beta_j}} \\
&= \sqrt{\frac{2\pi}{n}}\prod_{j=1}^{M}\frac{1}{(1+C(\beta_j))\sqrt{2\pi\beta_j}} \\
&\geq \sqrt{\frac{2\pi}{n}}e^{\frac{-M}{12}}\prod_{j=1}^{M}\frac{1}{(1+C(\beta_j))\sqrt{2\pi\beta_j}} \\
&\geq \sqrt{\frac{2\pi}{n}}e^{\frac{-M}{12}}\frac{1}{\sqrt{2\pi}^M} \times \left(\frac{M}{n}\right)^{\frac{M}{2}} \\
&\geq \frac{e^{-\frac{M}{12}}M^{\frac{M}{2}}}{\sqrt{2\pi}^{M-1}}n^{-\frac{M+1}{2}} \\
&\geq \sqrt{2\pi}\left(\frac{M}{2.13}\right)^{\frac{M}{2}} \times n^{-\frac{M+1}{2}} \\
&= C_M n^{-\frac{M+1}{2}} ,
\end{aligned}$$

where we used that $C(n)$ is maximum when $n = 1$ and that $1/12 \geq \log(25/288)$, and on the other hand that $\prod_{j=1}^{M}\beta_j$ is minimized when all $\beta_j$ are equal to $n/M$ (if we allow continuous values). □

The results presented in this section contains most of the difficulty induced by the replacement of the expectation by the CVaR in the proofs. Extending these results to other criterion is an interesting future work and may help generalize the Non Parametric Thompson Sampling algorithms to broader settings. In the next section we use these results to complete the proof of the asymptotic optimality of B-CVTS for bounded distributions.

### 5.4.2 Application to derive upper bounds on Post-CV and Pre-CV

We now use the results from previous section in order to prove that both Post-CV and Pre-CV are upper bounded by constants.

#### Upper bounding (Post-CV)

We start with this term because the result is direct using Lemma 5.8 and the continuity of $\mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}$. The continuity ensures that for any $\varepsilon_0 > 0$ there exists $\varepsilon_1 > 0$ such that if $D_L(F_k(t), F_k) \leq \varepsilon \Rightarrow$

$\mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}(F_k(t), c_1 - \varepsilon_2) \geq \mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}(F_k, c_1) - \varepsilon_0$. For that reason, for any $\varepsilon$, with well chosen $(\varepsilon_1, \varepsilon_2)$ we can write

$$\begin{aligned}\text{(Post-CV)} \leq & \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(\mathcal{G}_k(t)) N_k(t) \exp\left(-N_k(t)\mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}(F_k(t), c_1)\right)\right] \\ \leq & T n_0(T) \exp\left(-n_0(T)\left(\mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}(F_k, c_1) - \varepsilon_0\right)\right) ,\end{aligned}$$

if $n_0(T) \geq \frac{1}{\mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}(F_k,c_1)-\varepsilon_0}$. We now choose $n_0(T) = \frac{\log(T)+\log(\log T)}{\mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}(F_k,c_1)-\varepsilon_0}$ to obtain that for $T \geq 3$,

$$\text{(Post-CV)} \leq \frac{1}{\mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}(F_k, c_1) - \varepsilon_0}\left(1 + \frac{\log(T)}{\log(\log(T))}\right) \leq \frac{2}{\mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}(F_k, c_1) - \varepsilon_0} .$$

**Upper bounding (Pre-CV)**

We recall that the (Pre-CV) term is expressed as

$$A \coloneqq \sum_{n=1}^{T} \mathbb{E}_{\mathcal{Y}_n \sim \nu_1^n}\left[\frac{\mathbb{P}_{w\sim\mathcal{D}_n}\left(C_\alpha(\mathcal{Y}_n, w) \leq c_1^\alpha - \varepsilon_1\right)}{1 - \mathbb{P}_{w\sim\mathcal{D}_n}\left(C_\alpha(\mathcal{Y}_n, w) \leq c_1^\alpha - \varepsilon_1\right)}\right] .$$

Inspired by (Riou and Honda, 2020) we split this expectation into different regions depending of the value of the CVaR of the empirical distribution (including the term $y_0 = B$ added at the beginning of the history of observations).

We split the upper bound on $A$ into three terms

$$A \leq A_1 + A_2 + A_3 ,$$

where $A_1$ corresponds to the region $\{C_\alpha(\mathcal{Y}_n) \geq c_1^\alpha - \varepsilon_1/2\}$, $A_2$ to the region $\{c_1^\alpha - \varepsilon_1 \leq C_\alpha(\mathcal{Y}_n) \leq c_1^\alpha - \varepsilon_1/2\}$ and $A_3$ to $\{C_\alpha(\mathcal{Y}_n) \leq c_1^\alpha - \varepsilon_1\}$.

We now upper bound each of these three terms, for any value of $\varepsilon_1$.

**Upper bounding** $A_1$ We do not detail this part because this region corresponds to the most favorable case, as the CVaR of arm 1 is reasonably estimated. Hence, the BCP admits an exponential upper bound that can be derived with similar techniques as for Lemma 5.8, and so $A_1$ is upper bounded by a constant.

**Upper bound on $A_2$** In order to control the term $A_2$ we can upper bound the numerator by 1 and use Lemma 5.9 to obtain

$$A_2 \leq \sum_{n=1}^{T} \mathbb{E}_{y_1,\dots,y_n}\left[\mathbb{1}\left(c_1^\alpha - \varepsilon_1 \leq C_\alpha(\mathcal{Y}_n) \leq c_1^\alpha - \varepsilon_1/2\right) \frac{25n^3 \mathbb{1}(Y_1 < c_1^\alpha - \varepsilon_1)}{Y_{\lceil n\alpha \rceil} - Y_1}\right] .$$

Here, we have introduced $Y_1, \dots, Y_n$ to denote the ordered list of $(y_1, \dots, y_n)$ (i.e $Y_1 \leq Y_2 \leq \cdots \leq Y_n$). We also added the indicator $\mathbb{1}(Y_1 \leq c_1^\alpha - \varepsilon_1)$ because if $Y_1 \geq c_1^\alpha - \varepsilon_1$ then $C_\alpha(\mathcal{Y}_n, w) \geq c_1^\alpha - \varepsilon_1$ for any $w \in \mathcal{P}^n$. Furthermore, under the events we consider it also holds that

$$Y_{\lceil n\alpha \rceil} \geq C_\alpha(\mathcal{Y}_n) \geq c_1^\alpha - \varepsilon_1 \ .$$

Note that it is impossible to conclude at this step in general because the variable $Y_{\lceil n\alpha \rceil} - Y_1$ may be arbitrarily small in case all the $n$ observations are very concentrated. However, if $n$ is large and the distribution is *continuous* this event can only happen with a very low probability. This is a place in the proof where continuity is crucial. To do so, we upper bound the rest of the terms with a peeling argument on the values of $Y_1$. This is done using the closed-form formulas for the distribution of the minimum of $n$ random variable that are independent and identically distributed. Indeed, if $f_1$ denotes the density of arms 1, and we write the cdf and pdf of the minimum of $n$ independent observations of $\nu_1$ respectively $L_n$ and $l_n$, then it holds that $\forall x \in [0, B]$

$$L_n(x) = 1 - (1 - F_1(x))^n \ .$$

Now, since $\nu_1$ is continuous it follows that in each point the density is $l_n(x) = nf_1(x)(1 - F_1(x))^{n-1}$. The next step consists in defining a strictly decreasing sequence $(a_k)_{k\geq 0}$, and to look at the intervals $S_k = [c_1^\alpha - a_k - \varepsilon_1, c_1^\alpha - a_{k+1} - \varepsilon_1]$. On each of these intervals we obtain by construction that $Y_{\lceil n\alpha \rceil} \geq c_1^\alpha - \varepsilon_1 \geq Y_1 + a_{k+1}$, and thus

$$\mathbb{E}_{\mathcal{Y}_n}\left[\frac{25n^3}{Y_{\lceil n\alpha \rceil} - Y_1} \mathbb{1}\left(Y_1 \in S_k\right)\right] \leq \frac{25n^3}{a_{k+1}} \times \mathbb{P}\left(Y_1 \in S_k\right) \ .$$

Using the properties of the density $l_n$ it holds that

$$\begin{aligned}
\mathbb{P}\left(y_1 \in S_k\right) &= \int_{c_1^\alpha - \varepsilon_1 - a_k}^{c_1^\alpha - \varepsilon_1 - a_{k+1}} nf_1(x)(1 - F_1(x))^{n-1} dx \\
&\leq \sup_{x\in[0,B]} f_1(x) \int_{c_1^\alpha - \varepsilon_1 - a_k}^{c_1^\alpha - \varepsilon_1 - a_{k+1}} n(1 - F_1(x))^{n-1} dx \\
&\leq \sup_{x\in[0,B]} f_1(x)(a_k - a_{k+1}) n(1 - F_1(c_1^\alpha - \varepsilon_1 - a_k))^{n-1} \ .
\end{aligned}$$

With these results we can now upper bound $A_2$ by writing

$$A_2 \leq \underbrace{\sum_{n=1}^{T} \mathbb{E}_{y_1,\dots,y_n}\left[\mathbb{1}\left(c_1^\alpha - \varepsilon_1 \leq C_\alpha(\mathcal{Y}_n) \leq c_1^\alpha - \varepsilon_1/2\right) \frac{25n^3}{Y_{\lceil n\alpha\rceil} - Y_1} \mathbb{1}(y_1 \leq c_1^\alpha - \varepsilon_1 - a_0)\right]}_{A_{21}}$$
$$+ \underbrace{\sum_{n=1}^{T} \mathbb{E}_{y_1,\dots,y_n}\left[\mathbb{1}\left(c_1^\alpha - \varepsilon_1 \leq C_\alpha(\mathcal{Y}_n) \leq c_1^\alpha - \varepsilon_1/2\right) \frac{25n^3}{Y_{\lceil n\alpha\rceil} - Y_1} \mathbb{1}(y_1 \geq c_1^\alpha - \varepsilon_1 - a_0)\right]}_{A_{22}} .$$

The left-hand side term can be handled thanks to Brown's inequality (Brown, 2007), which is the equivalent of Hoeffding's inequality for CVaR. Using that $Y_{\lceil n\alpha\rceil} - Y_1 \geq a_0$ on the considered interval, we obtain

$$A_{21} \leq \sum_{n=1}^{T} \frac{25n^3}{a_0} e^{-2n\left(\frac{\alpha(a_0+\varepsilon_1)}{B_k}\right)^2} = \mathcal{O}(1) .$$

Regarding the second term $A_{22}$ we have

$$A_{22} \leq \sum_{n=1}^{T} \sup_{x\in[0,B]} n f_1(x) \times \sum_{k=0}^{+\infty} \frac{a_k - a_{k+1}}{a_{k+1}} (1 - F_1(c_1^\alpha - \varepsilon_1 - a_k))^{n-1} .$$

We first use that the cdf is increasing, which enables to upper bound $(1-F_1(c_1^\alpha-\varepsilon_1-a_k))^{n-1}$ by the quantity $(1 - F_1(c_1^\alpha - \varepsilon_1 - a_0))^{n-1}$. It remains to choose the sequence $(a_k)$ in order to make the sum $\sum_{k=0}^{+\infty} \frac{a_k - a_{k+1}}{a_{k+1}}$ converge. We define recursively the sequence as $a_{k+1} = \frac{2^k}{2^k+1} a_k$, starting from $a_0 = \frac{c_1^\alpha - \varepsilon_1}{2}$. This way, $\sum_{k=0}^{+\infty} \frac{a_k - a_{k+1}}{a_{k+1}} = \sum_{k=0}^{+\infty} \frac{1}{2^k} = 2$. This shows that

$$A_{22} \leq 50 \sum_{n=1}^{T} n^4 \sup_{x\in[0,B]} f_1(x) \exp\left(-n \log(1 - F_1(c_1^\alpha - \varepsilon_1))\right) = \mathcal{O}(1) .$$

Hence, both terms $A_{21}$ and $A_{22}$ are upper bounded by constants, so $A_2 = \mathcal{O}(1)$.

**Upper bound on $A_3$** To upper bound $A_3$ we use the same discretization arguments as in (Riou and Honda, 2020). More precisely, we introduce a number of bins $M$ that is specified later in the proof, and for any $i \in \{1, \dots, n+1\}$ we define $\widetilde{y}_i = \frac{\lfloor M y_i \rfloor}{M}$ and $\widetilde{\mathcal{Y}}_n$ the corresponding set of truncated observations. Thanks to these definitions we can upper bound $\bar{A}_3$ as

$$A_3 \leq \sum_{n=1}^{T} \mathbb{E}_{\mathcal{Y}_n}\left[\mathbb{1}\left(C_\alpha\left(\widetilde{\mathcal{Y}}_n, w\right) < c_1^\alpha - \varepsilon_1\right) \frac{1}{\mathbb{P}_{w\sim\mathcal{D}_n}(C_\alpha(\tilde{\mathcal{Y}}_n, w) \geq c_1^\alpha - \varepsilon_1 - 1/M)}\right]$$
$$\leq \sum_{n=1}^{T} \mathbb{E}_{\mathcal{Y}_n}\left[\mathbb{1}\left(C_\alpha\left(\widetilde{\mathcal{Y}}_n, w\right) < c_1^\alpha - \varepsilon_1 - \frac{1}{M}\right) \frac{1}{\mathbb{P}_{w\sim\mathcal{D}_n}(C_\alpha(\tilde{\mathcal{Y}}_n, w) \geq c_1^\alpha - \varepsilon_1 - 1/M)}\right].$$

Now, we use the first result of Lemma 5.10 that provides with the additional observation on the last item

$$\mathbb{P}_{w\sim\mathcal{D}_n}(C_\alpha(\tilde{\mathcal{Y}}_n, w) \geq c_1^\alpha - \varepsilon_1 - 1/M) \geq \frac{1}{n}\frac{n!}{\prod_{i=1}^M \beta_i} \prod_{j=1}^M (p_j^\star)^{\beta_j},$$

where $p^\star \in \mathcal{P}^M$ is a well chosen weight vector. We choose $\varepsilon_1, M$ small enough so that the order of the CVaRs of the arms is preserved with the discretization. Writing $\varepsilon_1 + 1/M = \varepsilon_1'$. Using with a slight abuse of notations $\mathcal{M}$ the set of allocations considered in the expectation, we can further write

$$\begin{aligned} A_3 &\leq \sum_{n=1}^{T} \sum_{\beta\in\mathcal{M}} n \frac{\frac{n!}{\prod_{i=1}^M \beta_i!} \prod_{j=1}^M (\widetilde{p}_i)^{\beta_i}}{\frac{n!}{\prod_{i=1}^M \beta_i!} \prod_{j=1}^M (p_i^\star)^{\beta_i}} \\ &= \sum_{n=1}^{T} n \sum_{\beta\in\mathcal{M}} \prod_{j=1}^M \left(\frac{\widetilde{p}_i}{p_i^\star}\right)^{\beta_i} \\ &= \sum_{n=1}^{T} n \sum_{\beta\in\mathcal{M}} \exp\left(-n\left(\mathrm{KL}\left(\frac{\beta}{n}, \widetilde{p}\right) - \mathrm{KL}\left(\frac{\beta}{n}, p^\star\right)\right)\right) \end{aligned}$$

Interestingly, the discretization allows to consider for each possible allocation of $n$ data the ratio between the probability of pulling arm 1 in this situation, and the probability of drawing this allocation. This quantity is easy to interpret, and can be upper bounded by two terms involving the KL between the empirical frequency and some probability vector.

We can now choose $p^\star$ in order to have $\mathrm{KL}\left(\frac{\beta}{n}, p^\star\right) = \mathcal{K}_{\inf}^{\alpha,\mathcal{Y}}\left(\frac{\beta}{n}, c_1^\alpha - \varepsilon_1 - \frac{1}{M}\right)$. Furthermore, it holds by definition that $\mathrm{KL}\left(\frac{\beta}{n}, \widetilde{p}\right) \geq \mathcal{K}_{\inf}^{\alpha,\tilde{\mathcal{Y}}}\left(\frac{\beta}{n}, c_1^\alpha - \frac{1}{M}\right)$. We can hence write

$$\mathrm{KL}\left(\frac{\beta}{n}, \widetilde{p}\right) - \mathrm{KL}\left(\frac{\beta}{n}, p^\star\right) \geq \mathcal{K}_{\inf}^{\alpha,\tilde{\mathcal{Y}}}\left(\frac{\beta}{n}, c_1^\alpha - \frac{1}{M}\right) - \mathcal{K}_{\inf}^{\alpha,\mathcal{Y}}\left(\frac{\beta}{n}, c_1^\alpha - \varepsilon_1 - \frac{1}{M}\right),$$

which is strictly positive using basic properties of $\mathcal{K}_{\inf}$ on the set of bounded distributions. This is however not sufficient to conclude, but we can further introduce

$$\delta = \inf_{p\in\mathcal{P}^M : C_\alpha(\mathcal{Y},p) \leq c_1^\alpha - \varepsilon_1 - \frac{1}{M}} \left[\mathcal{K}_{\inf}^{\alpha,\tilde{\mathcal{Y}}}\left(p, c_1^\alpha - \frac{1}{M}\right) - \mathcal{K}_{\inf}^{\alpha,\tilde{\mathcal{Y}}}\left(p, c_1^\alpha - \varepsilon_1 - \frac{1}{M}\right)\right],$$

where $\widetilde{\mathcal{Y}}$ is $\widetilde{\mathcal{Y}}_n$ but with each item only repeated once (the set built from $\widetilde{\mathcal{Y}}_n$). The compacity of the set of distributions supported in $[0, B]$ ensures that $\delta > 0$, and we finally obtain

$$A_3 \leq \sum_{n=1}^{T} n^{M+1} \exp(-n\delta) = \mathcal{O}(1) \ .$$

Hence, $A_3$ is also upper bounded by a constant. This final result concludes the proof of optimality of B-CVTS for continuous bounded distribution, as we finally have

$$\text{(Pre-CV)} \leq A_1 + A_2 + A_3 = O(1) \ .$$

## 5.5 Additional theoretical results of practical interest

In this section we provide two results that are not presented in the initial paper on CVTS (Baudry et al., 2021a) but can be of interest for the practitioner. The first one is that the theoretical guarantees are preserved if CVTS runs in a batch setting where the number of participants in each batch is upper bounded by some constant, while the second shows that if the learner cannot provide an upper bound on the support setting $B = +\infty$ still leads to logarithmic regret guarantees when $\alpha < 1$.

### 5.5.1 Optimality in the batch setting

With the application of agriculture in mind, we consider running B-CVTS on $T$ time steps that we call *seasons* and assume that at the end season $t$ a number of arms $n_t \in [M]$ will be drawn, where $M \in \mathbb{N}$ is an unknown upper bound on the possible number of draws for each season.

More precisely, for each season $t \in [T]$ and for each integer $f \in [n_t]$ the B-CVTS algorithm chooses an action $A_{t,f}$ to play, where the choice of $A_{t,f}$ can depend only on the observations made up to (and including) season $t-1$. We call this algorithm *Batched–B-CVTS*.

With the following result we quantify the impact of the batched feedback on the CVaR regret of Batched–B-CVTS, and show that if the upper bound $M$ is independent of the time horizon the algorithm remains asymptotically optimal, as in the purely sequential setting.

**Theorem 5.11** (CVaR Regret of Batched–B-CVTS). *Consider a bandit problem $(F_1, \dots, F_K) \in \mathcal{F}^K$, with respective CVaR$_\alpha$ denoted by $(c_1, \dots, c_K)$ with $c_1 = argmax_{k=1,\dots,K} c_k$. Assume that the algorithm runs for $T$ seasons, and that at each season the size of the batch is $n_T \leq M \in \mathbb{N}$. Then, for any $\varepsilon > 0$ small enough there exists some $\varepsilon_1 > 0, \varepsilon_2 > 0$ such that the regret of Batched–B-CVTS satisfies*

$$\mathcal{R}_T^\alpha \leq \sum_{k=2}^{K} \Delta_k^\alpha \left( m_T^k + \mathbf{M} + 2\mathbf{M} \frac{e^{-2m_T^k \varepsilon_1^2}}{1 - e^{-2\,\varepsilon_1^1}} + C_{1,\varepsilon_2}^\alpha \right) ,$$

*where $m_T^k = \frac{\log(T) + \log(\mathbf{M})}{\mathcal{K}_{\inf}^\alpha(F_k, c_1) - \varepsilon}$ and $C_{1,\varepsilon_2}$ is a constant depending only on the distribution $F_1$, the family $\mathcal{F}$ and $\varepsilon_2$.*

We see that if $M$ is indeed a constant (i.e does not depend on the time horizon) when $T$ is large enough it has not impact on the scaling of the regret. Before proving this result we explain the intuition behind the new terms depending on $M$. First, the dominant term in $m_T^k = \mathcal{O}(\log(TM))$ is expected since it corresponds to what we would obtained by playing at most $TM$ times in the purely sequential setting. The main impact of the batch setting lies instead in the additive $M$, and corresponds to the sub-optimal arm being sampled for the whole season just before being sufficiently sampled to be identified as sub-optimal. Finally, the multiplicative $M$ term corresponds to the *empirical distribution* of arm $k$ begin mis-estimated and "costing" a full batch.

*Proof.* We decompose the expected number of pulls of each sub-optimal arm inside the cohort in a similar fashion as in the proof of Theorem 5.4. First, the expected number of pulls of arm $k$ during the total duration of the experiment is

$$\mathbb{E}[N_k(T)] = \mathbb{E}\left[ \sum_{t=1}^{T} \sum_{f=1}^{n_t} \mathbb{1}(A_{t,f} = k) \right] ,$$

where $A_{t,f}$ denotes the recommendation to farmer $f$ at season $t$. For any $m_T \in \mathbb{R}$ we decompose the expectation according to whether $N_k(t-1) \leq m_T$ or not. We handle the corresponding first term by considering the random variable $\tau = \{\sup_{t \leq T} : N_k(t-1) \leq m_T\}$. By construction, $\tau$ is the last season for which the total number of observations for arm $k$ is smaller than $m_T$. We obtain that

$$\begin{aligned} E_{T,1} &\coloneqq \sum_{t=1}^{T} \sum_{f=1}^{n_t} \mathbb{1}(A_{t,f} = k, N_k(t-1) \leq m_T) \\ &\leq \sum_{t=1}^{\tau} \sum_{f=1}^{n_t} \mathbb{1}(A_{t,f} = k, N_k(t-1) \leq m_T) + \sum_{t=\tau+1}^{T} \sum_{f=1}^{n_t} \mathbb{1}(A_{t,f} = k, N_k(t-1) \leq m_T) \end{aligned}$$

$$\leq N_k(\tau) + \sum_{f=1}^{n_{\tau+1}} \mathbb{1}(A_{\tau+1,f} = k)$$
$$\leq m_T + M \ ,$$

which is fully deterministic and gives the first two terms of the theorem. We now consider the case $N_k(t-1) \geq m_T$, that we further analyze according to the same events as for Theorem 5.4 , i.e using that

$$\mathbb{1}(F_k(t-1) \notin \mathcal{B}_{\varepsilon_1}(F_k)) + \mathbb{1}(F_k(t-1) \in \mathcal{B}_{\varepsilon_1}(F_k), \widetilde{c}_{k,t,f} \geq c_1 - \varepsilon_2) + \mathbb{1}(\widetilde{c}_{1,t,f} \leq c_1 - \varepsilon_2) \geq 1 \ ,$$

where $\mathcal{B}_{\varepsilon_1}(F_k)$ is an $\varepsilon_1$-Levy ball around $F_k$, and $\varepsilon_1, \varepsilon_2$ are two small positive constants. We denoted by $(\widetilde{c}_{k,t,f})$ the noisy CVaRS computed by the algorithm for any arm $k$, season $t$ and farmer $f$. Using the notation $E_{T,2} = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{f=1}^{n_t} \mathbb{1}(A_{t,f} = k, N_k(t-1) \geq m_T)\right]$ this leads to

$$E_{T,2} \leq \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{f=1}^{n_t} \mathbb{1}(A_{t,f} = k, N_k(t-1) \geq m_T, F_k(t-1) \notin \mathcal{B}_{\varepsilon_1}(F_k))\right]}_{e_1}$$
$$+ \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{f=1}^{n_t} \mathbb{1}(A_{t,f} = k, N_k(t-1) \geq m_T, F_k(t-1) \in \mathcal{B}_{\varepsilon_1}(F_k), \widetilde{c}_{k,t,f} \geq c_1 - \varepsilon_2)\right]}_{e_2}$$
$$+ \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{f=1}^{n_t} \mathbb{1}(A_{t,f} = k, N_k(t-1) \geq m_T, \widetilde{c}_{1,t,f} \leq c_1 - \varepsilon_2)\right]}_{e_3} \ .$$

**Upper bounding** $e_1$ Denoting by $F_{k,n}$ the empirical distribution of arm $k$ after a total number of *pulls* $n$ (instead of after season $t$), we obtain

$$e_1 \coloneqq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{f=1}^{n_t} \mathbb{1}(A_{t,f} = k, N_k(t-1) \geq m_T, F_k(t-1) \notin \mathcal{B}_{\varepsilon_1}(F_k))\right]$$
$$\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(N_k(t-1) \geq m_T, F_k(t-1) \notin \mathcal{B}_{\varepsilon_1}(F_k)) \sum_{f=1}^{n_t} \mathbb{1}(A_{t,f} = k)\right]$$
$$\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{n=m_T}^{T} \mathbb{1}(N_k(t-1) = n, F_{k,n} \notin \mathcal{B}_{\varepsilon_1}(F_k)) \sum_{f=1}^{n_t} \mathbb{1}(A_{t,f} = k)\right] \ ,$$

with a union bound on the number of pulls. Under $N_k(t-1) = n$ it holds that $F_k(t-1) = F_{k,n}$, and so we can further write that

$$\begin{aligned}
e_1 &\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{n=m_T}^{T} \mathbb{1}(N_k(t-1)=n, F_{k,n} \notin \mathcal{B}_{\varepsilon_1}(F_k)) \sum_{f=1}^{n_t} \mathbb{1}(A_{t,f}=k)\right] \\
&\leq \mathbb{E}\left[\sum_{n=m_T}^{T} \mathbb{1}(F_{k,n} \notin \mathcal{B}_{\varepsilon_1}(F_k)) \sum_{t=1}^{T}\sum_{f=1}^{n_t} \mathbb{1}(A_{t,f}=k, N_k(t-1)=n)\right] \\
&\leq M\mathbb{E}\left[\sum_{n=m_T}^{T} \mathbb{1}(F_{k,n} \notin \mathcal{B}_{\varepsilon_1}(F_k))\right] \\
&= M \sum_{n=m_T}^{+\infty} \mathbb{P}(F_{k,n} \notin \mathcal{B}_{\varepsilon_1}(F_k))
\end{aligned}$$

Finally, using the Dvoretzky–Kiefer–Wolfowitz inequality (Massart, 1990) we obtain

$$\begin{aligned}
e_1 &\leq M \sum_{n=m_T}^{+\infty} 2e^{-2n\,\varepsilon_1^2} \\
&\leq \frac{2Me^{-2m_T\,\varepsilon_1^2}}{1-e^{-2\,\varepsilon_1^2}} \,.
\end{aligned}$$

This upper bound holds for any choice of $m_T, \varepsilon_1$, and we remark that if $m_T \to +\infty$ then $e_1 \to 0$.

**Upper bounding** $e_2$ The term $e_2$ can be upper bounded with the exact same steps as in the purely sequential setting. Further using the continuity of $\mathcal{K}_{\inf}^{\alpha,\mathcal{D}}$ for any $\varepsilon > 0$ we can choose $\varepsilon_1, \varepsilon_2$ small enough such that the proof leads to

$$e_2 \leq M \times T \times e^{-m_T\left(\mathcal{K}_{\inf}^{\alpha,\mathcal{D}}(F_k,c_1)-\varepsilon\right)} \,.$$

Choosing $m_T = \frac{\log(TM)}{\mathcal{K}_{\inf}^{\alpha,\mathcal{D}}(F_k,c_1)-\varepsilon}$ then ensures that $e_2 \leq 1$ and $m_T$ becomes the dominant term in the regret bound.

**Upper bounding** $e_3$ The final term is the one leading to the most complicated part of the analysis of B-CVTS. Fortunately, the batch setting have no impact on this part, so we can directly reuse the upper bound of $e_3$ in the proof of Theorem 5.4. Indeed, we can re-write $e_3$ to make it equivalent to the corresponding term in the purely sequential problem:

$$e_3 = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{f=1}^{n_t} \mathbb{1}(\widetilde{c}_{1,t,f} \leq c_1 - \varepsilon_2)\right] = \mathbb{E}\left[\sum_{r=1}^{S_T} \mathbb{1}(\widetilde{c}_1(r) \leq c_1 - \varepsilon_2)\right] ,$$

where $\widetilde{c}_{1,t,f}$ denotes the noisy CVaR computed at season $t$ for farmer $f$, while $(\widetilde{c}_1(r))_{r\geq 1}$ represent the same quantities but assigning CVaRs computed in the same batch an arbitrary order. We used the notation $S_T = \sum_{t=1}^{T} n_t \leq MT$.

We further remark that contrarily to the previous terms, the upper bound of $A_3$ does not depend on $M$ at all since the upper bound is a convergent series. □

### 5.5.2 Logarithmic regret with $B = +\infty$

This result comes from a discussion with Junya Honda, after talking about the experimental results of the next section reporting that the performance of B-CVTS does not seem to be altered by setting a conservative upper bound. The question is then to determine how far this observation can hold, and pushing it to the limit what guarantees we would obtain by setting $B = +\infty$. The striking result that we obtained is that if $\alpha < 1$ then B-CVTS still attains logarithmic regret, and the upper bound of the expected number of pulls still matches the lower bound in some "easy enough" cases. However, this comes at the price of losing asymptotic optimality in general.

Before stating the result we explain how $B = +\infty$ can be allowed in B-CVTS. First, we recall that for a probability distribution defined by a probability $w = (w_1, \ldots, w_{n+1})$ and a set $\mathcal{X} = (X_1 < \cdots < X_{n+1})$ the CVaR can be expressed using $n_\alpha(w) = \inf\{j \in [n+1] : \sum_{i=1}^{j} w_i \geq \alpha\}$ as

$$C_\alpha(\mathcal{X}, w) = \frac{1}{\alpha}\left(\sum_{i=1}^{n_\alpha - 1} w_i X_i + \left(\alpha - \sum_{i=1}^{n_\alpha - 1} w_i\right) X_{n_\alpha}\right)$$

if $n_\alpha(w) \geq 2$, otherwise $C_\alpha(\mathcal{X}, w) = X_1$. Hence, we remark that if $n_\alpha \leq n+1$ (or equivalently, $w_{n+1} > 1 - \alpha$) then the value of exact $X_{n+1}$ does not matter in the computation of $C_\alpha(\mathcal{X}, w)$, it just needs to be larger than $X_n$. This allows us to adapt B-CVTS as follows: for a given arm, draw $w \sim \mathcal{D}_{n+1}$ and compute $n_\alpha(w)$. If $n_\alpha(w) = n+1$ the bonus is included in the computation, we then set the noisy CVaR of this arm to $+\infty$. Otherwise, we use the above formula to compute the perturbed CVaR by replacing $(X_1, \ldots, X_{n_\alpha(w)})$ by their corresponding values in the history of the arm. These two steps are actually at the core of the proof of the following Theorem 5.12.

**Theorem 5.12** (Regret of B-CVTS with $B = +\infty$)**.** *Consider $\boldsymbol{\nu} = (\nu_1, \dots, \nu_K)$ a bandit, and denote the best CVaR by $c_\star^\alpha = \max_{k\in\{1,\dots,K\}} CVaR_\alpha(\nu_k)$. Assume that $\forall k \in \{1, \dots, K\}$, $\nu_k$ belongs the family $\mathcal{C}^{B_k}$ of continuous distributions supported in $[a_k, B_k]$ for some **unknown** $B_k > 0$ and $a_k \leq B_k$. Then, for any $\varepsilon > 0$ the regret of B-CVTS **using the exploration bonus** $B = +\infty$ on $\boldsymbol{\nu}$ satisfies*

$$\mathcal{R}_{\boldsymbol{\nu}}^\alpha(T) \leq \sum_{k:\Delta_k^\alpha>0} \Delta_k^\alpha \frac{1}{\min\left(\log(1/\alpha), \mathcal{K}_{\inf}^{\alpha,\mathcal{C}^{B_k}}(\nu_k, c_\star^\alpha) - \varepsilon\right)} \log T + \mathcal{O}_\varepsilon(1) \ .$$

*Proof.* The analysis again relies on the upper and lower bounds we can obtain for the BCP if $B = +\infty$. First, the pre-convergence term is trivially smaller than with a finite $B$, and does not need to be further examined here. We now have a look at the upper bound of the BCP, with a quite straightforward proof. Denoting by $\mathcal{Y}^B$ the history augmented by $B$ ($\mathcal{Y}^\infty$ for $B = +\infty$) we first distinguish the case whether $B$ is actually "used" in the noisy CVaR or not,

$$\begin{aligned}\mathbb{P}(C_\alpha(\mathcal{Y}^\infty, w) \geq c) &= \mathbb{P}(C_\alpha(\mathcal{Y}^\infty, w) = +\infty) + \mathbb{P}(C_\alpha(\mathcal{Y}^\infty, w) \geq c, C_\alpha(\mathcal{Y}^\infty, w) < +\infty) \\ &= \mathbb{P}(w_{n+1} \geq 1-\alpha) + \mathbb{P}(C_\alpha(\mathcal{Y}^\infty, w) \geq c, w_{n+1} < 1-\alpha) \ .\end{aligned}$$

Now we can use that the marginal distribution of $w_{n+1}$ is a beta distribution $\beta(1, n)$, and that under the events considered in the second term the $+\infty$ item is not considered in the computation of the CVaR, hence it could in fact be replaced by anything larger than the maximum value collected so far. For this reason, for bounded distributions we can simply write

$$\begin{aligned}\mathbb{P}(C_\alpha(\mathcal{Y}^\infty, w) \geq c) &= \alpha^n + \mathbb{P}(C_\alpha(\mathcal{Y}^\infty, w) \geq c, w_{n+1} < 1-\alpha) \\ &= \alpha^n + \mathbb{P}(C_\alpha(\mathcal{Y}^B, w) \geq c, w_{n+1} < 1-\alpha) \\ &\leq \alpha^n + \mathbb{P}(C_\alpha(\mathcal{Y}^B, w) \geq c) \\ &\leq \alpha^n + \exp\left(-n\mathcal{K}_{\inf}^{\alpha,\mathcal{C}_B}(\widetilde{F}_B, c)\right) \\ &\leq 2\exp\left(-n\min\left\{\log(1/\alpha), \mathcal{K}_{\inf}^{\alpha,\mathcal{C}_B}(\widetilde{F}_B, c)\right\}\right) \ ,\end{aligned}$$

where the second line comes from the fact that if $w_{n+1} < 1-\alpha$ the last term is not used in the CVaR and could be anything larger than the other items, and the second term on the last line comes from the upper bound we provided in previous works. $\widetilde{F}_B$ is the empirical distribution of the history augmented by $B$. Finishing the proofs requires only to use this

upper bound on the BCP and follow the same steps as for Theorem 5.4 for the post-convergence term. □

We believe this result is somehow surprising, or at least quite original in bandits as it is known that in the standard setting ($\alpha = 1$) it is impossible to obtain logarithmic regret without knowledge of the upper bound (Hadiji and Stoltz, 2020). A very interesting remaining question is to determine whether this algorithm could have guarantees for a broader family of distributions (e.g light-tailed?). This would require to derive new techniques at different steps of the proof. A related question is to analyze if the function $\mathcal{K}_{\inf}^{\alpha,+\infty}$ would be actually well defined, continuous in all the arguments and if the upper bound we provide matches it, or could be improved to match it.

## 5.6 Experiments

In this section we report the results of experiments with the algorithms presented in the previous sections, first on synthetic examples, and then on a use-case study in agriculture based on the DSSAT agriculture simulator. With this second set of experiments we aim at showing the potential of B-CVTS to tackle the realistic problem introduced in the preamble of this thesis.

The experiments presented here are a selection of those displayed in (Baudry et al., 2021a), and we chose to focus on the B-CVTS algorithms as in the rest of this chapter. The conclusions made in this section are however still valid for M-CVTS.

### 5.6.1 Preliminary Experiments on Synthetic Examples

We first performed various experiments on synthetic data in order to check the good practical performance of B-CVTS on settings that are simple to implement and are good illustrative examples of the performance of the algorithms.

**Truncated Gaussian Mixtures (TGM)** In this section we consider bounded multi-modal distributions, built by truncated Gaussian Mixture models in $[0, 1]$. We call these distributions Truncated Gaussian Mixtures (TGM for short). We first remark that these distributions are not continuous because they can have a positive mass in $0$ and $1$, but it is still a good illustrative example to check the performance of B-CVTS. Indeed, as a sanity check we performed all the experiments presented in this part by making the distributions continuous (instead of truncating, we re-sampled observations until they were in $(0, 1)$) and the results were deemed to be exactly the same.

**First set of experiments** We first consider experiments with two modes, each mode being equi-probable and having the same variance for simplicity ($\sigma = 0.1$) in all experiments. It is for example interesting to set up experiments with arms whose modes are relatively close, and other arms that have a large mass of probability close to the two support bounds (e.g one mode close to 1 and one close to 0).

We experiment 4 possible configurations of the modes, denoted by parameters $(\mu_i)_{i\in\{1,\dots,4\}}$: $\mu_1 = (0.2, 0.5)$ (arm 1), $\mu_2 = (0, 1)$ (arm 2), $\mu_3 = (0.3, 0.6)$ (arm 3), $\mu_4 = (0.1, 0.65)$ (arm 4).

Before detailing the experiments we can look at the CVaRs of these distributions for different values of parameter $\alpha$. For example, arm 2 has a larger mean than the one with arm 1, but the $50\%$ CVaR of 1 is larger. We represented the CVaR for each parameter for different values of $\alpha \in (0, 1]$ in Figure 5.1, with the thresholds $\alpha \in \{0\%, 10\%, 90\%\}$ (used in our experiments) represented by the vertical lines. Interestingly, with these arms the most difficult problems are not necessarily those with smallest values of $\alpha$. Indeed, for $\alpha = 80\%$ it may be particularly difficult to choose between arm 2 and 3, or between arm 1 and 4, while arm 3 is the clear winner for $\alpha = 10\%$ due to the distribution being very concentrated around 0.5. Furthermore, the arm 2 provides observations mostly around 0 and 1 but has a larger mean than the others, so it becomes the best arm for values of $\alpha$ that are close to 1.

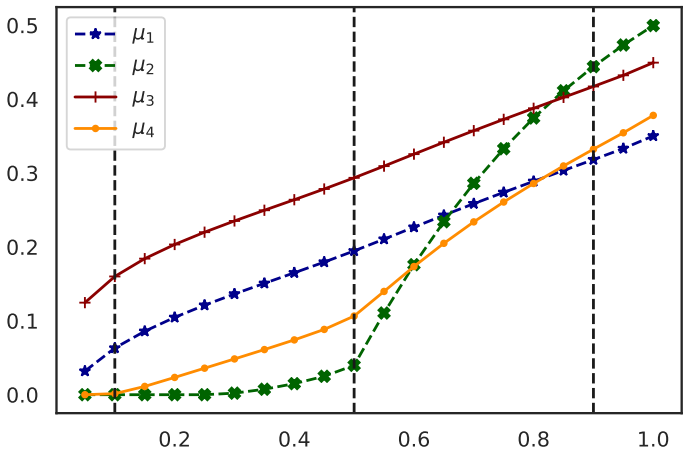


**Figure 5.1** – CVaR of each TGM distribution $\nu_i$ (with centers $\mu_i$), $i = 1, \dots, 4$ for different values of $\alpha$

We run the algorithms for $\alpha = 10\%, 50\%$ and $90\%$ on four bandit problems with the following characteristics : arm 1 and 2 (Exp. 1), arm 1 and 3 (Exp. 2), arm 1 and 4 (Exp. 3), and all arms from 1 to 4 (Exp. 4). In Tables 5.1, 5.2, 5.3 and 5.4 we report the results for the four considered problems (mean regret and standard deviation at $T = 10000$). On these examples B-CVTS significantly outperforms the two UCB algorithms for all levels of $\alpha$.

**Table 5.1** – CVaR regret (average and std) for Exp. 1 at $T = 10000$ for 5000 replications.

| $\alpha$ | U-UCB | CVaR-UCB | B-CVTS |
|---|---|---|---|
| 10% | 274.9 (1.8) | 5.3 (1.5) | **1.1 (0.5)** |
| 50% | 127.0 (19.3) | 135.3 (41.1) | **29.8 (17.2)** |
| 90% | 80.5 (10.4) | 53.5 (6.7) | **10.2 (17.9)** |

**Table 5.2** – CVaR regret (average and std) for Exp. 2 at $T = 10000$ for 5000 replications.

| $\alpha$ | U-UCB | CVaR-UCB | B-CVTS |
|---|---|---|---|
| 10% | 373.7 (4.1) | 72.8 (9.6) | **4.1 (2.3)** |
| 50% | 135.8 (8.9) | 37.9 (7.5) | **5.5 (2.7)** |
| 90% | 62.6 (7.1) | 43.9 (5.1) | **5.0 (1.8)** |

**Table 5.3** – CVaR regret (average and std) for Exp. 3 at $T = 10000$ for 5000 replications.

| $\alpha$ | U-UCB | CVaR-UCB | B-CVTS |
|---|---|---|---|
| 10% | 269.4 (1.8) | 23.2 (4.8) | **2.8 (1.5)** |
| 50% | 138.5 (12.4) | 71.8 (19.0) | **14.7 (8.3)** |
| 90% | 53.1 (6.6) | 34.5 (6.6) | **20.2 (22.4)** |

**Table 5.4** – CVaR regret (average and std) for Exp. 4 at $T = 10000$ for 5000 replications.

| $\alpha$ | U-UCB | CVaR-UCB | B-CVTS |
|---|---|---|---|
| 10% | 958.9 (4.8) | 230.5 (25.3) | **10.4 (3.2)** |
| 50% | 318.4 (12.2) | 147.7 (17.9) | **21.2 (6.4)** |
| 90% | 154.3 (11.9) | 119.5 (11.7) | **25.1 (14.1)** |

**Testing $\alpha = 1\%$** We then check the robustness of B-CVTS to a smaller value of the parameter $\alpha$ by setting $\alpha = 1\%$. The bandit of Experiment 5 (Exp. 5) has six TGM arms with respective mean and variance parameters $\mu_{135} = (0.3, 0.6)$, $\mu_{246} = (0.25, 0.65)$, $\sigma_{12} = 0.05$, $\sigma_{34} = 0.06$, $\sigma_{56} = 0.07$. This experiment allows to additionally check if adding different variances to the arms affects the performance of the algorithms. However, we keep the probability of each mode to $0.5$. This problem provides the following CVaR values for each arm at level $1\%$: $c_{1:6}^{0.01} = [0.18, 0.13, 0.15, 0.10, 0.13, 0.08]$. The results are reported in Table 5.5, in which we observe a very large performance gap between B-CVTS and UCB algorithms. This is particularly interesting because it shows that the UCB algorithms are not really able to learn for very small values of $\alpha$ (indeed $\alpha = 1\%$ is very small when drawing only a total number of $10^4$ observations) before the horizon becomes extremely large. We already observed this behavior for CVaR-UCB in previous experiments, but this time we can see as well that its average regret is even higher than the one of U-UCB, and its variance spiked. On the other hand, B-CVTS seems to learn smoothly even for $\alpha = 1\%$, as its average regret only doubles between $T = 1000$ and $T = 5000$, and increases even less between $T = 5000$ and $T = 10000$.

**Random Problems with more modes and more arms** Finally, we further check the robustness of B-CVTS to more arms and more diverse distribution profiles by increasing the number of possible modes. To do so, we implement an experiment with $K = 30$ arms, with TGM distributions with 10 modes exhibiting different means and variances, which covers a large variety of shapes of distributions. All of those parameters are drawn uniformly at random, and we summarize their distributions as $(\mu, \sigma) \sim \mathcal{U}([0.25, 1]^{10} \times [0, 0.1]^{10})$, and $p \sim \mathcal{D}_{10}$ (uniform

**Table 5.5** – CVaR regret (average and std) for Exp. 5 at $\alpha = 1\%$, for $T \in \{10^3, 5 \times 10^3, 10^4\}$ for 5000 replications.

| T | U-UCB | CVaR-UCB | B-CVTS |
|---|---|---|---|
| 1000 | 49.1 (0.3) | 53.2 (5.6) | **18 (37)** |
| 5000 | 245 (1.1) | 263.2 (24.7) | **35.5 (51)** |
| 10000 | 489.1 (2.2) | 518.4 (45.0) | **41 (66)** |

distribution on the simplex, presented in Section 5.3). We name this setting TGM Experiment 6. The results of this experiment are reported in Table 5.6 for a parameter $\alpha = 0.05$ averaged over 400 random instances. Again, we choose a smaller value for $\alpha$ than in the previous extensive sets of experiments because problems with small $\alpha$ seem to be more challenging. The results highlight that best performances are obtained by B-CVTS.

**Table 5.6** – CVaR regret (average and std) for Exp. 6, $\alpha = 5\%$, averaged over 400 random instances for $T \in \{10^4, 2 \times 10^4, 4 \times 10^4\}$ for 5000 replications.

| T | U-UCB | CVaR-UCB | B-CVTS |
|---|---|---|---|
| 10000 | 2149.9 (263) | 2016.0 (265) | **210.9 (6.4)** |
| 20000 | 4276.4 (538) | 3781.3 (521) | **237.1 (15.4)** |
| 40000 | 8493.4 (1085) | 6894.1 (985) | **263.5 (17.9)** |

**Summary** We preliminary evaluated the CVaR bandit algorithms on synthetic problems before testing them on a realistic-world bandit environment in the next section. These experiments seem to highlight a greater robustness of B-CVTS to many different settings regarding several parameters: the risk-level $\alpha$, the number of arms $K$ and the different possible shapes of the distributions (materialized by the number of modes and variances in our synthetic experiments). In particular, B-CVTS is the only algorithm that has not shown to be affected by the value of $\alpha$, as the two UCB algorithms had their respective performances degraded to some extent depending on $\alpha$ values.

### 5.6.2 Experiments with `DSSAT` crop-model

**Setting** In this section we study a simplified version of the problem in agriculture we introduced in the preamble of this thesis. We consider the choice of the best planting date for a maize crop in conditions emulating the types of soil and climate that can be found in Southern

Mali. We use the `DSSAT` simulator, to test our algorithms on this problem *in silico* decision. We specifically address maize planting date decision, as maize is a crucial crop for global food security (Shiferaw et al., 2011). Each simulation is assumed to be realistic, and starts from the same field initial conditions as ground measured. The simulator takes as input historical weather data, field soil measures, crop specific genetic parameters and a given crop management plan. Modeling is based on simulations of atmospheric, soil and plants compartments and their interactions. In the considered experiments, after a decision is made on planting date in the simulator, daily stochastic meteorologic features are generated according to historical data (Richardson and Wright, 1984) and injected in the complex crop model. At the end of crop cycle, a maize grain yield is measured to evaluate decision-making. We parameterized the crop-model under challenging rainfed conditions on shallow sandy soils, i.e. with poor water retention and fertility. Such experiment intends to be representative of realistic conditions faced by small-holder farmers under heavy environmental constraints, such as in Sub-Saharan Africa. Thus, this setting can help picturing how CVaR bandits may perform in real-world conditions. Furthermore, we recall that that risk-aware bandits are particularly relevant for this experiment since bad trials are deleterious to the farmers, that rely on their harvest for the subsistence of their household. Depending on her profile, a farmer may be more or less risk averse, and the *Conditional Value at Risk* can be used to personalize her level of risk-aversion. For instance, a small-holder farmer looking for food security may seek to avoid very poor yields compromising auto-consumption (e.g $\alpha \leq 20\%$), while a market-oriented farmer may be more prone to risky choices in order to increase her profit but still not risk neutral (e.g $\alpha = 80\%$). Furthermore, yield distributions are supposed to be *bounded*. Indeed, a finite yield potential can be defined under non-stressing conditions for a given crop and environment (Evans and Fischer, 1999; Tollenaar and Lee, 2002). Observed yields can be modeled as following Von Liebig's law of minimum (Paris, 1992): limiting factors will determine how much of the yield potential can be expressed. Hence, this experiment fits the theoretical setting we consider in this chapter. Finally, we built a bandit-oriented Python wrapper to `DSSAT` that we made available[1] to the bandit community for reproducibility.

**Experiment with 4 planting dates** We test bandit performances on the 4 armed `DSSAT` environment described in Table 5.7. To illustrate the non-parametric nature of these distributions, we report in Figure 5.2 estimations of their density obtained with Monte-Carlo simulations, as well as of their CVaRs. The resulting distributions are typically *multi-modal*, with one of their mode very close to zero (years of bad harvest), and with upper tails that cannot be properly characterized. However the practitioner can realistically assume that the distributions are upper-bounded, due to the physical constraints of crop-farming. The yield upper-bound is set to 10 t/ha thanks to expert knowledge for the considered conditions.

[1] https://github.com/rgautron/DssatBanditEnv

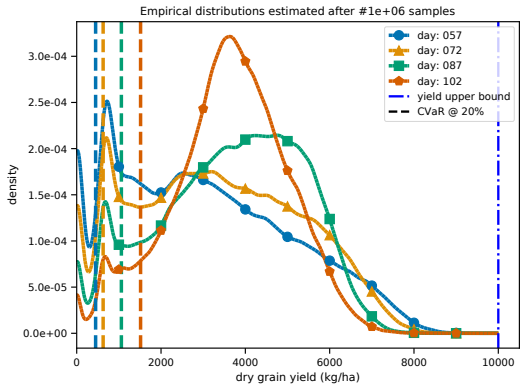


**Figure 5.2** – Empirical simulated yields and respective CVaRs at 20% estimated after $10^6$ samples in `DSSAT` environment.

| day (action) | $\text{CVaR}_\alpha$ | | | |
|---|---|---|---|---|
| | 5% | 20% | 80% | 100% |
| 057 | 0 | 448 | 2238 | 3016 |
| 072 | 46 | 627 | 2570 | 3273 |
| 087 | 287 | 1059 | 3074 | **3629** |
| 102 | **538** | **1515** | **3120** | 3586 |

**Table 5.7** – Empirical yield distribution metrics in kg/ha estimated after $10^6$ samples in `DSSAT` environment

The presented `DSSAT` environment advocates for the use of algorithms specifically designed for CVaR bandits, as the optimal arm can change depending on the value of the parameter $\alpha$. Our experiment consists in running 64 trajectories for three algorithms U-UCB, CVaR-UCB and B-CVTS defined in Section 5.2. Experiments are carried out with an horizon of $10^4$ time steps, and we compare the results for each algorithm for $\alpha \in \{5\%, 20\%, 80\%\}$ to see how the parameter impacts their performance. Indeed we want a strategy to perform well on all $\alpha$ choices, allowing to freely model any farmer's risk aversion level. As shown in Figure 5.3 and Table 5.8, B-CVTS appears to be consistently better than its UCB counterparts in `DSSAT` environment for all tested $\alpha$ values, which is encouraging for real-life applications.

| $\alpha$ | U-UCB | CVaR-UCB | B-CVTS |
|---|---|---|---|
| 5% | 3128 (3) | 760 (14) | **192 (11)** |
| 20% | 4867 (11) | 1024 (17) | **202 (10)** |
| 80% | 1411 (13) | 888 (13) | **287 (12)** |

**Table 5.8** – Empirical yield regrets at horizon $10^4$ in t/ha in `DSSAT` environment, for 1040 replications. Standard deviations in parenthesis.

**Experiment with 7 planting dates** We consider a bandit instance that consists of 7 arms, each arm corresponds to a planting date spaced of 15 days from the previous one. An illustration of the underlying distributions is given in Figure 5.4. In this case, the best arm is consistent with all values of $\alpha$, as shown in Table 5.9. Nevertheless, arms exhibit different gaps when considering different values of $\alpha$. This experiment intends to evaluate B-CVTS performance with a greater number of real-world alike arms with a diversity of reward distribution shapes.

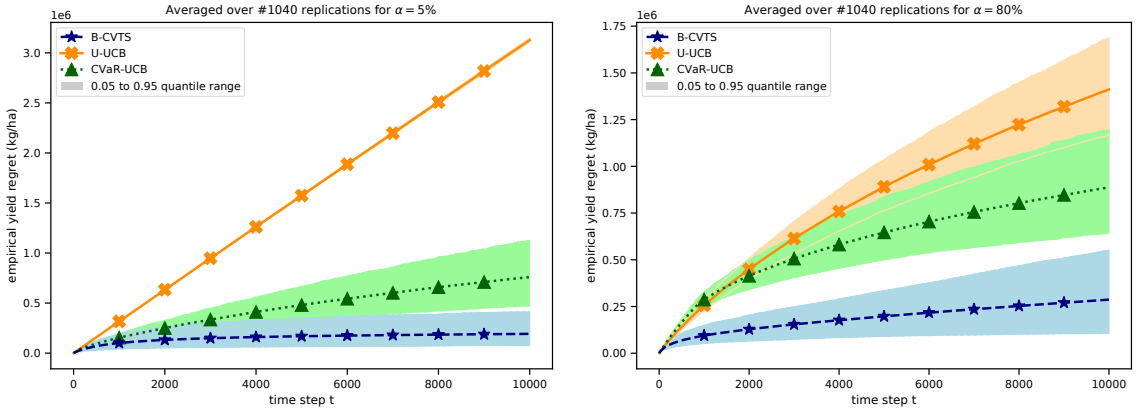


**Figure 5.3** – Regret comparison in `DSSAT` environment, averaged over 1040 experiment replications, $\alpha = 5\%$ (Left) and $\alpha = 80\%$, along with $90\%$ confidence intervals.

The results are reported in Table 5.10. Furthermore, the regret curves for the three algorithms, with $\alpha \in \{20\%, 80\%\}$ parameter are illustrated in Figure 5.5.

In this experiment, by exhibiting superior performances B-CVTS appears to be more robust than the algorithms based on UCB for CVaR bandits when we increase the number of arms. In practice for the planting-date problem, a global, few months planting-window is known but needs further refinements e.g. to identify the best two-week time slot for planting. That is to say, the number of arms is unlikely to be greater that what has been tested in this experiment, making B-CVTS a particularly fit-for-purpose candidate in this setup.

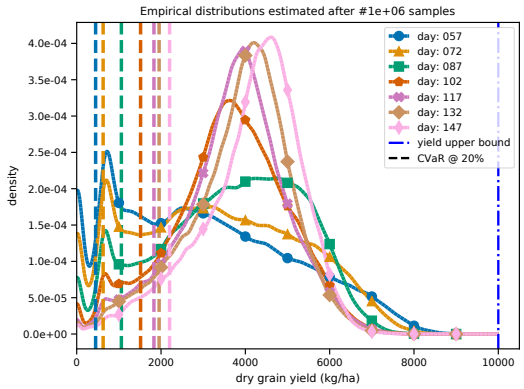


**Figure 5.4** – Monte-Carlo estimate of the distributions using $10^6$ samples from DSSAT; 7-armed problem (Left) and 4-armed problem with over-estimated upper bound.

| day (action) | $\text{CVaR}_\alpha$ | | | |
|---|---|---|---|---|
| | 5% | 20% | 80% | 100% |
| 057 | 0 | 448 | 2238 | 3016 |
| 072 | 46 | 627 | 2570 | 3273 |
| 087 | 287 | 1059 | 3074 | 3629 |
| 102 | 538 | 1515 | 3120 | 3586 |
| 117 | 808 | 1832 | 3299 | 3716 |
| 132 | 929 | 1955 | 3464 | 3850 |
| 147 | **1122** | **2203** | **3745** | **4112** |

**Table 5.9** – 7-armed distributions CVaRs for different levels of $\alpha$

| $\alpha$ | U-UCB | CVaR-UCB | B-CVTS |
|---|---|---|---|
| 5% | 5687 (5) | 1891 (18) | **700 (22)** |
| 20% | 6445 (10) | 1795 (19) | **489 (17)** |
| 80% | 3367 (14) | 1580 (15) | **293 (8)** |

**Table 5.10** – Results for DSSAT 7-armed experiment, empirical regret at $T = 10000$ in t/ha for 1040 replications. Standard deviations in parenthesis.

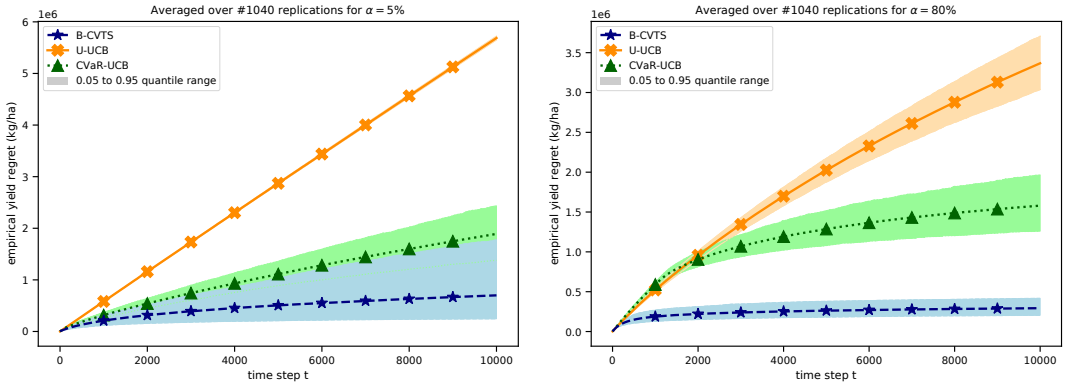


**Figure 5.5** – Regret comparison with the 7-armed DSSAT environment, averaged over 1040 experiment replications, $\alpha = 5\%$ (Left) and $\alpha = 80\%$, along with 90% confidence intervals.

**Impact of support upper bound over-estimation** In this experiment we get back to the 4-armed problem, but here we largely over-estimate the yield upper-bound to 30 t/ha, when a close to reality yield upper bound is about 10 t/ha. From an agronomic point of view, this yield value is a very unlikely over-estimation in the given conditions. This experiment intends to empirically evaluate how a rough arms' upper-bound estimation affects algorithms' performances, when little expert knowledge is available. An illustration of the underlying distributions and how the upper-bound estimation is exaggerated is given in Figure 5.4.

We provide the results of this experiment in Table 5.11, and display the regret curves in Figures 5.7.

This experiment addresses one possible concern for practitioners: the prerequisite of rewards' support upper bound. We empirically demonstrate that with realistic simulations, when a highly over-estimated, unrealistic support upper-bound is given to all algorithms – triple of expert's estimation –, B-CVTS keeps outperforming UCB-like CVaR bandit algorithms. We show that this over-estimation did not affect B-CVTS performances compared to the situation of correct support upper-bound identification as presented in Section 5.6. In particular, it even slightly improved its performance for $\alpha = 80\%$. This result is counter-intuitive, but it can be explained by the fact that the extra exploration induced by the larger upper bound may have sped up learning in this particular case, improving overall performances. On the other hand,

CVaR-UCB seems much more impacted by this over-estimation (regret is respectively increased by about $150\%$, $75\%$ and $78\%$ for $\alpha \in \{5\%, 20\%, 80\%\}$). Similarly U-UCB's performance is altered, despite its already unsatisfying results when fed the true upper bound.

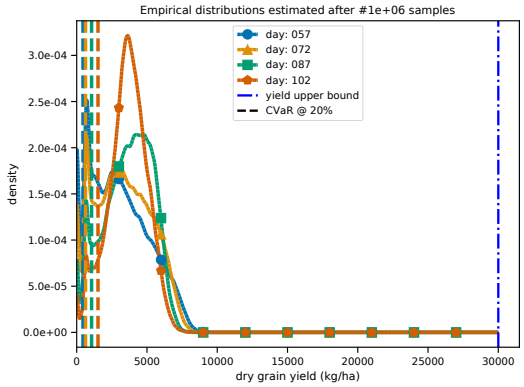


**Figure 5.6** – Illustration of the over-estimated upper bound with the empirical distributions of Figure 5.2.

| $\alpha$ | U-UCB | CVaR-UCB | B-CVTS |
|---|---|---|---|
| $5\%$ | 3179 (2) | 759 (14) | **195 (11)** |
| $20\%$ | 5644 (6) | 1020 (17) | **202 (10)** |
| $80\%$ | 2642 (10) | 888 (13) | **284 (12)** |

**Table 5.11** – Results for `DSSAT` Empirical regret at $T = 10000$ in t/ha for 1040 replications for the $4$-armed experiment with over-estimated upper bound.

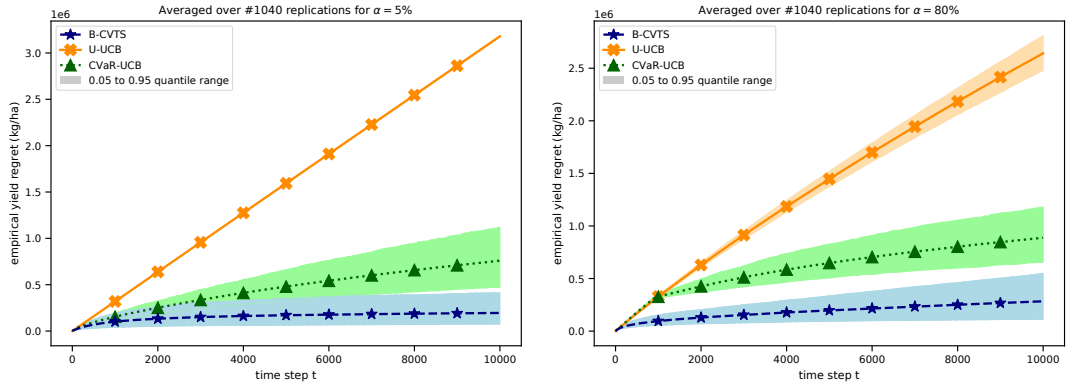


**Figure 5.7** – Regret comparison with the $4$-armed `DSSAT` environment and an over-estimated upper bound, averaged over 1040 experiment replications, $\alpha = 5\%$ (Left) and $\alpha = 80\%$, along with $90\%$ confidence intervals.

**Conclusion of the experiments** B-CVTS appeared to be a satisfying candidate for real-world alike problems, as shown with the planting date bandits. We empirically showed that the B-CVTS algorithm was best able to deal with a greater number of planting date arms than its UCB counterparts. We showed as well that B-CVTS remained the best performer despite considering a very unlikely support upper-bound estimation. We think that in many physical resource-based problems, this should be reassuring for practitioners, in particular when compared with UCB algorithms' sensibility to the input upper bound.

## 5.7 Appendix A: Basic properties of the Dirichlet distribution

We consider the Dirichlet distribution Dir($\alpha$) for some parameter $\alpha = (\alpha_1, \dots, \alpha_n)$. Let $w = (w_1, \dots, w_n)$ be a random variable drawn from the distribution Dir($\alpha$). We first recall that $w$ takes its values in the probability simplex $\mathcal{P}^n = \{p \in [0,1]^n : \sum_{i=1}^n p_i = 1\}$. The distribution admits the following density,

$$f(w_1, \dots, w_n) = \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\prod_{i=1}^n \Gamma(\alpha_i)} \prod_{i=1}^n w_i^{\alpha_i - 1} ,$$

where $\Gamma$ denotes the Gamma function. In this manuscript we only consider integer values for the coefficient $(\alpha_i)_{i \in \mathbb{N}}$, and for any $m \in \mathbb{N}$ $\Gamma(m) = (m-1)!$.

This distribution has convenient properties. First, using the notation $\sum_{i=1}^n \alpha_i = N$ and interpreting $\alpha_i / N$ as the frequency of an item in a set of observations drawn from a finite collection (empirical distribution), and $w$ a random re-weighting of these observation providing a "noisy" empirical distribution, the Dirichlet distribution ensures that the new frequency of each item is unbiased with respect to the observed frequency, with a variance that is inversely proportional to the total number of items collected. For any $i \in [1, n]$,

$$\mathbb{E}[w_i] = \frac{\alpha_i}{N} , \qquad \text{and} \qquad \mathbb{V}(w_i) = \frac{\alpha_i (N - \alpha_i)}{N^2 (N+1)} ,$$

and the marginal density of each component of $w$ is actually a distribution Beta($\alpha_i, N - \alpha_i$). This explains the use of the Dirichlet distribution to generalize the Beta-Bernoulli Thompson Sampling.

In this manuscript we also use two main properties of the Dirichlet distribution, both using the relation between the Dirichlet distribution and the Exponential distribution. Let $R_1, \dots, R_n$ be $n$ i.i.d random variables drawn from exponential distributions with respective parameters $\alpha_i$, $R_i \sim \mathcal{E}(\alpha_i)$. Then the vector $w = (w_1, \dots, w_n)$ with $w_i = \frac{R_i}{\sum_{j=1}^n R_j}$ follows a Dirichlet distribution Dir($\alpha$).

The second property is a consequence of the first one, and is that the components of a random variable drawn from a Dirichlet distribution can be aggregated, providing another Dirichlet distribution: if $w \sim$ Dir($\alpha$), then $w' = (w_1, \dots, w_i + w_j, \dots, w_n) \sim$ Dir($(\alpha_1, \dots, \alpha_i + \alpha_j, \dots, \alpha_n)$) (putting the sum in the $i-th$ slot and removing the $j$-th slot without changing the other indices).

# Chapter 6

# Dirichlet Sampling Beyond Bounded Rewards

In Chapter 5 we showed that NPTS can be successfully of extended to CVaR bandits when distributions are bounded. However, the existence and knowledge of the upper bound may sometimes not be precisely accessible to the practitioner, raising the question of the robustness of bandit algorithms to model misspecification. In this chapter we extend this strategy for alternative assumptions on the distributions. We study a generic *Dirichlet Sampling* (DS) algorithm, based on pairwise comparisons of empirical indices computed with *re-sampling* of the arms' observations and a data-dependent *exploration bonus*. We propose variants of this strategy achieving respectively optimal regret when the distributions are bounded and logarithmic regret for semi-bounded distributions with a mild quantile condition. Furthermore, a simple tuning can lead to consistent guarantees inside a large class of unbounded distributions, at the cost of slightly larger than logarithmic regret. We finally provide numerical experiments further showing the merits of DS in the decision-making problem on synthetic agriculture data introduced in Chapter 5. The results we present were published in (Baudry et al., 2021c).

## Contents

## 6.1 Introduction

In this chapter we get back to the standard bandit problem, where the learner sequentially chooses an action (arm) and collect a reward, with the objective of maximizing the expected sum of rewards. We recall that this is equivalent to minimizing the regret, defined as

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \mu^\star - \mu_{A_t}\right] = \sum_{k=1}^{K} \Delta_k \mathbb{E}\left[N_k(T)\right] \ , \tag{6.1}$$

where $N_k(T) = \sum_{t=1}^{T} \mathbb{1}(A_t = k)$ denotes the number of selections of arm $k$ after $T$ time steps, $\mu^\star = \max_{j\in\{1,\dots,K\}} \mu_j$ and $\Delta_k = \mu^\star - \mu_k$ is called the *gap* between arm $k$ and the largest mean. We also recall that if the arms' distributions $(\nu_1, \dots, \nu_K)$ all belong to a family of distributions $\mathcal{F}$, a uniformly efficient[1] bandit algorithm on $\mathcal{F}$ must satisfy

$$\liminf_{T\to\infty} \frac{\mathcal{R}_T}{\log(T)} \geq \sum_{k:\Delta_k>0} \frac{\Delta_k}{\mathcal{K}_{\inf}^{\mathcal{F}}(\nu_k, \mu^\star)} \ , \quad \mathcal{K}_{\inf}^{\mathcal{F}}(\nu_k, \mu^\star) = \inf_{G\in\mathcal{F}} \{\mathrm{KL}(\nu_k, G) : \mathbb{E}_G(X) > \mu^\star\} \ . \tag{6.2}$$

A bandit algorithm is then called *asymptotically optimal* for a family of distributions $\mathcal{F}$ when its regret matches this lower bound. In Chapter 1 we showed that such algorithm exists for instance if $\mathcal{F}$ is a *Single-Parameter Exponential Family* (SPEF) (Cappé et al., 2013; Kaufmann et al., 2012), for which $\mathcal{K}_{\inf}^{\mathcal{F}}$ is simply the Kullback-Leibler divergence between the distribution of mean $\mu_k$ and that of mean $\mu^\star$ in $\mathcal{F}$; and for bounded distributions with a known upper bound (Cappé et al., 2013; Honda and Takemura, 2015; Riou and Honda, 2020). In this chapter we will use the notation $\mathcal{K}_{\inf}^{B}$ for this family for an upper bound $B$.

**Motivations** While many algorithms achieve optimal regret for bounded distributions with the sole knowledge of the upper bound, the algorithms tackling unbounded distributions (e.g SPEF, sub-Gaussian, sub-exponential) generally assume a known parametric model for the tails. While such assumption entails convenient properties on the theoretical side, the practitioner may have some difficulty to determine which setting/parameters correspond to her problem. Furthermore, this uncertainty raises the question of robustness with respect to these hypotheses. Several works have considered this question: Hadiji and Stoltz (2020) show that adapting to an unknown bounded range requires a tradeoff between instance-dependent and worst-case regret, and recently (Agrawal et al., 2021b; Ashutosh et al., 2021) proved the impossibility of an instance-dependent logarithmic regret for light-tailed distributions without further assumptions on the tail parameters. The root cause for this is the lack of compactness of such families $\mathcal{F}$, which allows mass to "leak" at infinity so that confusing distributions with mean $\mu^*$ exist arbitrarily close to $\nu_k$, meaning $\mathcal{K}_{\inf}^{\mathcal{F}}(\nu_k, \mu^*) = 0$. Ashutosh et al. (2021) also

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

introduce a robust variant of UCB, that trades off logarithmic regret for $\mathcal{O}\left(f(T)\log(T)\right)$, where $f$ essentially tracks the possible mass leakage at infinity. These results puts into question the usual hypotheses under which bandit algorithms are designed: considering a *parametric* control of the tails is indeed sensitive to model mis-specification, but on the other hand the examples chosen to prove infeasability results seem a bit extreme for the practitioner.

In the first part of this thesis we proposed a family of algorithms based on *sub-sampling*, that can achieve strong theoretical guarantees under non-parametric assumptions on the arms (see Assumption 2.20). The kind of structure it requires on the arms, while being more flexible than the SPEF assumption, can still be somehow difficult to verify in some case-studies, e.g for the distributions we presented in Figure 3. Hence, in this chapter we propose a complementary approach, that is inspired by an optimal algorithm for bounded distributions, *Non-Parametric Thompson Sampling* (NPTS), that we already studied in Chapter 5 in the context of CVaR bandits. Building on NPTS, we propose in this chapter simple alternative setups allowing unspecified tail shapes but avoiding "mass leakage" to infinity, for instance with mild conditions linking the quantiles and the means of the distributions. As in the rest of this manuscript we consider the case of *light-tailed* distributions (that we define below). This problem is already non-trivial, so we let possible extensions for heavy-tail distributions for future work (e.g with tools like median-of-means, see (Bubeck et al., 2013)).

**Definition 6.1** (Light-tailed distribution). *We say that a distribution $\nu$ is light-tailed if there exists $\lambda_0 > 0$ such that $\forall \lambda : |\lambda| \leq \lambda_0$ it holds that*

$$\mathbb{E}_{X\sim\nu}\left[e^{\lambda X}\right] < +\infty \ .$$

**Outline** Following the central question that we consider in this thesis, we want to deign algorithms that require as little knowledge on the tails of distributions as possible. To this extent, NPTS (Riou and Honda, 2020) is a good candidate, considering its simple scheme that is sufficient to reach asymptotic optimality for bounded distributions with known bounds, and that in particular does not explicitly rely on the computation of the $\mathcal{K}_{\inf}^{B}$ function contrarily to other asymptotically optimal algorithms. Furthermore, the flexibility of this algorithm has been demonstrated in Chapter 5 with its adaptation in a risk-aware setting. We provide an extension of the principle of NPTS that we call *Dirichlet Sampling* (DS): we combine the core elements of NPTS and the duel-based framework that we already used in the first part of this thesis for the SDA algorithms and that is inspired by (Chan, 2020). This framework allows to introduce data-dependent exploration bonuses using the history of two arms. We present the resulting algorithm and detail the technical motivations of this approach in Section 6.2. We then introduce in Section 6.3 a first decomposition of the regret of DS algorithms under

general assumptions, and the technical results that allow to fine-tune the algorithm for different families (see Section 6.3.1). Then, we detail three instances of DS algorithms and their regret guarantees in Section 6.4: *Bounded Dirichlet Sampling* (BDS) tackles bounded distributions with possibly unknown upper bounds, *Quantile Dirichlet Sampling* proposes a first generalization to the unbounded case using truncated distributions. Last, *Robust Dirichlet Sampling* (RDS) has a slightly larger than logarithmic regret for any unspecified *light-tailed* unbounded distributions, making it a competitor to the Robust-UCB algorithm of (Ashutosh et al., 2021). Finally, we consider in Section 6.6 the use-case in agriculture introduced in Chapter 5 using the DSSAT simulator (see Hoogenboom et al. (2019)), which naturally faces all the questions (robustness, model specification) that motivate this work and shows the merit of DS over state-of-the-art methods for this problem.

## 6.2 Dirichlet Sampling Algorithms

In this section we introduce Dirichlet Sampling, a strategy extending the Non-Parametric Thompson Sampling algorithm of Riou and Honda (2020) outside the scope of bounded distributions with a known support upper bound. For this purpose, we build an adaptive strategy in a duel-based framework, already used in the sub-sampling based algorithms we introduced in Part I. We motivate this choice in the following.

**Background** We introduced NPTS in Chapter 1 (see Algorithm 1.6), and recall that it is identical to the B-CVTS algorithm introduced in the previous chapter when the risk parameter considered is $\alpha = 1$. The simplicity and strong theoretical guarantees of this algorithm are appealing for further generalization. As we fully depart from the Bayesian approach, considering alternative exploration bonuses, we derive a new family of algorithms under the name of *Dirichlet Sampling*. We keep the two principles of re-weighting the observations using a Dirichlet distribution and helping exploration by adding a bonus to the collected data, and explore how to apply them to more general (e.g unbounded) distributions. In particular, we allow in DS some pre-processing of the observations before re-weighting (see section 6.3.1 and 6.4) and motivate in Section 6.3.1 the use of a *data-dependent* bonus, that use information from several arms. The complexity introduced by such bonus in the analysis requires a change of algorithm structure, dropping the index policy for a *leader vs challenger* approach (Chan, 2020).

**Round-based algorithm** We define a round as a step of the algorithm at the end of which a set of (possibly several) arms are selected to be pulled. Let $\mathcal{A}_r \subset \{1, \dots, K\}$ be the subset of the arms pulled at the beginning of a round $r$, as in Chapters 2-3 we call $T$-round regret the

quantity

$$\bar{\mathcal{R}}_T = \mathbb{E}\left[\sum_{r=1}^{T}\sum_{k=1}^{K}\Delta_k \mathbb{1}(k \in \mathcal{A}_r)\right] = \sum_{k=1}^{K}\Delta_k \mathbb{E}[N_k(T)]\,, \tag{6.3}$$

where we slightly change the definition of $N_k$ (compared to the one in 6.1) to $N_k(T) = \sum_{r=1}^{T}\mathbb{1}(k \in \mathcal{A}_r)$. We recall that the $T$-round regret is an upper bound of the regret after $T$ pulls. At the beginning of each round we define a reference arm (leader), and then organize pairwise comparisons called *duels* between this arm and the other arms (challengers). The leader is chosen as the arm with largest sample size, that is $\ell(r) \in \text{argmax}_{k\in\{1,\dots,K\}} N_k(r)$.We choose to break ties first in favor of the best empirical arm, then with a random choice. A major motivation for this choice is that the leader will have a sample size that is *linear* in the number of rounds, as at least one arm is chosen at each round. This ensures strong statistical properties that we will exploit to design the exploration bonus of DS strategies. For that reason, randomizing the index of the leader is also unnecessary: it competes against each challenger with its *empirical mean*. We suggest to do the same with all the arms $k$ satisfying $N_k(r) = N_\ell(r)$. These choices have a practical interest as they avoid the computation time of drawing the largest weight vectors. We believe this can be an alternative of independent interest to computationally intensive index policies. For instance, most of the computational cost of NPTS (Riou and Honda, 2020) comes from drawing the random weights for the arm we define as the leader. Finally, we can remark that this use of the duel-based structure differs from SSMC (Chan, 2020) and SDA (see Part I of this thesis) since these algorithms use the empirical means of the challengers and sub-samples from the leader in the duels.

**Challenger's index** We define an index that is not dependent on the round, but only on the history of the challenger and the leader available at this round, that we denote respectively by $\mathcal{X} = (X_1, \dots, X_n)$, $\mathcal{Y} = (Y_1, \dots, Y_N)$ for simplicity of notation. We denote by $\bar{X}_n$ and $\bar{Y}_N$ their respective averages. We propose a duels in two steps, with a first comparison of the means and then a comparison using a Dirichlet re-sampled mean for the challenger. This is summarized in Algorithm 6.1 below.

**Input:** History $\mathcal{X} = (X_1, \dots, X_n)$ of the challenger, history $\mathcal{Y} = (Y_1, \dots, Y_N)$ of the leader
Draw a Dirichlet Sampling (DS) mean $\mu(\mathcal{X}, \mathcal{Y})$ for the challenger.
**if** $\max\{\bar{X}_n, \mu(\mathcal{X}, \mathcal{Y})\} \geq \bar{Y}_N$ **then**
  Challenger wins ; ▷ `Two chances to win: with empirical mean and DS mean`
**end**
Otherwise, leader wins.

**Algorithm 6.1:** Generic Dirichlet Sampling duel step

In Dirichlet Sampling, these duels take place inside a round-based framework that we summarize in Algorithm 6.2. We write it for a generic "Dirichlet Sampled mean" $\mu$ that must be computed by a re-weighting of the observations augmented by an exploration bonus. As in NPTS, the weights are drawn with a Dirichlet distribution. For instance, we propose a standard way to define a Dirichlet Sampling mean with a data-dependent (instead of constant) bonus $\mathcal{B}(\mathcal{X}, \mathcal{Y})$.

**Example 6.2.** *Consider a bonus $\mathcal{B}(\mathcal{X}, \mathcal{Y})$ and weights $(w_1, \dots, w_{n+1}) \sim \mathcal{D}_{n+1}$, the following expression is a possible re-sampled mean with Dirichlet Sampling,*

$$\mu(\mathcal{X}, \mathcal{Y}) = \sum_{i=1}^{n} w_i X_i + w_{n+1} \mathcal{B}(\mathcal{X}, \mathcal{Y}) \ .$$

However, the algorithm structure in Algorithm 6.2 could be combined with any randomized index, which is of independent interest as we will see in Section 6.3. In the next section we study the theoretical properties of Dirichlet Sampling, and discuss the choice of the index $\mu$ for different families of distributions.

**Input:** $K$ arms, horizon $T$
**Init.:** $t = 1$, $r = 1$, $\forall k \in \{1, ..., K\}$: $\mathcal{Y}_k = \{Y_1^k\}$, $N_k = 1$; ▷ Draw each arm once
**while** $t < T$ **do**
  $\mathcal{A} = \{\}$ ; ▷ Arm(s) to pull at the end of the round
  $\ell = \text{Leader}((\mathcal{Y}_1, N_1), \dots, (\mathcal{Y}_k, N_k))$ ; ▷ Choose a Leader
  **for** $k \in \{1, \dots, K\} : N_k < N_\ell$ **do**
    **if** *$k$ wins the duel* **then**
      $\mathcal{A} = \mathcal{A} \cup \{k\}$ ; ▷ Play the duels
    **end**
  **end**
  Draw arms from $|\mathcal{A}|$ if $\mathcal{A}$ is non-empty, else draw arm $\ell$.
  Update $t, r, (N_k)_{k \in \{1,\dots,K\}}, (\mathcal{Y}_k)_{k \in \{1,\dots,K\}}$. ; ▷ Collect Reward(s) and update data
**end**

**Algorithm 6.2:** Generic round-based strategy

## 6.3 Regret Analysis and Technical Results

In this section, we analyze the regret of DS algorithms. We first derive a general regret decomposition for the generic round-based strategy we described in Algorithm 6.2, that holds only

thanks to the duel-based structure and a standard assumption on the concentration of the arms' means. We then introduce several properties of Dirichlet sampling, that theoretically guide proper tuning of the exploration bonus used in DS. We finally instantiate the algorithm for three different problems and provide regret bounds in these settings. Starting with the regret decomposition, we exhibit general conditions to ensure guarantees that are independent of the index used. Allowing a different family of distribution $\mathcal{F}_k$ for each arm $k$, the first one concerns the concentration of the mean of each distribution.

**Assumption 6.3** (Concentration of means)**.** *For all $\nu_k \in \mathcal{F}_k$, there exists a good rate function $I_k$ satisfying $I_k(x) > 0$ for $x \neq \mu_k$ and for all $x > \mu_k$, $y < \mu_k$, and any i.i.d sequence $Y_1, \dots, Y_n$ drawn from $\nu_k$*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \geq x\right) \leq e^{-nI_k(x)} \ , \text{ and } \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \leq y\right) \leq e^{-nI_k(y)} \ . \tag{6.4}$$

This hypothesis is standard in the bandit literature, and is for instance satisfied by any *light-tailed* distributions (see Definition 6.1). We refer to (Dembo and Zeitouni, 2010) for techniques to derive the rate function of a distribution. We recall that we already made this assumption in the first part of this thesis (Assumption 2.4).

We now provide an upper bound on the round-regret presented in Section 6.2 for Algorithm 6.2. To simplify the notation we consider that there is only one optimal arm and, without loss of generality, that $\forall k > 1, \mu_k < \mu_1$. Furthermore, for simplicity we write the following theorem for an index $\mu(\mathcal{Y}, \widehat{\mu})$, that depends on $\mathcal{Y}$ through the empirical mean of the history, denoted by $\widehat{\mu}$.

**Theorem 6.4** (Generic regret decomposition of DS)**.** *Consider a bandit model $\nu = (\nu_1, \dots, \nu_K)$, where all distributions in $\nu$ satisfy (C1). Then for any Dirichlet sampled mean depending only on the history of the leader through its empirical mean, the expected number of pulls of each arm $k \in \{2, \dots, K\}$ under the round-based strategy of Algorithm 6.2 is upper bounded for each $\varepsilon \in [0, \Delta_k)$ by*

$$\mathbb{E}\left[N_k(T)\right] \leq n_k(T) + B^k_{T,\varepsilon} + C_{\nu,\varepsilon} \ ,$$

*where $C_{\nu,\varepsilon}$ is independent on $T$ and,*

$$n_k(T) = \mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, \ell(r) = 1)\right] \ ,$$

*and denoting $\mathcal{Y}_n$ the set of $n$ first observations of arm 1,*

$$B^k_{T,\varepsilon} = \sum_{k'=2}^{K} \sum_{n=1}^{\lceil 2\log(T)/I_1(\mu_k+\varepsilon)\rceil} \sup_{\widehat{\mu}\in[\mu_{k'}-\varepsilon,\mu_{k'}+\varepsilon]} \mathbb{E}\left[\frac{\mathbb{1}\left(\mu(\mathcal{Y}_n)\leq\widehat{\mu}\right)}{\mathbb{P}(\mu(\mathcal{Y}_n,\mu)\geq\widehat{\mu})}\right] .$$

The proof follows the general outline of Chan (2020) that we already used in the analysis of LB-SDA in Chapter 3, and details all the components of $C^k_{\nu,\varepsilon}$. This term is related to deviations of sample means for arm $k$ and arm 1 and is typically bounded by a (problem-dependent) constant under light-tail concentration (Assumption 6.3) so it does not depend on $\mu$ but only on the rate functions and the means of each arm. The other two terms of the upper bound reflect the exploration strategy. $n_k(T)$ is the expected number of pulls of arm $k$ when the best arm is the leader; we interpret it as the sample size required to statistically separate both arms at horizon $T$. On the other hand, $B^k_{T,\varepsilon}$ measures the capacity of the best arm to recover from a bad (small-sized) sample.

We now prove the theorem, but skip some details that were provided in Chapter 3 for the analysis of LB-SDA to avoid redundancy.

*Proof.* Thanks to the duel structure of DS, the fact that an arm is pulled or not depends of its status as a leader or a challenger. Furthermore, a challenger can be pulled only if it wins its duel against the leader. Consider an arm $k \in \{2,\dots,K\}$, we first write

$$\mathbb{E}[N_k(T)] \leq \underbrace{\mathbb{E}\left[\sum_{r=1}^{T-1}\mathbb{1}(k\in\mathcal{A}_{r+1},\ell(r)=1)\right]}_{n_k(T)} + 1 + \underbrace{\mathbb{E}\left[\sum_{r=1}^{T-1}\mathbb{1}(\ell(r)\neq 1)\right]}_{E_T} .$$

We already extracted the first term $n_k(T)$ of Theorem 6.4, and further work on the term $E_T$ corresponding to upper bounding the expected number of rounds with a sub-optimal leader. As in Chapter 3 we consider two alternatives, defining the sequence $a_r = \lceil r/4\rceil$ for $r\in\mathbb{N}$ and the event

$$\mathcal{D}_r = \{\exists u\in[a_r,r] : \ell(u)=1\} .$$

We recall that for any round $s \geq a_r$ the number of pulls of the leader is larger than $b_r \coloneqq \lceil a_r/K\rceil$, and consider the event $\ell(r)\cap\mathcal{D}_r$, that can only happen in case of a *leadership takeover*: a sub-optimal arm has the same sample size as arm 1 (larger than $b_r$) and a better empirical average. Summarizing this, starting the sum at any round $r_0$ we have

$$\sum_{r=1}^{T-1} \mathbb{P}\left(\ell(r)=1, \mathcal{D}_r\right) \leq r_0 + \sum_{r=1}^{T-1}\sum_{k'=2}^{K}\sum_{u=a_r}^{r}\sum_{n=b_r}^{r} \mathbb{E}\left[\mathbb{1}\left(N_{k'}(u)=N_1(u)=n, \bar{Y}_{k',n} \geq \bar{Y}_{1,n}\right)\right] .$$

We classically use that for any $x_{k'} \in \mathbb{R}$, $\{\bar{Y}_{k',n} \geq \bar{Y}_{1,n}\} \subset \{\bar{Y}_{k',n} \geq x_{k'} \cup \bar{Y}_{1,n} \leq x_{k'}\}$, and thanks to Assumption 6.3 we finally obtain

$$\sum_{r=1}^{T-1} \mathbb{P}\left(\ell(r)=1, \mathcal{D}_r\right) \leq r_0 + \sum_{k'=2}^{K}\sum_{r=1}^{T-1} r^2\left(e^{-b_r I_1(x_{k'})} + e^{-b_r I_k(x_{k'})}\right) = \mathcal{O}(1) .$$

This term is hence part of the constant $C_{\nu,\varepsilon}$. We now consider $\sum_{r=1}^{T-1}\mathbb{P}(\ell(r)=1,\bar{\mathcal{D}}_r)$. As in Chapter 3, we can upper bound these events by the number of total duels lost by arm 1 against a sub-optimal leader, using that

$$\sum_{r=1}^{T-1}\mathbb{P}(\ell(r)=1,\bar{\mathcal{D}}_r) \leq 9 \times \mathbb{E}\left[\sum_{r=1}^{T-1}\sum_{k'=2}^{K}\mathbb{1}(\ell(r)=k',\mathcal{C}_r^{k'})\right] ,$$

where $\mathcal{C}_r^{k'}$ is the event corresponding to arm 1 losing a duel against the leader $k'$. We can now fix any sub-optimal leader $k'$ and upper bound the term $\mathcal{C}_r^{k'}$. We recall that arm 1 has two chances to win the duel: first with its empirical mean, and then with a Dirichlet sampled mean. Furthermore, the bad outcome can come from either bad estimation of arm $k'$ or of arm 1. For any $\varepsilon > 0$, it holds that

$$\begin{aligned}\mathcal{C}_r^{k'} \subset & \left\{\left|\bar{Y}_{k',N_{k'}(r)} - \mu_{k'}\right| \geq \varepsilon, \ell(r)=k'\right\} \\ & \cup \left\{\left|\bar{Y}_{k',N_{k'}(r)} - \mu_{k'}\right| \leq \varepsilon, \ell(r)=k', \bar{Y}_{1,N_1(r)} \leq \mu_{k'}+\varepsilon, \mu\left(\mathcal{Y}_{N_1(r)}, \bar{Y}_{k',N_{k'}(r)}\right) \leq \bar{Y}_{k',N_{k'}(r)}\right\} .\end{aligned}$$

Thanks to the concentration of the leader, the expected number of pulls caused by the first term can be upper bounded by

$$\begin{aligned}\sum_{r=1}^{T-1}\mathbb{P}\left(\left|\bar{Y}_{k',N_{k'}(r)} - \mu_{k'}\right| \geq \varepsilon, \ell(r)=k'\right) &= \sum_{r=1}^{T-1}\sum_{n=\lceil r/K\rceil}^{r}\mathbb{P}\left(\left|\bar{Y}_{k',n} - \mu_{k'}\right|, \geq \varepsilon\right) \\ &= \mathcal{O}(1).\end{aligned}$$

For the simplicity of notation we keep the notation $\mathcal{C}_r^{k'}$ to define the remaining term. We then continue the analysis of $\mathcal{C}_r^{k'}$ by considering whether $N_1(r) \geq n_1(T)$ or not, for some new function $n_1(T)$. The idea is to choose $n_1(T)$ such that $\mathcal{C}_r^{k'}$ is unlikely for $n \geq n_1(T)$ thanks to the

first step of the duel with empirical means. Writing $C_{k'} \coloneqq \sum_{r=1}^{T-1} \mathbb{P}\left(\mathcal{C}_r^{k'}, \ell(r) = k'\right)$ we obtain

$$\begin{aligned} C_{k'} &\leq \sum_{r=1}^{T-1} \mathbb{P}\left(\mathcal{C}_r^{k'}, N_1(r) \geq n_1(T), \ell(r) = k'\right) + \sum_{r=1}^{T-1} \mathbb{P}\left(\mathcal{C}_r^{k'}, N_1(r) \leq n_1(T), \ell(r) = k'\right) \\ &\leq \sum_{r=1}^{T-1} \sum_{n=n_1(T)}^{T-1} \mathbb{P}(\bar{Y}_{1,n} \leq \mu_{k'} + \varepsilon) + \sum_{r=1}^{T-1} \mathbb{P}\left(\mathcal{C}_r^{k'}, N_1(r) \leq n_1(T), \ell(r) = k'\right) \\ &\leq \sum_{r=1}^{T-1} \mathbb{P}\left(\mathcal{C}_r^{k'}, N_1(r) \leq n_1(T), \ell(r) = k'\right) + \mathcal{O}(1) \ , \end{aligned}$$

if $n_1(T) \geq \frac{\log(T)}{I_1(\mu_k' + \varepsilon)}$. We then consider the remaining term, that we denote by

$$\mathcal{H}_{k'}^{r,n} = \left\{ N_1(r) = n, \left| \bar{Y}_{k',N_{k'}(r)} - \mu_{k'} \right| \leq \varepsilon, \bar{Y}_{1,n} \leq \bar{Y}_{k',N_{k'}(r)}, \mu\left(\mathcal{Y}_n, \bar{Y}_{k',N_{k'}(r)}\right) \leq \bar{Y}_{k',N_{k'}(r)} \right\} \ ,$$

and use it to write that

$$\sum_{r=1}^{T-1} \mathbb{P}\left(\mathcal{C}_r^{k'}, N_1(r) \leq n_1(T), \ell(r) = k'\right) \leq \sum_{r=1}^{T-1} \sum_{n=1}^{n_1(T)} \mathbb{P}\left(\mathcal{H}_{k'}^{r,n}\right) \ .$$

Following for instance (Riou and Honda, 2020) we can further state that

$$\sum_{r=1}^{T-1} \mathbb{1}(\mathcal{H}_{k'}^{r,n}) = \sum_{m=1}^{T-1} \mathbb{1}\left( \sum_{r=a_{r_0}}^{T-1} \mathbb{1}(\mathcal{H}_{k'}^{r,n}) \geq m \right) \ ,$$

and define as $\tau_1^n, \ldots, \tau_m^n$ the $m$ first rounds for which $\mathcal{H}_{k'}^{r,n}$ hold. If $\mathbb{1}\left(\sum_{r=a_{r_0}}^{T-1} \mathbb{1}(\mathcal{H}_{k'}^{r,n}) \geq m\right)$ is true then $\mathcal{H}_j^{\tau_j,n}$ holds for any $i \leq m$ and all these $\tau_i$ are finite, which provides

$$\mathbb{1}\left( \sum_{r=a_{r_0}}^{T-1} \mathbb{1}(\mathcal{H}_{k'}^{r,n}) \geq m \right) \leq \prod_{i=1}^{m} \mathbb{1}\left(\mathcal{H}_{k'}^{\tau_i^n,n}\right) \ .$$

The remaining term can be upper bounded as

$$\begin{aligned} D_{T,\varepsilon}^{k'} &\coloneqq \sum_{n=1}^{n_1(T)} \sum_{m=1}^{T-1} \mathbb{E}\left[ \prod_{i=1}^{m} \mathbb{1}\left(\mathcal{H}_{k'}^{\tau_i^n,n}\right) \right] \\ &= \sum_{n=1}^{n_1(T)} \sum_{m=1}^{T-1} \mathbb{E}_{\mathcal{Y}_n}\left[ \prod_{i=1}^{m} \mathbb{P}\left(\mu\left(\mathcal{Y}_n, \bar{Y}_{k',N_{k'}(\tau_i^n)}\right) \leq \bar{Y}_{k',N_{k'}(\tau_i^n)} \middle| \mathcal{Y}_n\right) \mathbb{1}\left(\mathcal{H}_{k'}^{\tau_i^n,n}\right) \right] \ . \end{aligned}$$

We remove the dependency of the index in $\mathcal{Y}_{N_{k'}(\tau_i^n)}^k$, knowing that it only depends of its mean that is located in a small range around $\mu_j$. We finally obtain

$$D_{T,\varepsilon}^{k'} \leq \sum_{n=1}^{n_1(T)} \sum_{m=1}^{T-1} \sup_{\widehat{\mu}\in[\mu_{k'}-\varepsilon,\mu_{k'}+\varepsilon]} \mathbb{E}_{\mathcal{Y}_n}\left[\mathbb{P}\left(\mu\left(Y_{1,n},\widehat{\mu}\right)\leq\widehat{\mu}\right)^m \mathbb{1}\left(\bar{Y}_{1,n}\leq\widehat{\mu}\right)\right]$$
$$\leq \sum_{n=1}^{n_1(T)} \sup_{\widehat{\mu}\in[\mu_{k'}-\varepsilon,\mu_{k'}+\varepsilon]} \mathbb{E}_{\mathcal{Y}_n}\left[\frac{\mathbb{P}\left(\mu\left(\mathcal{Y}_n,\widehat{\mu}\right)\leq\widehat{\mu}\right)}{\mathbb{P}\left(\mu\left(\mathcal{Y}_n,\widehat{\mu}\right)\geq\widehat{\mu}\right)}\mathbb{1}\left(\bar{Y}_{1,n}\leq\widehat{\mu}\right)\right] .$$

This concludes the proof if we define $B_{T,\varepsilon}^k = \sum_{k'=2}^K D_{T,\varepsilon}^{k'}$ in Theorem 6.4.

□

Theorem 6.4 is formulated to be as general as possible and can be regarded as a counterpart of existing results for the analysis of other randomized strategy, such as Theorem 1 of Kveton et al. (2019b) for *General Randomized Exploration* or Theorem 36.2 in (Lattimore and Szepesvári, 2020) for Thompson Sampling.

**Remark 6.5** (Further generalizations). *Theorem 6.4 still holds under more general assumptions. First, we could replace the empirical averages by any* robust estimator *of the means in the first comparison step, and require Assumption 6.3 to hold for this estimator instead (e.g with median-of-means or truncated means for heavy-tailed bandits). Then, the exploration bonus could use any statistics on the leader's history that have concentration properties similar to Assumption 6.3 with slight adaptations of the proof. This can include quantiles (that we can concentrate e.g thanks to DKW inequality, see (Massart, 1990)), or moments of higher order under some assumptions on the distributions.*

*Finally, Assumption 6.3 could actually be relaxed and require inequalities in $o(n^{-3})$ (instead of exponential decay in $n$) to obtain the same final result (but different constant terms) with our proof scheme.*

We will later analyze instances of Dirichlet Sampling where the first-order term of the regret is driven entirely by $n_k(T)$. We therefore introduce the following assumptions to control the contribution of $B_{T,\varepsilon}^k$ to the regret.

**Assumption 6.6** (Sufficient exploration of arm 1). *For any $\varepsilon > 0$, and any $n_1(T) = \mathcal{O}(\log T)$ it holds that*

$$\sum_{n=1}^{n_1(T)} \mathbb{E}_{\mathcal{Y}_n\sim\nu_1^n}\left[\frac{\mathbb{1}(\bar{Y}_{1,n}\leq\mu_1-\varepsilon)}{\mathbb{P}_{w\sim\mathcal{D}_{n+1}}\left(\mu(\mathcal{Y}_n,\mu_1-\varepsilon)\geq\mu_1-\varepsilon\right)}\right] = o(\log T) .$$

The LHS represents the expected cost in terms of regret of underestimating the optimal arm; intuitively, it measures the expected number of lost duels before finally winning one when starting with low rewards. This is a classic decomposition in bandit analysis, and a counterpart of Assumption 6.6 holds for most index policies with provable regret guarantees, e.g Theorem 1 in Kveton et al. (2019b) (GIRO) or Lemma 4 in Agrawal and Goyal (2012a)) (Bernoulli Thompson Sampling). We find it noteworthy that this regret decomposition depends only on the distribution of the best arm and its randomized Dirichlet Sampling index when it is a challenger.

**Corollary 6.7** (Conditions for controlled regret)**.** *If Assumptions 6.3 and 6.6 hold for the DS index on the families of distribution $(\mathcal{F}_k)_{k\in\{1,\dots,K\}}$, the regret of the DS algorithm satisfies*

$$\mathcal{R}_T \leq \sum_{k=2}^{K} \Delta_k n_k(T) + o(\log T) \ .$$

Up to this point this result is quite abstract, but this standardized analysis allows us to instantiate the Dirichlet Sampling algorithm on different class of problems and calibrate it in order to ensure that Assumption 6.6 holds and to make $n_k(T)$ explicit. In particular if $n_k(T) = \mathcal{O}(\log T)$, we recover the logarithmic regret. In the next section, we present technical results to justify calibrations of the DS index for several kind of families.

### 6.3.1 Technical tools: boundary crossing probability of a DS index

In this section, we highlight some key properties of a sum of random variables re-weighted by a Dirichlet weight vector that help us suggest a sound tuning of the bonus $\mathcal{B}(\mathcal{Y}, \mathcal{Y})$ for different kind of families. We then detail such tuning. We first recall the definition of *Boundary Crossing Probabilities (BCP)*, that we already used in Chapter 5.

**Definition 6.8** (Boundary Crossing Probability (BCP))**.** *Consider a set of $n+1$ observation points $\mathcal{Y} = (Y_1, \dots, Y_{n+1}) \subset \mathbb{R}^{n+1}$. Then, for any $\mu \in \mathbb{R}$, a "Boundary Crossing Probability" (BCP) conditionally on $\mathcal{Y}$ is defined as*

$$[\text{BCP}] := \mathbb{P}_{w\sim\mathcal{D}_{n+1}}\left(\sum_{i=1}^{n+1} w_i Y_i \geq \mu\right) \ ,$$

*where we recall that $\mathcal{D}_{n+1}$ is the Dirichlet distribution with parameter $(1, \dots, 1)$ of size $n+1$, i.e the uniform distribution on the $(n+1)$-simplex. We emphasize that here $\mathcal{Y}$ is considered fixed, and the only source of randomness comes from the weights $w$.*

When all observations are *distinct* from each other this BCP has a closed formula, which has been derived for instance in (Cho and Cho, 2001) as

$$\mathbb{P}_{w\sim\mathcal{D}_{n+1}}\left(\sum_{i=1}^{n+1} w_i Y_i \geq \mu\right) = \sum_{i=1}^{n+1} \frac{(Y_i-\mu)_+^n}{\prod_{j=1,j\neq i}^{n+1}(Y_i-Y_j)} \,. \tag{6.5}$$

This expression is obtained by computing the volume of the half-space of the simplex defined by the hyperplane $\sum_{i=1}^{n+1} w_i Y_i \geq \mu$. Unfortunately, this formula is not very informative: for sorted data the terms are alternatively positive and negative, and can take large values (compensating each other). This makes the exact formula hardly tractable even for numerical simulations. We also add that the closed formula does not exist for a Dirichlet distribution with some parameters larger than 1.

This quantity is of much interest as both the growth of $n_k(T)$ and checking that Assumption 6.6 holds can be performed by respectively upper and lower bounding for the BCP. In Chapter 5 we provided such bounds for the CVaR case, that are still valid here because the standard setting is a special case of CVaR bandits with a risk level $\alpha = 1$. These results resort on usual properties of the Dirichlet distribution that were given in Appendix 5.7. The lower bounds suggest non-trivial tuning of the bonus. We first exhibit a necessary condition when the bonus is not allowed to depend on the set of observations $\mathcal{Y}$.

**Lemma 6.9** (Necessary condition with a data-independent bonus)**.** *Consider a fixed bonus $B_\mu$, and a distribution $F$ (with CDF also denoted $F$). Assumption 6.6 holds only if*

$$B_\mu > \mu + \frac{1}{1-F(\mu)}\mathbb{E}_{Y\sim F}\left[(\mu - Y)_+\right] \,.$$

*Proof.* When all the observations are below the threshold Equation (6.5) provides

$$\mathbb{P}_{w\sim\mathcal{D}_n}\left(\sum_{i=1}^{n} w_i Y_i + w_{n+1}B(\mu) \geq \mu\right) = \prod_{i=1}^{n}\frac{B-\mu}{B-Y_i} \,,$$

so plugging this term in Assumption 6.6 gives the expression

$$\mathbb{E}\left[\prod_{i=1}^{n}\left(\frac{B-Y_i}{B-\mu}\right)\mathbb{1}(Y_i\leq\mu)\right] = \mathbb{E}_{Y_1\sim F}\left[\left(\frac{B-Y_1}{B-\mu}\right)\mathbb{1}(Y_1\leq\mu)\right]^n \,.$$

The assumption can then hold only if the expectation is smaller than 1, which is equivalent to

$$(B-\mu)(1-F(\mu)) \geq \mathbb{E}\left[(\mu-Y)_+\right] \,,$$

which gives the result. □

This result is obtained using a "worst-case" scenario in which all the observations are below the threshold $\mu$. Hence, it does not cover all possible trajectories, yet it suggests to investigate the properties of bonuses with a similar form. Since the right-hand side of the inequality requires a knowledge on the arms distributions that we would like to avoid, we use an empirical estimator for the expectation. This suggests to introduce some parameter $\rho$ and data-dependent bonuses of the form

**Definition 6.10** (Canonical bonus)**.** *For a dataset $\mathcal{Y} = (Y_1, \dots, Y_n)$, a parameter $\rho$ and a threshold $\mu$ we define*

$$B(\mathcal{Y}, \mu, \rho) = \mu + \rho \times \frac{1}{n} \sum_{i=1}^{n} (\mu - Y_i)^+ \,. \tag{6.6}$$

We interpret $\rho$ as the **leverage** of the empirical excess gap $\frac{1}{n}\sum_{i=1}^{n}(\mu - Y_i)^+$ w.r.t the threshold $\mu$. We then tune $\rho$ according to the hypothesis we make on the arm distributions, which is much less constraining than assuming knowledge of the shape of the entire tail. In all DS algorithms we proposed (see next section), we use Equation (6.6) as the basis for defining the appropriate bonus. Finally, we provide in Lemma 6.11 a novel lower bound on the BCP that reveals that, in the general light-tailed unbounded case, without further processing of the data, DS cannot achieve a logarithmic regret when the empirical maximum of the observations tends to $+\infty$ at some rate $g(n)$.

**Lemma 6.11** (Lower bound for the BCP)**.** *Consider a set $\mathcal{Y} = (Y_1, \dots, Y_{n+1}) \in \mathbb{R}^{n+1}$ and any threshold $\mu$, and assume that $\max_{i \in \{1,\dots,n+1\}} Y_i \geq g(n) \geq \mu$ for some function $g$. Denoting by $\bar{\Delta}_n^+ = \frac{1}{n}\sum_{i=1}^{n+1}(\mu - Y_i)^+$ the empirical excess gap, it holds that*

$$\mathbb{P}_{w \sim \mathcal{D}_{n+1}} \left( \sum_{i=1}^{n+1} w_i Y_i \geq \mu \right) \geq \exp\left( -n \frac{\bar{\Delta}_n^+}{g(n) - \mu} \right) \,.$$

*Proof.* We obtain this lower bound by truncating all the observations that are larger than the threshold except the maximum, allowing to use Equation 6.5. Combining this property with $\log(1 + x) \leq x$ we obtain

$$\mathbb{P}_{w \sim \mathcal{D}_{n+1}} \left( \sum_{i=1}^{n+1} w_i Y_i \geq \mu \right) \geq \mathbb{P}_{w \sim \mathcal{D}_{n+1}} \left( \sum_{i=1}^{n} w_i \min(Y_i, \mu) + w_{n+1} \max_{j=1,\dots,n+1} Y_j \geq \mu \right)$$

$$
\begin{aligned}
&= \frac{(\max_{j=1,\dots,n+1} Y_j - \mu)^n}{\prod_{i=1}^n (\max_{j=1,\dots,n+1} Y_j - \min(Y_i, \mu))} \\
&= \exp\left(-\sum_{i=1}^n \log\left(\frac{\max_{j=1,\dots,n+1} Y_j - \min(Y_i, \mu)}{\max_{j=1,\dots,n+1} Y_j - \mu}\right)\right) \\
&= \exp\left(-\sum_{i=1}^n \log\left(1 + \frac{\mu - \min(Y_i, \mu)}{\max_{j=1,\dots,n+1} Y_j - \mu}\right)\right) \\
&\geq \exp\left(-\sum_{i=1}^n \frac{\mu - \min(Y_i, \mu)}{\max_{j=1,\dots,n+1} Y_j - \mu}\right) \\
&= \exp\left(-\sum_{i=1}^n \frac{(\mu - Y_i)_+}{\max_{j=1,\dots,n+1} Y_j - \mu}\right) \ ,
\end{aligned}
$$

which yields the result. □

In particular, we see in this expression that $g(n)$ may hinder the exponential rate in $n$. In the next section we discuss three examples of DS algorithms and their theoretical guarantees. Before that, we state a result that will be useful in the analysis of DS algorithms using the canonical exploration bonus.

**Corollary 6.12.** *Let $\mathcal{Y} = (Y_1, \dots, Y_n, Y_n)$ be a set and $Y_{n+1} = B(\mathcal{Y}, \mu, \rho)$ for some parameters $\rho, \mu$. Then, it holds that*

$$
\mathbb{P}_{w \sim \mathcal{D}_{n+1}}\left(\sum_{i=1}^{n+1} w_i Y_i \geq \mu\right) \geq \exp\left(-\frac{n}{\rho}\right) \ .
$$

The result is direct by replacing $g(n)$ by the expression of the bonus in Lemma 6.11.

## 6.4 From optimality to robustness: three instances of DS

Building on the results from the previous section, we now instantiate the DS algorithms for three bandit problems. We first prove that optimal guarantees can be derived for DS with bounded distributions under a non-standard definition of the problem (i.e unknown upper bound but alternative assumptions), motivated by practical considerations. Then, we consider a natural extension to unbounded distributions using a simple truncation mechanism, ensuring logarithmic regret under assumptions on some quantile of the distributions. Finally we consider a simple DS algorithm, securing slightly larger-than-logarithmic regret for the entire family of *light-tailed distributions*. In the following we denote by $B(\mathcal{X}, \mu, \rho)$ the bonus defined in Equation 6.6 that we call *canonical bonus* (Definition 6.10) for a set $\mathcal{X}$, a mean $\mu$ and some

parameter $\rho$. For simplicity we will keep a generic $\mu$ in our exposition, while its value is in practice the empirical mean of the leading arm. We will use this expression when detailing the value of the re-sampled means used in the algorithms that we propose.

A sketch of the proofs of the three theorems can be found in Section 6.5. They consist in deriving an expression for $n_k(T)$ and showing that Assumption 6.6 (sufficient exploration) holds for the algorithms in the settings they tackle.

### 6.4.1 Bounded Dirichlet Sampling (BDS): optimality for an alternative family of bounded distributions

Let $\mathcal{F}_{[b,B]}$ be the set of distributions supported in $[b, B]$, and consider a bandit $\nu = (\nu_1, \dots, \nu_K)$ with $\nu_k \sim \mathcal{F}_{[b_k,B_k]}$ for some $B_k \in \mathbb{R}$. If we assume that $B_k$ is known (case 1), then simply defining $B_k$ as the exploration bonus ensures an asymptotically optimal regret, with a direct adaptation of the proof of NPTS (Riou and Honda, 2020). However, the precise knowledge of the upper bound for each arm is sometimes inaccessible to the practitioner (e.g if the environment is new, or if no expert is available to provide a reasonable upper bound). We propose an alternative setting, with the family $\mathcal{F}_{\mathcal{B}}^{\gamma,p} = \{\exists B : \nu \in \mathcal{F}_{[b,B]}, \mathbb{P}_\nu([B-\gamma, B]) \geq p\} \subset \mathcal{F}_{[b,B]}$. $B_k$ is *unknown* but we assume it is *detectable* in the sense that we will observe a sample from its neighborhood $[B_k - \gamma, B_k]$ with a reasonable probability of at least $p$, with known $\gamma, p$. In this case we propose the following bonus, allowing to obtain theoretical results in this setting for some values of $\rho$ that we will precise later,

$$B(\mathcal{Y}, \widehat{\mu}) \coloneqq \max\{\mathcal{Y}^+ + \gamma, B(\mathcal{Y}, \widehat{\mu}, \rho)\}\,, \quad \text{where } \mathcal{Y}^+ = \max\{y : y \in \mathcal{Y}\}. \tag{6.7}$$

We summarize the re-sampled mean used by BDS in Algorithm 6.3.

**Input:** Set $\mathcal{Y} = (Y_1, \dots, Y_n)$, mean of the leader $\widehat{\mu}$, parameters $\gamma, \rho$
Draw $w = (w_1, \dots, w_{n+1}) \sim \mathcal{D}_{n+1}$
**return** $\sum_{i=1}^n w_i Y_i + w_{n+1} \max\left(\max_{i=1}^n Y_i + \gamma, B(\mathcal{Y}, \widehat{\mu}, \rho)\right)$

**Algorithm 6.3:** Bounded Dirichlet Sampling re-sampled mean

We then provide in Theorem 6.13 the theoretical guarantees obtained for BDS under this alternative "bounded distributions" assumption.

**Theorem 6.13** (Theoretical guarantees of BDS). *If* $\forall k \in \{2, \dots, K\}$ , $\nu_k \sim \mathcal{F}_{\mathcal{B}}^{\gamma,\rho}$, *choosing the exploration bonus of Equation 6.7 with* $\rho \geq -1/\log(1-p)$ *ensures that*

$$\mathbb{E}[N_k(T)] \leq \frac{\log(T)}{\mathcal{K}_{\inf}^{B_{\rho,\gamma}}(\nu_k, \mu_1)} + O(1) \ ,$$

*where* $B_{\rho,\gamma} = \max\left(B + \gamma, \mu_1 + \rho\mathbb{E}_{X\sim\nu_k}[(\mu_1 - X)_+]\right)$.

This setting is a first example of the interest of data-dependent bonuses. It makes sense in practice by avoiding for instance distributions with a small mass arbitrarily far from the rest of their support, which may not be likely in a real-world application. We now consider the unbounded case. Before that, we make a remark related to the B-CVTS algorithm of Chapter 5.

**Remark 6.14.** *The changes of the proof allowed by the round-based structure allows to remove the assumption that the distributions are continuous that was needed in the theoretical guarantees of B-CVTS, as the term* $A_2$ *in the proof of Theorem 5.4 is no more considered here thanks to the first comparison of sample means. We could thus analyze a duel-based variant of B-CVTS for arbitrary bounded distributions.*

### 6.4.2 Quantile Dirichlet Sampling (QDS): truncating the upper tail for logarithmic regret with unbounded distributions

Let us consider the family $\mathcal{F}_{[b,+\infty]}$ for some unknown $b \in \mathbb{R}$. A natural way to extend algorithms designed for $\mathcal{F}_{[b,B]}$ (where $B < +\infty$) is to truncate the upper tail of the distributions. We propose a simple way to do this, by considering (as a parameter of the algorithm) a quantile $\alpha$, denoted by $q_\alpha(\nu)$ for a distribution $\nu$, and a truncation operator $\mathcal{T}_\alpha$ that (1) do not change a distribution below its $\alpha$ quantile, and (2) "summarizes" its upper tail by its expectation, known as *Conditional Value at Risk* (CVaR)[2]. Formally, we obtain

$$\forall A \subset [b, q_\alpha(\nu)] : \ \mathcal{T}_\alpha(\nu)(A) = \nu(A) \ \ , \text{ and } \forall x > q_\alpha(\nu) : \ \mathcal{T}_\alpha(\nu)\left(\{x\}\right) = \alpha\mathbb{1}(x = C_\alpha(\nu)) \ ,$$

with $C_\alpha(\nu) = \mathbb{E}[X|X > q_\alpha(\nu)]$. We then propose *Quantile Dirichlet Sampling* (QDS), that computes the index of a challenger (say arm $k$, with observations $\mathcal{Y}_k$) during a duel as follow: (1) apply $\mathcal{T}_\alpha$ to the empirical distribution, (2) compute the bonus $B(\mathcal{Y}_k, \widehat{\mu}, \rho)$, and (3) *re-sample* the truncated empirical distribution with weights drawn according to $\mathrm{Dir}(1, \dots, 1, n_\alpha)$ where parameter $n_\alpha$ is for the weight used with the empirical CVaR, and is the number of observations used to compute it (to avoid a bias in the re-sampled mean). This procedure is summarized

[2]Defined on the upper tail, contrarily to Chapter 5 where we consider the lower tail

in Algorithm 6.4. If presenting this algorithm and its guarantees is rather technical, its implementation is in fact quite simple. Furthermore, the computation time of these steps can be optimized in practice (keeping in memory the sorted data, quantile and CVaR).

**Input:** Sorted data $\mathcal{Y} = (Y_1 \leq \cdots \leq Y_n)$, leader mean $\widehat{\mu}$, quantile $\alpha$, $\rho$
Set quantile index $n_\alpha = \lceil n\alpha \rceil / n$
Set $C_\alpha = \frac{1}{n-n_\alpha+1} \sum_{i=n_\alpha}^{n} Y_i$ ; ▷ `Compute the CVaR := average of largest data`
Draw $w = (w_1, \dots, w_{n_\alpha+1}) \sim \mathrm{Dir}((1, \dots, 1, \mathbf{n}_\alpha, 1))$ ; ▷ `Parameter 1 except for` $w_{n_\alpha}$
**return** $\sum_{i=1}^{n_\alpha-1} w_i Y_i + w_{n_\alpha} C_\alpha + w_{n_\alpha+1} B(\mathcal{Y}, \widehat{\mu}, \rho)$

**Algorithm 6.4:** Quantile Adaptive Dirichlet Sampling re-sampled mean

We can analyze this algorithm for the subset of distributions

$$\mathcal{F}^{\alpha}_{[b,+\infty)} = \{\nu \in \mathcal{F}_{[b,+\infty)} : \forall \mu > \mathbb{E}_\nu(X), \mathcal{K}_{\inf}^{\mathcal{F}_{[b,+\infty)}}(\nu, \mu) \geq \mathcal{K}_{\inf}^{\mathfrak{M}_k}(\mathcal{T}_\alpha(\nu), \mu)\} \ ,$$

where $\mathfrak{M}_k = \max\{q_\alpha(\nu_k), \mu_1 + \rho \mathbb{E}_{\nu_k}[(\mu_1 - X)^+]\}$, and the second $\mathcal{K}_{\inf}$ is taken on the family $\mathcal{F}_{[b,\mathfrak{M}_k]}$ (using previously introduced notations). Although technical, this condition essentially states that the bandit problem taken on the complete family $\mathcal{F}_{[b,+\infty)}$ is no harder than an alternative bandit problem considering the truncated distributions and a bounded family, with an upper bound depending on the $1-\alpha$ quantile and the leverage $\rho$ of the exploration bonus.

**Theorem 6.15** (Logarithmic Regret of QDS). *Consider a bandit model $\nu = (\nu_1, \dots, \nu_K)$ satisfying $\forall k$, $\nu_k \in \mathcal{F}^{\alpha}_{[b,+\infty)}$ for some $b > -\infty$ (lower-bounded support) and a known $\alpha > 0$. Then, for any $\varepsilon_0 > 0$ small enough QDS with any parameters $\alpha' < \alpha$ and $\rho \geq (1+\alpha')/\alpha'^2$ satisfies*

$$\mathbb{E}[N_k(T)] \leq \frac{\log T}{\mathcal{K}_{\inf}^{\mathfrak{M}_k^C}(\mathcal{T}_\alpha(\nu_k), \mu_1) - \varepsilon_0} + \mathcal{O}(1) \ ,$$

*with $\mathfrak{M}_k^C = \max\{C_\alpha(\nu_k), \mu_1 + \rho \mathbb{E}_\nu[(\mu_1 - X)^+]$, and $\mathcal{T}_\alpha$ is the truncation operator we defined.*

This result captures the continuum between bounded and light-tailed distributions. In our opinion, it sheds new light on the interpretation of infeasability results of e.g Ashutosh et al. (2021): logarithmic regret can be achieved *without specifying the tail with precise parameters*, but some simple realistic assumptions (e.g avoiding very small mass at a very large value, for instance after some quantile) can be enough to avoid pathological distributions that make little sense in practice.

**Remark 6.16.** *The restriction to the semi-bounded case $b > -\infty$ is due to our proof technique, based on a discretization of the support of the truncated distribution (see next section). Note that the actual value of $b$ is not known by the algorithm. This is intuitive since $\mathcal{K}_{\inf}^{\mathcal{F}_{-\infty,B}} = \mathcal{K}_{\inf}^{\mathcal{F}_{b,B}}$ for all $b, B \in \mathbb{R}$, as proved in Theorem 2 of (Honda and Takemura, 2015). Different theoretical tools could allow to prove a logarithmic regret for QDS in the doubly unbounded case, possibly with a symmetric treatment of the two tails. We leave this extension for future work.*

One may wonder whether the couple quantile condition/truncation is necessary to achieve theoretical results as well as good practical performance. Our last algorithm tackle this question.

### 6.4.3 Robust Dirichlet Sampling (RDS) for light-tailed distributions

We call *Robust Dirichlet Sampling* (RDS) the algorithm with bonus $B(\mathcal{Y}, \mu, \rho_n)$, where the leverage $\rho_n$ is a function of the sample size $n = |\mathcal{Y}|$. We prove that while being very simple, RDS achieves a robust sub-linear regret bound when each arm comes from **any** *unknown* light-tailed distribution, that we define following 6.1 as the family

$$\mathcal{F}_\ell = \{\nu \in \mathcal{F}_{(-\infty,+\infty)} : \exists \lambda_\nu > 0, \forall \lambda \in [-\lambda_\nu, \lambda_\nu], \mathbb{E}_\nu[\exp(\lambda X)] < +\infty\} \ .$$

RDS is the simplest instance of Dirichlet Sampling that we propose, as can be seen in Algorithm 6.5, and is also the one with the less restrictive family of distributions.

**Input:** Data $\mathcal{Y} = (Y_1, \dots, Y_n)$, mean $\mu$, $(\rho_n)_{n\in\mathbb{N}}$
Draw $w = (w_1, \dots, w_{n+1}) \sim \mathcal{D}_{n+1}$
**return** $\sum_{i=1}^n w_i Y_i + w_{n+1} B(\mathcal{Y}, \mu, \rho_{\mathbf{n}})$

**Algorithm 6.5:** Robust Dirichlet Sampling re-sampled mean

The regret bound of RDS only depends on the choice of an increasing sequence $\rho_n$ satisfying $\rho_n \to +\infty$ and $\rho_n = o(\sqrt{n})$, and Theorem 6.17 shows that RDS attain slightly larger than logarithmic regret under very general assumptions.

**Theorem 6.17** (Robust regret bound for RDS)**.** *Let $\nu = (\nu_1, \dots, \nu_K)$ a bandit model satisfying $\nu_k \in \mathcal{F}_\ell$ for all $k$. Consider **any** increasing sequence $(\rho_n)_{n\in\mathbb{N}}$ with $\rho_n \to +\infty$, $\rho_n = o(n)$. Then, for $T$ large enough the expected number of pull of any sub-optimal arm $k$ in RDS is upper bounded by*

$$\mathbb{E}[N_k(T)] \leq n_k^{\eta,\varepsilon_0}(T) + \mathcal{O}(1) \ ,$$

*where for any* $\eta \in (0,1], \varepsilon_0 > 0$, $n_k^{\eta,\varepsilon_0}(T)$ *is the sequence satisfying*

$$n_k^{\eta,\varepsilon_0}(T) = \frac{\log T}{\eta(\Delta_k - \varepsilon_0)}(M_{k,n_k^{\eta,\varepsilon_0}(T)} - \mu) \ ,$$

*with*

$$M_{k,n} = \max\left\{F_k^{-1}\left(\exp\left(-\frac{1}{n^2(\log n)^2}\right)\right), \rho_n\right\}.$$

*In particular, if* $\rho_n = \mathcal{O}(\log n)$ *then* $\mathbb{E}[N_k(T)] = \mathcal{O}(\log(T)\log\log(T))$ *for any light-tailed distribution* $\nu_k \in \mathcal{F}_\ell$.

The sequence $M_{k,n}$ is a high probability upper bound of the maximum of $n$ observations from $F_k$, that we discuss in next section. For light-tailed distributions, it holds that $M_{k,n} = \mathcal{O}(\log n)$ (using Jensen inequality as in the proof of Theorem 2.5 in (Boucheron et al., 2013)). Hence, choosing $\rho_n = \mathcal{O}(\log n)$ we can further obtain the simpler upper bound in $\mathcal{O}(\log(T)\log\log(T))$. This slightly larger-than-logarithmic rate is a consequence of Lemma 6.11. In our opinion this is a small cost compared to the generality of the guarantees of RDS. We call the algorithm *robust* because these theoretical guarantees are obtained on the broad class of light-tailed distributions, without any additional assumption. We recommend the leverage function $\rho_n = \mathcal{O}(\sqrt{\log(1+n)})$, which corresponds to the growth rate of the maximum of sub-Gaussian samples and is empirically validated (see Section 6.6). We emphasize that RDS thus avoids all hyperparameter tuning, a desirable feature for the practitioner with little information on the problem at hand. Furthermore, in the next section we show that this algorithm performs very well in practice despite its non-logarithmic asymptotic guarantees.

## 6.5 Proof Sketch

In this part we provide a general overview of the proofs of Theorem 6.13, 6.15 and 6.17. The full proofs can be found in Appendix D of (Baudry et al., 2021c). Here, we try to highlight the key ingredients that lead to the results provided in the previous section.

We recall that the three family of distributions are light-tailed and hence satisfy assumption 6.3. So, starting from Corollary 6.7 our objective is to derive the dominant term $n_k(T)$ and to prove that Assumption 6.6 is satisfied in the three examples we consider.

### 6.5.1 Deriving the dominant term: characterizing $n_k(T)$

In this section we consider an arm $k \in \{2, \dots, K\}$, of distribution $\nu_k$. We first derive a general proof sketch that will be used for each algorithm. The idea is simple, and inspired by the proof

for B-CVTS in Chapter 5: we want to find a *large probability confidence ball* for the empirical distribution of $\nu_k$ in which the BCP can be conveniently upper bounded. Lemma 6.18 summarizes this principle.

**Lemma 6.18.** *Assume that $\nu_k$ satisfies Assumption 6.3, and denote by $\mathcal{Y}_n^k = (Y_1^k, \dots, Y_n^k)$ a set of $n$ random variables drawn from $\nu_k$. For a given Dirichlet sampled mean $\mu$, assume that for any $n \in \mathbb{N}$ there exists a subset $\mathcal{B}_{k,n} \subset \mathbb{R}^n$ satisfying*

1. *There exists a strictly increasing function $f_k$ such that for any threshold $\widehat{\mu} < \mu_1$,*

$$\mathcal{Y}_n^k \subset \mathcal{B}_{k,n} \Rightarrow \mathbb{P}\left(\mu(\mathcal{Y}_n^k, \widehat{\mu}) \geq \widehat{\mu}\right) \leq \exp\left(-f_k(n, \mathcal{B}_{k,n}, \widehat{\mu})\right) \ .$$

2. $\sum_{n=1}^{T-1} \mathbb{P}\left(\mathcal{Y}_n^k \notin \mathcal{B}_{k,n}\right) = \mathcal{O}(1)$.

*If these two conditions hold, then one can obtain*

$$n_k(T) = m_k(T) + \mathcal{O}(1) \ ,$$

*where for any $\varepsilon > 0$, $m_k(T)$ is the sequence satisfying $f_k(n_k(T), \mathcal{B}_{k,n}, \mu_1 - \varepsilon) = \log T$.*

*Proof.* As in several proofs presented in previous chapter our initial upper bound is

$$n_k(T) \leq m_k(T) + \mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, \ell(r) = 1, N_k(r) \geq m_k(T))\right] \ ,$$

for any sequence $m_k(T)$. We then consider the remaining term according to the "good" event

$$\mathcal{G}_k^r = \left\{\mathcal{Y}_{N_k(r)}^k \in \mathcal{B}_{k,N_k(r)}\right\} \cap \left\{\ \bar{Y}_{1,N_1(r)} \leq \mu_1 - \varepsilon_1\right\} \ ,$$

where $\varepsilon_1 > 0$. Without detailing the steps (that are similar to those for B-CVTS), we claim that under the first assumption of Lemma 6.18 this event contribute to the upper bound of $n_k(T)$ with

$$C_{\mathcal{G}} = T \exp\left(-f_k(m_k(T), \mathcal{B}_{k,m_k(T)}, \mu_1 - \varepsilon_1)\right) \ .$$

Then, if $\mathcal{G}$ is not true one of the two events it contains does not hold. The first additional term (empirical distribution of $k$ not in the ball) is

$$C_{\bar{\mathcal{G}},1} \coloneqq \sum_{r=1}^{T-1} \sum_{n=m_k(T)}^{T-1} \mathbb{E}\left[\mathbb{1}(k \in \mathcal{A}_{r+1}, N_k(r) = n, \mathcal{Y}_{N_k(r)}^k \notin \mathcal{B}_{k,N_k(r)})\right]$$

$$\leq \sum_{n=m_k(T)}^{T-1} \mathbb{P}\left(\mathcal{Y}_n^k \notin \mathcal{B}_{k,n}\right) ,$$

and the second term (empirical mean of the leader over-estimated) can be upper bounded by $C_{\bar{\mathcal{G}},2} \coloneqq \sum_{r=1}^{T-1} re^{-\lceil r/K \rceil I_1(\mu_1 - \varepsilon_1)} = \mathcal{O}(1)$, using the concentration of the leader and a union bound on its sample size.

Combining these results, we obtain the following bound on $n_k(T)$ for arm $k$ as

$$n_k(T) \leq m_k(T) + Te^{-f_k(m_k(T), \mathcal{B}_{k,m_k(T)}, \mu_1 - \varepsilon_1)} + \sum_{n=m_k(T)}^{T-1} \mathbb{P}\left(\mathcal{Y}_n^k \notin \mathcal{B}_k\right) + \mathcal{O}(1) .$$

If the assumptions of the lemma are satisfied and $m_k(T)$ is chosen as suggested the result is proved. □

The objective is then to find the proper confidence ball $\mathcal{B}_{k,n}$ and associated rate function $f_k$ under the assumptions of the three theorems considered.

**BDS** As the distributions are bounded $\mathcal{B}_{k,n}$ can be chosen as a Levy ball, as in Chapter 5. Furthermore, Lemma 5.8 provides

$$\mathbb{P}_{w \sim \mathcal{D}_{n+1}}\left(\sum_{i=1}^{n} w_i Y_i + w_{n+1} B_{\text{BDS}}(\mathcal{Y}, \hat{\mu}) \geq \widehat{\mu}\right) \leq e^{-(n+1)\mathcal{K}_{\inf}^{B_{\text{BDS}}(\mathcal{Y},\widehat{\mu})}(F_{k,n},\widehat{\mu})} ,$$

where $B_{\text{BDS}}$ denotes the exploration bonus of BDS and $F_{k,n}$ is the empirical distribution associated to the data points *and* the bonus (counting as one observation). Setting $m_k(T)$ as the dominant term of the theorem and using the continuity of the $\mathcal{K}_{\inf}$ for bounded distributions (in the upper bound too) provides the result.

**QDS** We form $\mathcal{B}_{k,n}$ with a Levy ball around the true distribution, that we augment with the concentration of the bonus and CVaR,

$$\mathcal{B}_{k,n} = \{\mathcal{Y} \in \mathbb{R}^n : d_L(\nu_{\mathcal{Y}}, \nu_k) \leq \varepsilon, |B(\mathcal{Y}, \rho, \mu) - B_{k,\rho,\mu}| \leq \varepsilon_1, C_\alpha(\nu_{\mathcal{Y}}) \leq C_\alpha(\nu_k) + \varepsilon_2\} ,$$

for some $\varepsilon > 0, \varepsilon_1, \varepsilon_2 > 0$, denoting by $\nu_{\mathcal{Y}}$ the empirical distribution associated with a set $\mathcal{Y}$, $B_{k,\rho,\mu} = \mu + \rho \times \mathbb{E}_{\nu_k}[(\mu - X)_+]$. If $X \sim \nu_k$, $(\mu - X)_+$ is also a light-tailed variable, so the corresponding term in the upper bound of $n_k(T)$ is an additive constant. For the CVaR, we rely on the concentration of Wasserstein metrics for light-tailed distribution, using that (Lemma 2

from (Bhat and L.A., 2019b)) for two distributions $\nu$ and $\nu'$ it holds that

$$|C_\alpha(\nu) - C_\alpha(\nu')| \leq \frac{1}{1-\alpha} W_1(\nu, \nu') \ .$$

Then, Theorem 2 from (Fournier and Guillin, 2015) provides a concentration inequality for the Wasserstein distance. Combining these elements we obtain

$$\sum_{n=1}^{+\infty} \mathbb{P}_{\mathcal{Y}_n} \left( \mathcal{Y}_n \notin \mathcal{B}_{k,n} \right) < +\infty \ .$$

We can now consider the BCP under $\mathcal{Y}_n \in \mathcal{B}_{k,n}$. The upper bound of lemma 5.8 still holds, even if it is used with the truncated distribution. Hence, the QDS index satisfies

$$\mathbb{P}\left(\mu(\mathcal{Y}_n, \widehat{\mu}) \geq \widehat{\mu}\right) \leq \exp\left(-(n+1)\mathcal{K}_{\inf}^{M_{\mathcal{Y}_n}}(\mathcal{T}(\nu_{\mathcal{Y}_n}), \widehat{\mu})\right) \ .$$

If $\mathcal{Y}_n \in \mathcal{B}_{k,n}$, then $M_{\mathcal{Y}_n}$ is upper bounded by

$$M_{\mathcal{Y}_n} \leq \max\left(C_\alpha(\nu_k), B_{k,\rho,\mu}\right) + \max(\varepsilon_1, \varepsilon_2) \ ,$$

We then define $\mathfrak{M}_k^C = \max\left(C_{\alpha'}(\nu_k), B_{k,\rho,\mu}\right)$, that is independent of the run of the bandit algorithm. Finally, the definition of the Levy distance ensures that $d\left(\nu_{\mathcal{Y}_n}, \nu_k\right) \leq \varepsilon \Rightarrow d\left(\mathcal{T}(\nu_{\mathcal{Y}_n}), \mathcal{T}(\nu_k)\right) \leq \varepsilon$. Hence, we can use the continuity of $\mathcal{K}_{\inf}^{M_k}$ in all arguments (including $M_{\mathcal{Y}_n}$, see e.g (Honda and Takemura, 2015)) and obtain that for any $\varepsilon_0$ we can calibrate $\varepsilon, \varepsilon_1, \varepsilon_2$ in order to obtain for any $\widehat{\mu} \leq \mu_1 - \varepsilon_1$

$$\mathbb{P}\left(\mu(\mathcal{Y}_n, \widehat{\mu}) \geq \widehat{\mu}\right) \leq \exp\left(-(n+1)\left(\mathcal{K}_{\inf}^{\mathfrak{M}_k^C}(\mathcal{T}(\nu), \mu_1) - \varepsilon_0\right)\right) \ ,$$

which gives the first order term of the regret upper bound choosing $m_k(T) = \frac{\log T}{\mathcal{K}_{\inf}^{\mathfrak{M}_k^C}(\mathcal{T}(\nu_k), \mu_1) - \varepsilon_0}$ in Lemma 6.18.

**RDS** In this setting, we only need to control the sample $\mathcal{Y}_n$ through its mean, the "positive gap" used in the bonus, and a range on its maximum value. Hence, we fix some $\varepsilon > 0$ and consider

$$\mathcal{B}_{k,n} = \left\{ \mathcal{Y} \in \mathbb{R}^n : \bar{Y}_n \leq \mu_k + \varepsilon, \bar{Z}_n \leq \Delta_k^+ + \varepsilon, \sigma(\mathcal{Y}, \mu) \leq \sigma_{k,\mu} + \varepsilon, \mathcal{Y}^+ \in [m_n, M_n] \right\} \ ,$$

where $\mathcal{Y}^+$ the maximum of the set $\mathcal{Y}$, $(m_n)_{n\in\mathbb{N}}$, $(M_n)_{n\in\mathbb{N}}$ are two fixed sequences, and $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^n (\mu - Y_i)_+$ is the empirical excess gap.

We start by claiming that the two conditions $\{\mu(\mathcal{Y}) \leq \mu_k + \varepsilon\}$ and $\{\mu(\mathcal{Y}^+) \leq \Delta_k^+ + \varepsilon\}$ lead to a constant upper bound, just as in the previous paragraph. To obtain the same result with the event involving the standard deviation we consider the Wasserstein metric $W_p$, that is defined between two distributions $\nu$ and $\nu'$ of real random variables as

$$\mathcal{L}_p(\nu, \nu') = \inf \left\{ \int_{\mathbb{R}\times\mathbb{R}} |x-y|^p \xi(dx, dy) : \xi \in \mathcal{H}(\nu, \nu') \right\} ,$$

where $\mathcal{H}(\nu, \nu')$ is the set of all probability measures on $\mathbb{R} \times \mathbb{R}$ with marginals $\nu$ and $\nu'$. Then, the Wasserstein metric $W_p(\nu, \nu')$ is defined as $W_p(\nu, \nu') = \mathcal{L}_p(\nu, \nu')^{1/p}$ for $p > 1$. Two reasons motivate the use of this metric in our case: 1) concentration inequalities exist for $\mathcal{L}_p$ for light-tailed distribution, and 2) the moments of order $p$ are continuous with respect to the Wasserstein metric $W_p$ (see Theorem 6.9 in (Villani, 2008)). These two properties make $W_p$ a good substitute for the Levy metric used for bounded distributions. Here we choose $W_2$ as we want to control moments of order 2, and obtain with the parameters of Theorem 2 of (Fournier and Guillin, 2015) a concentration inequality that is sufficient to conclude.

We now investigate possible values for the sequence $m_n$ and $M_n$ that would allow $\mathcal{B}_{k,n}$ to happen with high probability. As we saw in Chapter 4, the maximum $\mathcal{Y}_n^+$ of a set of $n$ i.i.d random variables $\mathcal{Y}_n = (Y_1, \dots, Y_n)$ has an explicit distribution, which is (in terms of the cdf $F_k$ of $\nu_k$) for any $x \in \mathbb{R}$,

$$\mathbb{P}_{\mathcal{Y}_n \sim \nu_k^n}(\mathcal{Y}_n^+ \leq x) = F_k(x)^n .$$

First we calibrate the term $M_n$, we calibrate it to ensure that $\mathbb{P}(\mathcal{Y}_n^+ \leq M_n) \geq 1 - \frac{1}{n\log(n)^2}$, so that

$$M_n = F_k^{-1}\left(\left(1 - \frac{1}{n(\log n)^2}\right)^{\frac{1}{n}}\right) \leq F_k^{-1}\left(\exp\left(-\frac{1}{n^2(\log n)^2}\right)\right) .$$

This way, $\sum \mathbb{P}(\mathcal{Y}_n^+ \leq M_n)$ converges. Then we consider $m_n$, and this time we want $\mathbb{P}(\mathcal{Y}_n^+ \leq m_n) \leq \frac{1}{n\log(n)^2}$ to ensure the same convergence guarantees. We obtain

$$m_n = F_k^{-1}\left(\frac{1}{n(\log n)^2}^{\frac{1}{n}}\right) = F_k^{-1}\left(\exp\left(-\frac{\log n + 2\log\log n}{n}\right)\right) .$$

Combining all these results, we obtain

$$\sum_{n=1}^{T-1} \mathbb{P}_{\mathcal{Y}_n \sim \nu_k^n}(\mathcal{Y}_n \notin \mathcal{B}_{k,n}) = \mathcal{O}(1) .$$

We now use Lemma 5.8 and the fact that for any $\eta \in [0, 1)$ and $x \in (-\infty, \eta]$, $-\log(1-x) \leq x + \frac{1}{1-\eta}\frac{x^2}{2}$. Denoting $M_{\mathcal{Y}_n} = \max\left(\bar{\mathcal{Y}}_n, B(\mathcal{Y}_n, \rho_n, \mu)\right)$, $Y_{n+1} = B(\mathcal{Y}_n, \rho_n, \mu)$ and using the representation of Dirichlet samples as normalized exponential variables, the Chernoff inequality

provides

$$\begin{aligned}
&\mathbb{P}_{w\sim\mathcal{D}_{n+1}}\left(\sum_{i=1}^{n} w_i Y_i + w_{n+1} B(\mathcal{Y}_n, \rho_n, \mu) \geq \mu\right) \\
&\qquad \leq \exp\left(-\sum_{i=1}^{n+1} \log\left(1 - \eta \frac{Y_i - \mu}{M_{\mathcal{Y}_n} - \mu}\right)\right) \\
&\qquad \leq \frac{1}{1-\eta} \exp\left(-\sum_{i=1}^{n} \log\left(1 - \eta \frac{Y_i - \mu}{M_{\mathcal{Y}_n} - \mu}\right)\right) \\
&\qquad \leq \frac{1}{1-\eta} \exp\left(\sum_{i=1}^{n}\left(\eta \frac{Y_i - \mu}{M_{\mathcal{Y}_n} - \mu} + \frac{\eta^2}{2(1-\eta)}\left(\frac{Y_i - \mu}{M_{\mathcal{Y}_n} - \mu}\right)^2\right)\right) \\
&\qquad = \frac{1}{1-\eta} \exp\left(-n\eta \frac{\bar{\Delta}_n}{M_{\mathcal{Y}_n} - \mu} + n \frac{\eta^2}{2(1-\eta)} \frac{\bar{\sigma}_n(\mu)^2}{(M_{\mathcal{Y}_n} - \mu)^2}\right) ,
\end{aligned}$$

where $\bar{\Delta}_n = \frac{1}{n}\sum_{i=1}^{n} \mu - Y_i$, $\bar{\sigma}_n^2(\mu) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2$.

We recall that we consider this upper bound under the event $\mathcal{Y}_n \in \mathcal{B}_{k,n}$, which ensures that 1) $\bar{\mathcal{Y}}_n \in [m_n, M_n]$ with the sequences we defined, 2) $\bar{\Delta}_n \geq \mu - \mu_k + \varepsilon$, 3) the bonus is upper bounded by $\mu + \rho_n \times (\Delta_k^+ + \varepsilon)$, and 4) the quadratic deviation satisfies $\bar{\sigma}_n(\mu) \leq \sigma_{k,\mu} + \varepsilon$. For any $\varepsilon_0 > 0$, if we further assume that $M_n = o(m_n^2)$, for any $n$ large enough these results finally provide

$$\begin{aligned}
\mathbb{P}\left(\mu(\mathcal{Y}_n, \mu) \geq \mu\right) &\leq \frac{1}{1-\eta} \exp\left(-n\eta \frac{\Delta_k - \varepsilon}{\max(M_n, B_n) - \mu} + n \frac{\eta^2}{2(1-\eta)} \frac{(\sigma_{k,\mu} + \varepsilon)^2}{(m_n - \mu)^2}\right) \\
&\leq \frac{1}{1-\eta} \exp\left(-n\eta \frac{\Delta - \varepsilon_0}{\max(M_n, B_n) - \mu}\right) ,
\end{aligned}$$

where $B_n = \mu + \rho_n \left(\mathbb{E}_{\nu_k}\left[(\mu - X)_+\right] + \varepsilon\right)$. The condition $M_n = o(m_n^2)$ is satisfied for light-tailed distributions, as they generally have at most a poly-logarithmic growth of the maximum (e.g $\log(n)$ for exponential tails, $\sqrt{\log n}$ for gaussian tails, ...) and so $M_n$ and $m_n$ are actually of the same order of magnitude. We then recover all the terms of Theorem 6.17 by matching the exponent of the upper bound with $-\log T$.

### 6.5.2 Proof of sufficient exploration

In the following we denote by $E_n$ the term to upper bound to prove Assumption 6.6.

**BDS** We use the hypothesis $\mathbb{P}([B - \gamma, B]) \geq p$ and the second component of the bonus, $\bar{\mathcal{Y}}_n + \gamma := \max Y_i + \gamma$ along with Corollary 6.12 to obtain

$$
\begin{aligned}
E_n \leq & \mathbb{E}_{\mathcal{Y}_n}\left[\frac{\mathbb{1}(\mu(\mathcal{Y}_n) \leq \mu)(\mathbb{1}(\bar{\mathcal{Y}}_n \leq B-\gamma) + \mathbb{1}(\bar{\mathcal{Y}}_n \geq B-\gamma))}{\mathbb{P}\left(\widetilde{\mu}(\mathcal{Y}_n, \mu) \geq \mu\right)}\right] \\
\leq & \underbrace{(1-p)^n e^{\frac{n}{\rho}}}_{E_{n,1}} + \underbrace{\mathbb{E}_{\mathcal{Y}_n}\left[\frac{\mathbb{1}(\mu(\mathcal{Y}_n) \leq \mu)\mathbb{1}(\bar{\mathcal{Y}}_n + \gamma \geq B)}{\mathbb{P}\left(\widetilde{\mu}(\mathcal{Y}_n, \mu) \geq \mu\right)}\right]}_{E_{n,2}} .
\end{aligned}
$$

The two terms correspond to the two possible expressions for the bonus. The term $E_{n,1}$ gives the sufficient condition for the tuning of $\rho$ in Theorem 6.13 with

$$
\rho > \frac{-1}{\log(1-p)} \Rightarrow \sum_{n=1}^{+\infty} E_{n,1} = \mathcal{O}(1) .
$$

In the second term, the exploration bonus is larger than $B$, so we can use the proof technique of Riou and Honda (2020), based on a discretization scheme and on Lemma 5.10. Hence, $\sum_{n=1}^{T-1} E_n = \mathcal{O}(1)$ and so Assumption 6.6 is satisfied by BDS if $\rho$ satisfies the condition of Theorem 6.13.

**QDS** We use the assumption that rewards are semi-bounded with a range $[b, +\infty]$ to apply again the proof based on discretization of Riou and Honda (2020). We first find a value $y$ and a discretization step $\eta$ (unknown to the algorithm) such that truncating the values $Y_i$ to $\min(Y_i, y)$, and truncating each $Y_i < y$ to $\widetilde{Y}_i = \eta \left\lfloor \frac{Y_i}{\eta} \right\rfloor$ preserves the ranking of the arms. We denote by $S$ the number of items, and $\beta \in \mathbb{N}^S$ the vector of counts for each item. We directly use Lemma 5.10 and for any $\beta \in \mathbb{N}^S : ||\beta||_1 = n$ we want to lower bound

$$
K_\beta = \mathrm{KL}(\beta/n, \widetilde{\nu}_1) - \mathcal{K}_{\inf}^{m_\beta}(\beta/n, \mu_k) ,
$$

where $\widetilde{\nu}_1$ denote the discretized/truncated version of $\nu_1$ and $m_\beta$ denotes the maximum between the largest item with a non-zero coefficient in $\beta$ and the exploration bonus.

We recall that QDS summarizes the information larger than the empirical $(1-\alpha)$-quantile by their mean (i.e the $\mathrm{CVaR}_\alpha$ of the empirical distribution). The truncation in $y$ does not change that, and will simply makes this quantity smaller which will itself makes the BCP smaller. We use the result from Honda and Takemura (2010) (proof of Theorem 7) stating that for any $\beta$

$$
\mathcal{K}_{\inf}^{m_\beta}(\beta/n, \mu_k) \leq \frac{\bar{\Delta}_n}{m_\beta - \mu} \leq \frac{\bar{\Delta}_n}{\rho \bar{\Delta}_n^+} \leq 1/\rho ,
$$

since $m_\beta$ is at least larger than the exploration bonus. This means that for any $\xi > 0$ it holds that $K_B \geq \xi$ on all the sub-space of empirical distributions satisfying $\mathrm{KL}(\beta/n, \widetilde{\nu}_1) \geq (1+\xi)/\rho$. We now use Pinsker inequality to link the KL divergence with the total variation $\delta$, in the sub-space where $\mathrm{KL}(\beta/n, \widetilde{\nu}_1) \leq (1+\xi)/\rho$,

$$\delta(\beta/n, \widetilde{\nu}_1) \leq \sqrt{\frac{1+\xi}{2\rho}} \, .$$

If this quantity is small, we can control the probability of each measurable event. In particular, we want the quantile *used by the algorithm* to be strictly larger than the $(1-\alpha)$-quantile *of the assumption* of Theorem 6.15. If the parameter of the condition of the theorem is $\alpha$, and we run the algorithm with a parameter $\alpha' < \alpha$, we know that with proper tuning of $\rho$ we will have $F_{k,n}(q_{1-\alpha}(F_k)) < 1-\alpha$. This means that the *true quantile* $q_{1-\alpha}(\nu_k)$ *is present in the set* $\mathcal{Y}_n$ *and is not truncated by the algorithm*. In particular, if $\rho \geq \frac{1+\alpha'}{\alpha'^2}$ this is satisfied, and finally

$$\begin{aligned}
\mathrm{KL}(\beta/n, \widetilde{\nu}_1) - \mathcal{K}_{\inf}^{m_\beta}(\beta/n, \mu_k) &\geq \mathcal{K}_{\inf}^{\mathcal{F}}(\beta/n, \mu_1 - \eta) - \mathcal{K}_{\inf}^{q_{1-\alpha'}}(\beta/n, \mu_k) \\
&\geq \mathcal{K}_{\inf}^{q_{1-\alpha}}(\beta/n, \mu_1 - \eta) - \mathcal{K}_{\inf}^{q_{1-\alpha'}}(\beta/n, \mu_k) \\
&\geq \mathcal{K}_{\inf}^{q_{1-\alpha'}}(\beta/n, \mu_1 - \eta) - \mathcal{K}_{\inf}^{q_{1-\alpha'}}(\beta/n, \mu_k) \\
&\geq \kappa \, ,
\end{aligned}$$

for some $\kappa > 0$ and thanks to the definition of the family $\mathcal{F}^{\alpha}_{[b,+\infty]}$. This result concludes the proof as it ensures that Assumption 6.6 is satisfied by the QDS algorithm on $\mathcal{F}^{\alpha}_{[b,+\infty]}$.

**RDS** Corollary 6.12 gives a lower bound of the BCP in $e^{-\frac{n}{\rho_n}}$. We directly use this result and obtain

$$E_n \leq e^{-n(I_1(\mu) - 1/\rho_n)} \, ,$$

and for $n$ large enough $\rho_n > 2I_1(\mu)$, which is sufficient to obtain the convergence of $\sum_{n=1}^{+\infty} E_n$. Hence, for RDS the increasing exploration bonus directly provides larger-than-logarithmic exploration.

## 6.6 Experiments: crop-farming and synthetic problems

### 6.6.1 Application in agriculture: bandits with DSSAT

We benchmark our algorithms using the DSSAT crop yield simulator[3] (Hoogenboom et al., 2019), to emulate a simplified version of the decision problem in agriculture we introduced in the preamble of this manuscript. We consider the same use-case as in Section 5.6.2 of Chapter 5. More specifically, we model the problem of selecting a planting date for maize grains among 7 possible options, everything else being equal, as a 7-armed bandit. We recall that the resulting distributions incorporate historical variability as well as exogenous randomness coming from a stochastic meteorologic model. We illustrate this in Figure 6.1 with their histogram computed on $10^6$ samples. They are typically right-skewed, multimodal and exhibit a peak at zero corresponding to years of poor harvest, hence they hardly fit to a convenient parametric model (e.g SPEF/sub-Gaussian…). This problem is the same as the one studied in the previous chapter, but this time we consider standard regret minimization instead of CVaR bandits.

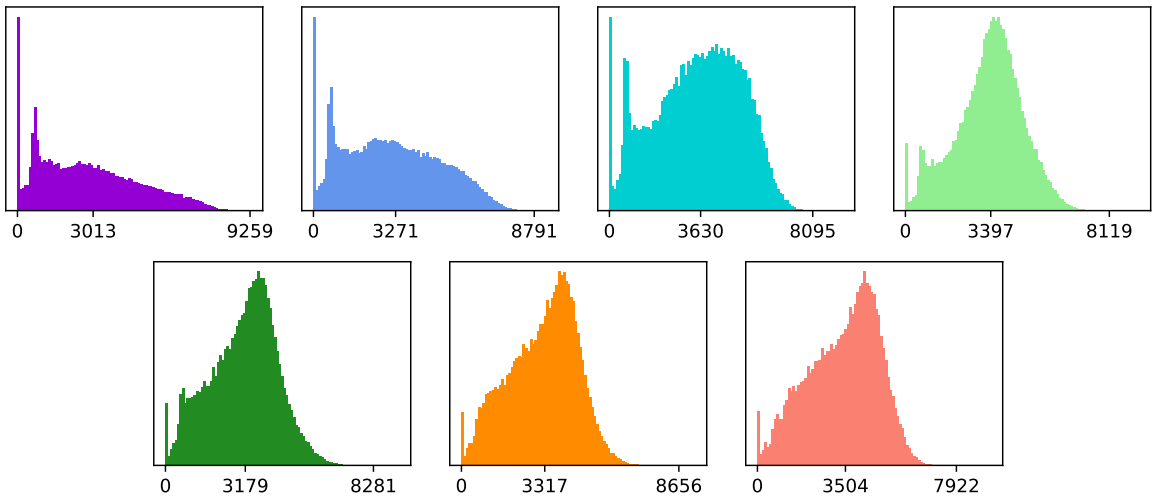


**Figure 6.1** – Distribution of simulated dry grain yield (kg/ha) for seven different planting dates over $10^6$ samples. Reported on the x-axis are the distribution minimum, mean and maximum values. In the setting considered in this chapter the optimal arm is the third one (mean 3630 kg/ha).

**Benchmarks** A natural choice for the learner would be to use algorithms adapted for bounded distributions with known support. Indeed, one could argue that crop yields are fundamentally bounded by a very large value, that can be provided with some expert knowledge. However this method may have limits when the upper bound cannot be estimated accurately (few data, new environment, …), as a conservative bound can have a cost on the regret. For this reason,

[3] *Decision Support System for Agrotechnology Transfer*, https://dssat.net/

we believe that the novel Dirichlet Sampling algorithms we introduce in Section 6.4 are a good alternative choice for this problem. In particular, the three algorithms we propose in this chapter are relevant in this setting: BDS keeps the bounded-support hypothesis but introduces the possible uncertainty on the bound, while the light-tailed hypothesis of RDS and the quantile condition of QDS look reasonable.

We compare the three DS algorithms with some of the algorithms introduced in Section 1.1 that assume bounded distributions. Under this assumption one can use classical algorithms such as UCB1 (Auer et al., 2002a) or Thompson Sampling with Beta prior using the binarization trick introduced in (Agrawal and Goyal, 2012a). These algorithms enjoy logarithmic regret without the optimal rate of (Burnetas and Katehakis, 1996). We also compare DS with two optimal algorithms: IMED (Honda and Takemura, 2015) and NPTS (Riou and Honda, 2020). We recall that the former is based on the explicit calculation of the $\mathcal{K}_{\inf}$ function for bounded distributions. These algorithms require the explicit knowledge of an upper bound on the support of the arms distributions. To represent the fact that a tight bound is sometimes unknown to the practitioner (uncertain environment, possibility of yet unobserved black swan events...) we run two variants of the above algorithms, one with the exact maximum yield across all simulated data, which we believe is a strong prior information, and one with the same bound inflated by $50\%$, which we deem a conservative estimate. Finally, we include RB-SDA (presented in Chapter 2) that requires no parameter and is then agnostic of the choice of the upper bound. However, its theoretical guarantees are not clear here.

**Tuning** For BDS we choose the parameters $\rho = 4, \gamma = 3500$, corresponding to $p \approx 20\%$ in the hypothesis of Theorem 6.13, which is conservative in our example. For QDS, we set $\rho = 4$ to be able to compare with BDS and a quantile $95\%$. Finally for RDS, we choose $\rho_n = \sqrt{\log(1+n)}$, which enters into the theoretical framework of Theorem 6.17.

**Results** We report the performance of DS algorithms and their competitors on the bandit problem using the DSSAT simulator we introduced in the two settings we propose: the first with a tight upper bound, and the second with a "conservative" estimate ($1.5$ times larger).

We present the empirical regret of the algorithms for $T = 10^4$ and $5000$ simulations in Figure 6.2. The first striking result is that the three DS algorithms perform similarly to the optimal algorithms using the exact upper bound (IMED and NPTS), and clearly outperform the non-optimal ones (UCB1 and binarized TS). The poor performance of non-optimal algorithms hints that this particular bandit instance is not easy and requires more sophisticated methods. Furthermore, Table 3 in (Baudry et al., 2021c) shows that if RB-SDA achieves good performance, it also exhibits larger dispersion than other methods ($95\%$ quantile is $0.99 \times 10^6$, standard

deviation is $0.26 \times 10^6$). It is not clear that RB-SDA operates in its theoretical scope in this setting.

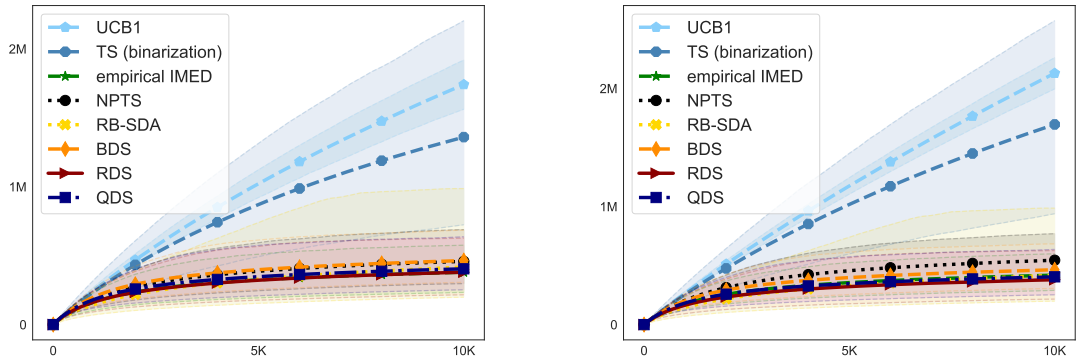


**Figure 6.2** – Average regret on 5000 simulations and horizon $T = 10^4$. Dashed lines correspond to 5%-95% regret quantiles. UCB1, Binarized Thompson Sampling, Empirical IMED and NPTS are run with exact upper bounds around $1.5 \times 10^4$ kg/ha (left) and the conservative upper bound $1.5 \times 10^4$ kg/ha (right). BDS: $\rho = 4$. RDS: $\rho_n = \sqrt{\log(1+n)}$. QDS: $\rho = 4, \alpha = 5\%$.

For these reason, we know focus on the comparison with IMED and NPTS only, plotting the regret of a selection of algorithms in Figure 6.3 for better visualization. Interestingly, RDS is the overall winner of the experiment considering the two settings (tight and conservative bounds). We see that if Dirichlet Sampling algorithms achieve similar regret to their competitors when the latter are allowed to use the "exact" upper bound, they compare favorably with the conservative estimate. Indeed, the performance of NPTS (and to a lesser extent IMED) is deteriorated by the conservative upper bound. Considering this, RDS seems to be the overall winner in both experiments. We think this demonstrates the merits of trading-off logarithmic regret (albeit only by a factor $\mathcal{O}(\log \log T)$) for finite-time adaptation to the tail behavior via the leverage $\rho_n$. As a side remark, note that our round-based implementation is more efficient than NPTS as it does not draw random weights for the leader, which is the most costly operation at each round.

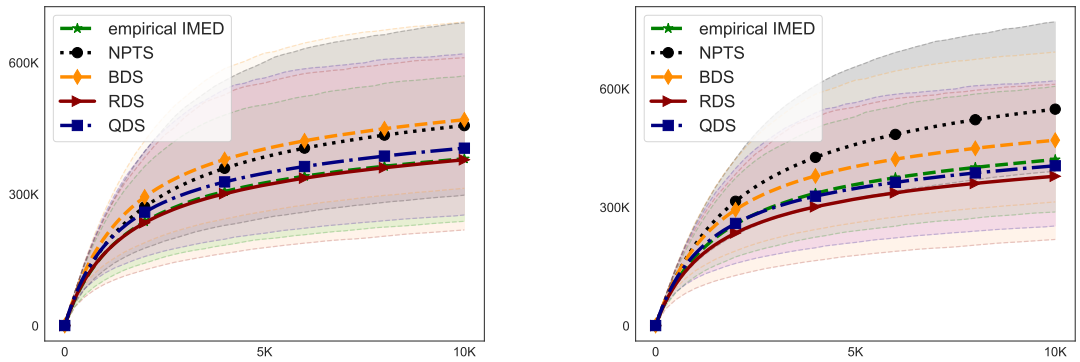


**Figure 6.3** – Average regret on 5000 simulations and horizon $T = 10^4$. Dashed lines correspond to 5%-95% regret quantiles. Empirical IMED and NPTS are run with exact upper bounds around $1.5 \times 10^4$ kg/ha (left) and the conservative upper bound $1.5 \times 10^4$ kg/ha (right).

### 6.6.2 Experiments on synthetic data

To further illustrate the properties of DS algorithms, we perform additional experiments on synthetic examples. First, we test the *sensitivity* of DS w.r.t its hyper-parameters, and check that their impact on the performance of the algorithms is moderate. Then, we show the merits of RDS in case of *model mis-specification*, inspired the robustness experiments of Ashutosh et al. (2021). Finally, we consider the case of *Gaussian mixtures*, a common tool to model non-parametric distributions via *kernel density estimation*, and show that they fit the scope of DS but not that of usual bandit algorithms.

**Sensitivity of BDS to its parameters** We study the sensitivity of BDS to its parameter $\rho$. Theorem 6.13 suggests to scale the exploration bonus $B_{\rho,\gamma}$ as $\rho = -1/\log(1-p)$, which is a proxy of an upper bound of $1/(1-F(\mu_1))$ in Lemma 6.9. We believe this bonus to be rather conservative when $p$ is small and the distributions considered exhibit little skewness; as an example, if a distribution is such that at most $25\%$ of its mass is located to the right of the optimal mean reward $\mu^*$, $\rho \approx 4$ should be a suitable tuning.

To investigate this, we consider a toy bandit instance with two arms following uniform distributions on $[0,1]$ and $[0.2, 0.9]$ respectively (note that the upper bound is different for each arm yet the distribution of mass near their respective bounds is the same, thus fitting the setting of BDS). These distributions are shown in Figure 6.4, and in particular their means are $0.5$ and $0.55$ respectively. For $\gamma = 0.1$, we compute the expected regret of BDS obtained with the theoretical tuning $\rho = -1/\log(1-p) \simeq 9.5$, and compare it with other choices of $\rho$. Figure 6.4 shows that only the most extreme tuning $\rho = 50$ exhibits significant, albeit still sublinear, regret. Small deviations from the theoretical tuning yields similar regret, the heuristic $\rho = 4$ discussed above being slightly better, which tends to confirm our belief that the analysis of Theorem 6.13 can be sharpened. Note that the exploration incentive given by $\rho$ is necessary since smaller values (e.g $\rho = 0.1$) tends to accumulate more regret.

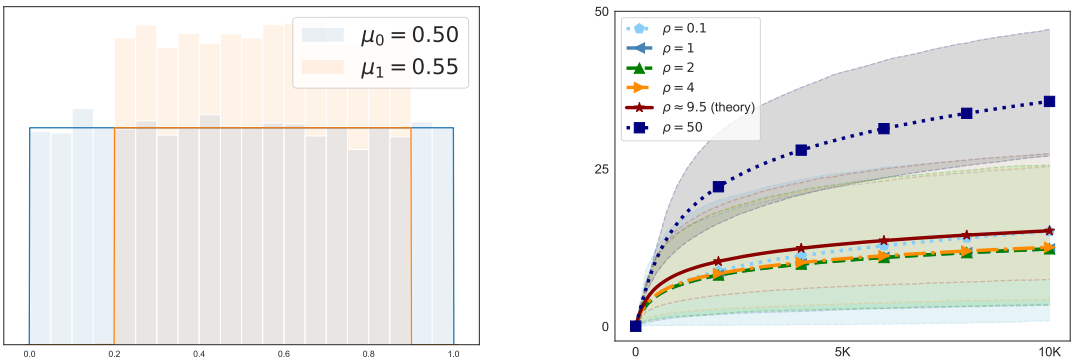


**Figure 6.4** – Left: bandit with two uniform arms $\mathcal{U}(0,1)$ and $\mathcal{U}(0.2,0.9)$ ($10^4$ samples each). Right: average regret on $5000$ simulations and horizon $T = 10^4$ of BDS for various values of $\rho$.

**Robustness for light-tailed bandits: comparison with R-UCB-LT** The study of statistically robust bandit algorithms is fairly recent, and as such is yet to have well-established benchmarks. Ashutosh et al. (2021) introduce R-UCB-LT, an adaptation of the standard sub-Gaussian UCB to enforce robustness w.r.t light-tailed distribution. We reproduce the setting of their experiment, namely two Gaussian arms $\mathcal{N}(1,1)$ and $\mathcal{N}(2,3)$, and compare several variants of both R-UCB-LT and RDS against a misspecified UCB1 (the misspecification takes the form of an overly optimistic 1-sub-Gaussian assumption, while the second arm is only $\sqrt{3}$-sub-Gaussian). Both R-UCB-LT and RDS rely on a slowly growing exploration bonus, denoted respectively by $f$ and $\rho$; we run both algorithms with $f$ and $\rho$ equal to $\log^2$, $\log$ and $\sqrt{\log}$.

Results are reported in Figure 6.5. As expected, the misspecified UCB1 exhibits much faster regret growth than the robust algorithms. However, RDS seems to outperform R-UCB-LT, the best average regret being achieved by RDS with $\rho_n = \sqrt{\log(1+n)}$ and $\rho_n = \log(1+n)$. Furthermore, the regret of RDS appears to be somewhat monotonic (slightly increasing) with respect to the hyperparameter $\rho$, and the best results are achieved by the one matching the asymptotic growth rate of the maximum of a i.i.d Gaussian samples, as recommended by Theorem 6.17. On the other hand, the best version of R-UCB-LT is obtained with $f \approx \log$ (for which we do not find a theoretical intuition) and the performance gap is significant when other bonuses are considered. We also tested R-UCB-LT with powers of $\log\log$ with similar results; we do not report these curves for the readability of the figures. In light of these results, RDS seems less sensitive to its parameter choice than R-UCB-LT.

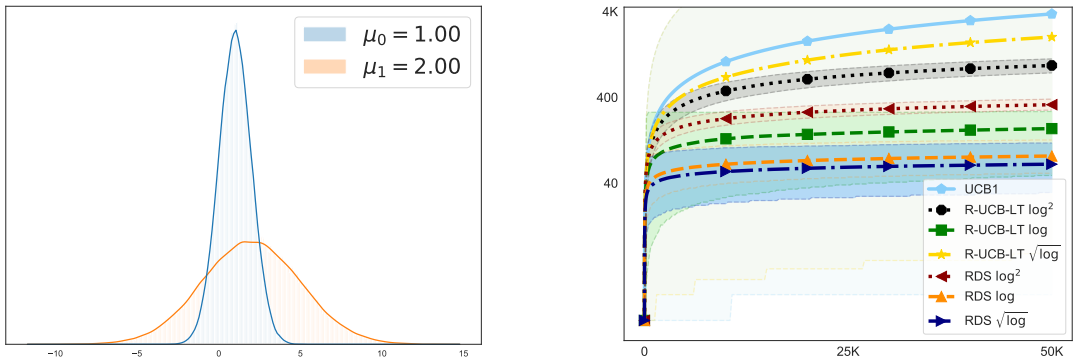


**Figure 6.5** – Left: Gaussian arms $\mathcal{N}(1,1)$ and $\mathcal{N}(2,3)$ ($5 \times 10^4$ samples each). Right: average regret (in log scale) on 5000 simulations and horizon $T = 5 \times 10^4$. UCB1 runs assuming a 1-sub-Gaussian instance.

**Gaussian Mixture** Many real-world situations (loss profile of a portfolio of financial assets, crop yields, statistics of heterogeneous populations...) exhibit multimodal distributions. The Gaussian mixture model is perhaps the simplest example of such distributions and is ubiquitous in many areas of machine learning and engineering (speech recognition, clustering...), in particular as a nonparametric model for kernel density estimation. Still, to the best of our knowledge, it escapes the scope of current optimal bandit methods as it is neither bounded

nor SPEF. Thanks to the different sets of assumptions in which they operate, both RDS and QDS are eligible algorithms to tackle the problem of sequential decision-making in a Gaussian mixture environment, at the cost of slightly larger-than-logarithmic regret and slightly lower $\mathcal{K}_{\inf}$ rate respectively.

We consider two arms distributed as a $50\%$-$50\%$ independent mixture of $\mathcal{N}(-0.3, 0.5^2)$ and $\mathcal{N}(1.3, 0.5^2)$ and a $10\%$-$80\%$-$10\%$ independent mixture of $\mathcal{N}(-1.5, 0.5^2)$, $\mathcal{N}(0.6, 0.5^2)$ and $\mathcal{N}(2.5, 0.5^2)$. Note that both mixtures have total variance equal to $0.5^2$. Due to the lack of theoretically grounded benchmark, we run three SPEF algorithms (kl-UCB, IMED and Thompson Sampling) assuming the arms belong to the SPEF of Gaussian distributions with fixed variance $0.5^2$. This is an example of *model misspecification*.

We run RDS with $\rho_n = \sqrt{\log(1+n)}$, which matches the asymptotic growth rate of the maximum of i.i.d Gaussian samples, and QDS with $\alpha = 5\%$, $\rho = 4$. Note that the use of QDS in this context is technically out of scope of Theorem 6.15 since Gaussian mixtures are not lower bounded; we believe however that this is an artifact of our proof technique that could be avoided with a finer analysis.

Results are reported in Figure 6.6. Both RDS and QDS outperform other existing methods; in particular, among the misspecified SPEF algorithms, only IMED exhibit comparable regret growth. This good performance of IMED is remarkable (across all our experiments), but we do not have the intuitions to explain its better consistency w.r.t for instance TS or kl-UCB. We finally note that as this bandit problem is complicated (small optimality gap, non-SPEF distributions), all algorithms have a relatively large variance.

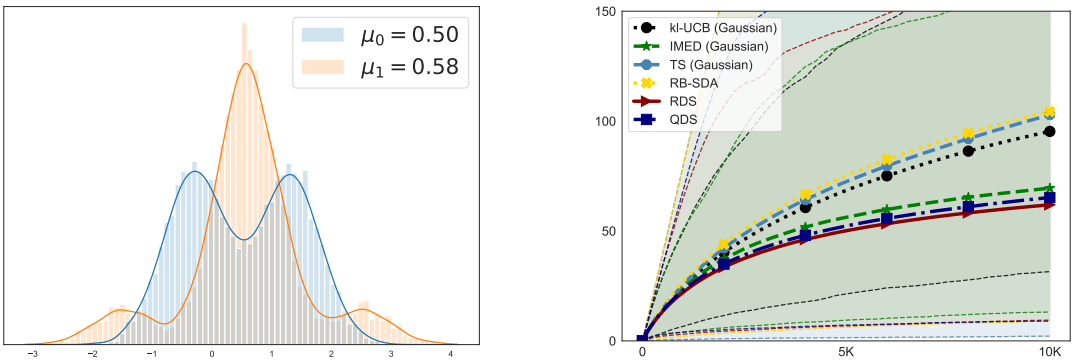


**Figure 6.6** – Left: Gaussian mixture arms ($10^4$ samples each). Right: average regret on $5000$ simulations and horizon $T = 10^4$. kl-UCB, IMED and Thompson Sampling are run assuming Gaussian arms with same variance as the mixtures. RDS: $\rho_n = \sqrt{\log(1+n)}$. QDS: $\rho = 4, \alpha = 5\%$.

# Chapter 7

# Conclusion and Perspectives

In this thesis we proposed novel algorithms to tackle several variants of the Multi-Armed Bandit problem, motivated by a case-study in agriculture. This application of bandits made us consider specifically the *regret minimization* setting, where the performance of the algorithm is evaluated throughout the whole duration of the experiment. In the preamble of this manuscript we introduced some of the challenges raised by this real-world problem, and among them two questions were particularly central in our works. The first one is about deriving algorithms with strong theoretical guarantees and practical performance while using as little prior knowledge as possible on the arms' distributions. The second question that we examined with care is the *evaluation* of bandit algorithms, that led us to consider alternative performance metrics to the expected sum of reward. This point is central for real-world applications of machine learning algorithms, where the agents involved may be *risk-averse*. This led us to consider *CVaR bandits*, where the learner wants to pull as often as possible the arm with the largest *Conditional Value at Risk* for some risk level $\alpha$, and its limit case when $\alpha \to 0$ that we related to the *Extreme Bandit* problem. Finally, to make algorithms more practical we also considered the performance of our approaches to incorporate *non stationary rewards* or *batch feedback*, that are commonly encountered in real problems. We developed two families of algorithms to tackle these questions, that we view as complementary approaches with their own merits and scopes.

## 7.1 Sub-Sampling Dueling Algorithms (SDA)

The family of algorithms considered in Part I of this thesis is based on the principle of *pairwise comparisons* between arms, that we call *duels*, where the arm that has been the most selected (leader) competes against each other arm (challengers) with *sub-samples* of its observations. The idea is that playing a variety of duels with "independent" sub-samples from the leader gives a *fair chance* for an under-sampled challenger to recover from potentially bad first rewards.

The algorithms that we proposed are inspired by BESA (Baransi et al., 2014) and SSMC (Chan, 2020), bridging the gap between the two approaches and avoiding some of their shortcomings.

In Chapter 2 we introduced in detail the SDA framework, and proposed several examples of sub-sampling algorithms, for instance Sampling Without Replacement (WR-SDA), Random Block Sampling (RB-SDA), or Last Block Sampling (LB-SDA). While WR-SDA matches BESA for two arms, RB-SDA and LB-SDA are closer in spirit to SSMC by using sequences of successively collected data, that we call *blocks* of observations. Hence, we call these sub-sampling algorithms *Block Samplers*. Our analysis can be divided in two parts. We obtained a first upper bound on the number of pulls of each sub-optimal arms for SDA using a Block Sampler, assuming only that the empirical means concentrate around the true means with an exponential rate. This property is quite general, as it is for instance satisfied by any *light-tailed* distribution. This upper bound showcases a term in $\mathcal{O}(\log(T))$ that would ideally be the dominant term of the upper bound, and the term $\sum_{r=1}^{T-1} \mathbb{P}(N_1(r) \leq C \log(r))$ for a constant $C$, that sums the probability that the *best arm* (denoted by arm 1 for simplicity) is *under-sampled*. In a second part of the analysis, we showed that these probabilities are small if:

1. The sub-sampling algorithm allows the arms to play a sufficient *diversity* of duels
2. Playing many "diverse" duels is enough to avoid under-exploration of the best arm: it must be unlikely that that the best arm loses a very large number of duels against a sub-optimal arm, even if its first draws were unlucky.

We formalized these two intuitions, that allow to conveniently separate the properties that need to be satisfied by the sub-sampling algorithm (first statement) and by the arms' distributions (second statement). We then established that the Last Block and Random Block samplers are suitable sub-sampling algorithms, before exhibiting a sufficient condition (Assumption 2.20) on the arms' distributions to obtain logarithmic regret for SDA. Intuitively, this result states that sub-sampling works if obtaining "low" rewards is more likely for the sub-optimal arms than for the best arm. We further showed that this assumption holds for some widely used families of distributions, such as *Single Parameter Exponential Families* (SPEF) or some models where a reward is obtained by adding a random noise (whose distribution is common for all arms) to the mean. In the former case LB-SDA and RB-SDA are even *asymptotically optimal*, in the sense that their regret upper bound matches the lower bound of Lai and Robbins (1985). Strikingly, these results hold without using any information on the arms' distributions for the implementation of the algorithms. Furthermore, their practical performance backs up the theory, and we showed that the instances of SDA introduced perform comparably to state-of-the-art bandit algorithms in many settings.

Considering these promising results, we decided to study possible extensions of these algorithms. In particular, in Chapters 3 and 4 we chose to work with LB-SDA due to its simplicity and low computational cost. In Chapter 3 we analyzed two variants of this algorithm

using a *limited memory*, i.e that do not keep in memory all the rewards collected, avoiding the main drawback of approaches based on sub-sampling. We first showed that a variant of LB-SDA with a *poly-logarithmic* memory in terms of the time horizon has the same asymptotic performance as the vanilla LB-SDA, allowing for a significant reduction of the memory usage (initially linear in the time horizon). Secondly, we proposed a natural adaptation of LB-SDA for *non-stationary* environments, equipping it with a *sliding window*. We proved that the resulting SW-LB-SDA algorithm achieves similar theoretical guarantees as comparable benchmarks (such as Sliding Window UCB), but in potentially broader settings where the mean of the distributions would not be the only feature evolving with time. In our experiments, we showed the interest of this approach with a simple example with Gaussian arms with both evolving means and variance: while the competitors require either to know the variance during each phase (strong knowledge) or at least an upper bound of possible variances (weaker knowledge, but deteriorated performance when the variance is actually smaller), SW-LB-SDA adapts naturally to changes. Hence, we believe that this adaptation of LB-SDA shows the potential of fully non-parametric algorithms in bandits.

Finally, LB-SDA inspired us to develop novel algorithms for *Extreme Bandits*, that we introduced in Chapter 4. We replaced the comparison of empirical means in previous chapters by the comparison of a robust estimator of the "heaviness" of a tail, that we call *Quantile of Maxima* (QoMax). We first proposed an *Explore-Then-Commit* strategy using this estimator (QoMax-ETC), before proposing QoMax-SDA, adapting LB-SDA for QoMax comparisons. We proved strong theoretical guarantees for these two algorithms, under minimal assumptions on the tails of the distributions. For instance, some of our results are obtained only assuming that one tail "dominates" (i.e is heavier than) the others asymptotically (we refer to Definition 4.3). To the best of our knowledge, this is the least restrictive assumption considered in this literature, showing again the power of non-parametric approaches, and especially of algorithms based on sub-sampling. Furthermore, both QoMax-based algorithms perform very well in practice in the experiments that we implemented, and are more efficient than their competitors in terms of computation and memory cost.

## 7.2 Dirichlet Sampling

In Part II of this thesis we proposed extensions of the *Non-Parametric Thompson Sampling* algorithm of Riou and Honda (2020). This algorithm computes noisy evaluations of the empirical means of each arm by re-weighting their observations with weights drawn from a Dirichlet distribution, initializing their history with the known upper bound of the distributions' support. In Chapter 5 we analyzed an extension of this algorithm for *CVaR bandits*, that we call *Bounded – CVaR Thompson Sampling* (B-CVTS). We proved that this algorithm is *asymptotically optimal* when distributions are bounded with a known upper bound, in the sense that the

expected number of pulls of each sub-optimal arm matches the adaptation of the lower bound of Burnetas and Katehakis (1996) for CVaR bandits. This result is interesting because B-CVTS is the first algorithm with such guarantees in CVaR bandits. Indeed, most existing competitors are adaptations of the *optimism in face of uncertainty* principle, that requires a careful design of confidence intervals on the empirical CVaR. Deriving such intervals is still an active research field, and so existing CVaR bandit algorithms may have sub-optimal performance because they are not optimally calibrated. For this reason, algorithms such as B-CVTS are particularly appealing in this setting, and the detailed experiments seem to prove the merits of this alternative approach compared to its competitors. In particular, we implemented an experiment using the DSSAT simulator that emulates the crop-management problem that we introduced in the foreword of this thesis, and for which B-CVTS clearly outperforms the other algorithms. In addition, we proved some results that further advocate for the use of B-CVTS in practice: we first showed that rewards coming in batches do not change the theoretical guarantees of B-CVTS, and we then showed that using over-estimated upper bound of the support does not alter significantly the performance of the algorithm. In fact, we even showed that for $\alpha < 1$ (i.e not in the "expectation" case) setting this upper bound to $+\infty$ still allows to prove a logarithmic CVaR-regret.

In Chapter 6 we extended NPTS in a different direction, back in the standard (expectation) setting, and analyzed possible generalizations of NPTS for alternative assumptions on the arms that could include unbounded distributions. We carefully analyzed *Boundary Crossing Probabilities* for the Dirichlet distribution, in order to provide sound theoretical tuning of an exploration bonus that would replace the support's upper bound. As the empirical version of this bonus uses the history of two arms, we proposed to use the *leader vs challenger* framework introduced in (Chan, 2020) already used for Sub-Sampling Dueling Algorithms. This algorithmic structure has computational advantages (we can avoid sampling weights for the most pulled arm), and allowed us to provide theoretical guarantees for this strategy. We call any algorithm combining these factors an instance of *Dirichlet Sampling* (DS). Motivated by our theoretical results, we proposed three DS algorithms that progressively relax the initial assumption of NPTS on the distributions' support: *Bounded Dirichlet Sampling* (BDS) achieves asymptotic optimality when the upper bound is *unknown but detectable*, for *Quantile Dirichlet Sampling* (QDS) we obtained a *logarithmic regret* when the arms satisfy a mild quantile assumption, and *Robust Dirichlet Sampling* (RDS) ensures a *slightly larger than logarithmic* regret when assuming only that the distributions are light-tailed. Hence, these three algorithms exhibit a theoretical trade-off between the level of generality of the assumption on the arms and the performance that can be guaranteed. However, our experiments showed that the three algorithms actually perform very similarly in practice, advocating for the use of the most robust one.

## 7.3 Conclusions on our contributions

We think that the works presented in the two parts of this thesis show the potential of non-parametric bandit algorithms, which we describe as algorithms that do not use a parametric model on the distributions for their implementation. For example, the Last Block Sub-sampling Dueling Algorithm (LB-SDA) and Robust Dirichlet Sampling (RDS) introduced respectively in Chapter 2 and 6 do not require any information on the arms for their implementation. As those two algorithms achieve strong theoretical and practical performance, the central message of this thesis is that using less information on the arms' distribution is not necessarily costly in terms of regret and/or practical performance. We even proved that in some cases the contrary can be true, as in some examples the algorithms are more robust to a potentially complicated or mis-specified model. We view the two families of algorithms that we studied as complementary: when the learner knows absolutely nothing on the arms' distribution except that they are light-tailed we recommend the use of RDS, if the distributions are bounded and can exhibit weird shapes (e.g as in Figure 3) NPTS or BDS are suitable, and when the learner can reasonably assume that Assumption 2.20 holds we recommend using either RB-SDA or LB-SDA. Indeed, we proved that these algorithms may have optimal theoretical guarantees even if the learner cannot precisely characterize the family of distributions.

We further showed that these two families of algorithms can be extended to broader settings, for instance considering *alternative performance metrics*. One of our main discoveries is that a simple adaptation of NPTS/DS can lead to optimal algorithms for CVaR bandits, and that similarly a variant of SDA with a robust estimator can achieve state-of-the-art performance for Extreme Bandits. Both settings could be of interest in practical use-cases such as our crop-management problem in agriculture. In the two cases the novel approaches that we introduced allowed to tackle the problems under a different perspective compared to existing literature, for instance by avoiding the conservative (and sometimes complicated) design of tight confidence intervals. These findings may open new doors for future improvements, and extensions to other settings that we did not consider during this thesis. Finally, we further demonstrated the flexibility of our algorithms by analyzing their natural extensions in usual variants of the bandit problems, respectively non-stationary bandits for LB-SDA and bandits with batch feedback for B-CVTS.

## 7.4 Perspectives

Our research was largely motivated by the crop-management problem introduced in the preamble in this thesis, and several challenges still remain open before implementing our algorithms in the real-world. For instance, we did not consider the use of contextual information in our algorithms. So far we would implement one model for each possible combination of

crop, soil, and climate types that we would encounter (or focus on one combination), while it may be possible to discover a general structure for these parameters that would allow to use a meta-model covering every combination. Some problems may also rise from the batch setting: spatial correlations between crop yields from the same years, or temporal correlations between crop yields from the same field across the seasons. Intuitively, we would like to reduce the weight of correlated data in the learning process. Furthermore, in such realistic application we may want to use all available information to improve our algorithms. For instance, the farmers may have access to weather predictions at the beginning of and during the season, or some ground measures may help adapt the strategies. So far, these considerations have been completely out of the scope of our works, that have been dedicated to solving some of the more fundamental problems we introduced. Finally, the non-stationarity (e.g due to climate change) may also be analyzed with more precise models: changes of distributions may have a structure (e.g following a climate model), and it may be possible to think about smarter ways to adapt to them if we knew something about this structure.

On the theoretical side, we did not consider some questions that could be of interest. For instance, we may want to analyze equivalents of the DS algorithms for CVaR bandits, and in particular what kind of performance could be expected from an algorithm like RDS. Also, we focused on the CVaR case as an example of alternative performance metric, but the NPTS/DS principle may work just as well for a broader variety of risk metrics.

Regarding the Sub-sampling Dueling Algorithms, several questions are also interesting for future research. For instance, we showed that Assumption 2.20 is sufficient to make SDA work, but we wonder if it is actually *necessary*. Furthermore, there may be a way to circumvent this assumption and to make SDA works in settings for which it currently fails. For instance, in (Baransi et al., 2014) the authors suggest a larger amount of forced exploration to ensure with large probability that the best arm will be estimated "well enough" so that sub-sampling is sufficient to balance exploration and exploitation. This is a possibility that we could consider. We could also imagine algorithms working in two steps: in a first step the arms are sampled enough times to ensure that some condition is met (e.g: the support is "well-covered" for all arms), and in a second step we would use SDA. A last open question that we did not have time to consider during this thesis is about the generalization of SDA to structured setting. While the number of pulls accurately represent the "quantity of information" collected for an arm in the standard setting, this is no more true in the structured case (e.g in linear bandits). Finding alternative "information measures" and new ways to implement pairwise comparisons in structured bandits is an interesting and challenging open question.

# List of Figures

# List of Algorithms

# List of Tables

# List of References


Y. Abbasi-Yadkori, A. György, and N. Lazic. A new look at dynamic regret for non-stationary stochastic bandits. *CoRR*, abs/2201.06532, 2022.

C. Acerbi and D. Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26: 1487–1503, 2002.

M. Achab, S. Clémençon, A. Garivier, A. Sabourin, and C. Vernade. Max k-armed bandit: On the extremehunter algorithm and beyond. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 389–404. Springer, 2017.

R. Agrawal. Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4), 1995.

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

S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012b.

S. Agrawal and N. Goyal. Further Optimal Regret Bounds for Thompson Sampling. In *Proceedings of the 16th Conference on Artificial Intelligence and Statistics*, 2013a.

S. Agrawal and N. Goyal. Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, pages 99–107. PMLR, 2013b.

S. Agrawal, S. Juneja, and W. M. Koolen. Regret minimization in heavy-tailed bandits. In M. Belkin and S. Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 26–62. PMLR, 2021a.

S. Agrawal, W. M. Koolen, and S. Juneja. Optimal best-arm identification methods for tail-risk measures. In *Advances in Neural Information Processing Systems*, pages 25578–25590, 2021b.

N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and system sciences*, 58(1):137–147, 1999.

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9: 203–228, 1999.

K. Ashutosh, J. Nair, A. Kagrecha, and K. Jagannathan. Bandit algorithms: Letting go of logarithmic regret for statistical robustness. In *International Conference on Artificial Intelligence and Statistics*, pages 622–630. PMLR, 2021.

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

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002b.

P. Auer, P. Gajane, and R. Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pages 138–158, 2019.

A. Baransi, O.-A. Maillard, and S. Mannor. Sub-sampling for multi-armed bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 115–131. Springer, 2014.

D. Baudry, E. Kaufmann, and O.-A. Maillard. Sub-sampling for efficient non-parametric bandit exploration. *Advances in Neural Information Processing Systems*, 33, 2020.

D. Baudry, R. Gautron, E. Kaufmann, and O. Maillard. Optimal thompson sampling strategies for support-aware cvar bandits. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 716–726. PMLR, 2021a.

D. Baudry, Y. Russac, and O. Cappé. On Limited-Memory Subsampling Strategies for Bandits. In *ICML 2021- International Conference on Machine Learning*, Vienna / Virtual, Austria, July 2021b.

D. Baudry, P. Saux, and O. Maillard. From optimality to robustness: Adaptive re-sampling strategies in stochastic bandits. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 14029–14041, 2021c.

D. Baudry, Y. Russac, and E. Kaufmann. Efficient algorithms for extreme bandits. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pages 2210–2248. PMLR, 2022.

O. Besbes, Y. Gur, and A. Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems*, pages 199–207, 2014.

L. Besson, E. Kaufmann, O.-A. Maillard, and J. Seznec. Efficient change-point detection for tackling piecewise-stationary bandits. *Journal of Machine Learning Research*, 2022.

S. P. Bhat and P. L.A. Concentration of risk measures: A wasserstein distance approach. In *Advances in Neural Information Processing Systems*, 2019a.

S. P. Bhat and P. L.A. Concentration of risk measures: A wasserstein distance approach. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b.

S. Bhatt, P. Li, and G. Samorodnitsky. Extreme bandits using robust statistics. *arXiv preprint arXiv:2109.04433*, 2021.

J. Bian and K. Jun. Maillard sampling: Boltzmann exploration done optimally. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, AISTATS 2022*, 2022.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities : a non asymptotic theory of independence*. Oxford University Press, 2013.

D. Brown. Large deviations bounds for estimating conditional value-at-risk. *Oper. Res. Lett.*, 35:722–730, 2007.

S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

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

Y. Cao, Z. Wen, B. Kveton, and Y. Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 418–427. PMLR, 2019.

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

A. Carpentier and M. Valko. Extreme bandits. In *Neural Information Processing Systems*, Montréal, Canada, Dec. 2014.

A. Cassel, S. Mannor, and A. Zeevi. A general approach to multi-armed bandits under risk criteria. In *Proceedings of the 31st Annual Conference On Learning Theory*, 2018.

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

J. Q. L. Chang, Q. Zhu, and V. Y. F. Tan. Risk-Constrained Thompson Sampling for CVaR Bandits. *CoRR*, Nov. 2020.

Y. Chen, C.-W. Lee, H. Luo, and C.-Y. Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Conference on Learning Theory*, pages 696–726. PMLR, 2019.

Y. Cho and E. Cho. The volume of simplices clipped by a half space. *Applied mathematics letters*, 14(6): 731–735, 2001.

V. A. Cicirello and S. F. Smith. The max k-armed bandit: A new model of exploration applied to search heuristic selection. In *The Proceedings of the Twentieth National Conference on Artificial Intelligence*, volume 3, pages 1355–1361, 2005.

Y. David and N. Shimkin. Pure exploration for max-quantile bandits. In P. Frasconi, N. Landwehr, G. Manco, and J. Vreeken, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 556–571, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46128-1.

A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*, volume 95. 2010.

M. Drmota and R. Tichy. *Sequences, discrepancies and applications*, volume 1651 of *Lecture Notes in Mathematics*. Springer Verlag, Deutschland, 1 edition, 1997.

B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

H. Esfandiari, A. Karbasi, A. Mehrabian, and V. S. Mirrokni. Regret bounds for batched bandits. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 7340–7348. AAAI Press, 2021.

K. J. Evans, A. Terhorst, and B. H. Kang. From data to decisions: helping crop producers build their actionable knowledge. *Critical reviews in plant sciences*, 36(2):71–88, 2017.

L. Evans and R. Fischer. Yield potential: its definition, measurement, and significance. *Crop science*, 39 (6):1544–1551, 1999.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707, Aug. 2015.

N. Galichet. *Contributions to multi-armed bandits: Risk-awareness and sub-sampling for linear contextual bandits*. PhD thesis, 2015.

N. Galichet, M. Sebag, and O. Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, 2013.

Z. Gao, Y. Han, Z. Ren, and Z. Zhou. Batched multi-armed bandits problem. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 501–511, 2019.

A. Garivier and E. Moulines. On upper-confidence bound policies for switching bandit problems. In *Algorithmic Learning Theory - 22nd International Conference, ALT 2011, Espoo, Finland, October 5-7, 2011. Proceedings*, 2011.

A. Garivier, P. Ménard, and G. Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Math. Oper. Res.*, 44:377–399, 2019.

R. Gautron, O.-A. Maillard, P. Preux, M. Corbeels, and R. Sabbadin. Reinforcement learning for crop management support: Review, prospects and challenges. *Computers and Electronics in Agriculture*, 200: 107182, 2022a.

R. Gautron, E. J. Padrón, P. Preux, J. Bigot, O.-A. Maillard, and D. Emukpere. gym-DSSAT: a crop model turned into a Reinforcement Learning environment. Research Report RR-9460, Inria Lille, July 2022b.

H. Hadiji and G. Stoltz. Adaptation to the range in $k$-armed bandits. *CoRR*, 2020.

J. H. Halton. Algorithm 247: Radical-inverse quasi-random point sequence. *Commun. ACM*, 7(12), Dec. 1964.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Z. Hochman and P. Carberry. Emerging consensus on desirable characteristics of tools to support farmers' management of climate risk in australia. *Agricultural Systems*, 104(6):441–450, 2011.

M. J. Holland and E. M. Haress. Learning with cvar-based feedback under potentially heavy tails, 2020.

J. Honda and A. Takemura. An Asymptotically Optimal Bandit Algorithm for Bounded Support Models. In *Proceedings of the 23rd Annual Conference on Learning Theory*, 2010.

J. Honda and A. Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Mach. Learn.*, 2011.

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

G. Hoogenboom, C. Porter, K. Boote, V. Shelia, P. Wilkens, U. Singh, J. White, S. Asseng, J. Lizaso, L. Moreno, et al. The dssat crop modeling ecosystem. *Advances in crop modelling for a sustainable agriculture*, pages 173–216, 2019.

S. Ito, T. Tsuchiya, and J. Honda. Adversarially robust multi-armed bandit algorithm with variance-dependent regret bounds. In P. Loh and M. Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 1421–1422. PMLR, 2022.

T. Jin, J. Tang, P. Xu, K. Huang, X. Xiao, and Q. Gu. Almost optimal anytime algorithm for batched multi-armed bandits. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5065–5073. PMLR, 2021.

M. Jourdan, R. Degenne, D. Baudry, R. de Heide, and E. Kaufmann. Top two algorithms revisited. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

C. Kalkanli and A. Ozgur. Batched thompson sampling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2021.

E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory - 23rd International Conference, ALT*, 2012.

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

R. Krishnamurthy and A. Gopalan. On slowly-varying non-stationary bandits. *CoRR*, abs/2110.12916, 2021.

B. Kveton, C. Szepesvari, M. Ghavamzadeh, and C. Boutilier. Perturbed-history exploration in stochastic multi-armed bandits. *arXiv preprint arXiv:1902.10089*, 2019a.

B. Kveton, C. Szepesvari, S. Vaswani, Z. Wen, T. Lattimore, and M. Ghavamzadeh. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 3601–3610. PMLR, 2019b.

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

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

A. W. Ledford and J. A. Tawn. Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2), May 1997.

E. Leurent. *Safe and Efficient Reinforcement Learning for Behavioural Planning in Autonomous Driving*. Theses, Université de Lille, Oct. 2020.

F. Liu, J. Lee, and N. Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. *arXiv preprint arXiv:1711.03539*, 2017.

O. Maillard. Robust risk-averse stochastic multi-armed bandits. In *Algorithmic Learning Theory - 24th International Conference, ALT*, 2013.

B. B. Mandelbrot. The variation of certain speculative prices. In *Fractals and scaling in finance*. Springer, 1997.

A. G. Manegueu, A. Carpentier, and Y. Yu. Generalized non-stationary bandits. *CoRR*, abs/2102.00725, 2021.

H. Markowitz. Portfolio selection*. *The Journal of Finance*, 7(1):77–91, 1952.

P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *Annals of Probability*, 18, 1990.

R. L. McCown. Changing systems for supporting farmers' decisions: problems, paradigms, and prospects. *Agricultural systems*, 74(1):179–220, 2002.

A. Mendelson, M. Zuluaga, B. Hutton, and S. Ourselin. What is the distribution of the number of unique original items in a bootstrap sample? *CoRR*, abs/1602.05822, 2016.

R. Nishihara, D. Lopez-Paz, and L. Bottou. No regret bound for extreme bandits. In *Artificial Intelligence and Statistics*, pages 259–267. PMLR, 2016.

I. Osband and B. V. Roy. Bootstrapped thompson sampling and deep exploration. *CoRR*, abs/1507.00300, 2015.

Q. Paris. The return of von liebig's "law of the minimum". *Agronomy Journal*, 84(6):1040–1046, 1992.

V. Perchet, P. Rigollet, S. Chassang, and E. Snowberg. Batched bandit problems. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, page 1456. JMLR.org, 2015.

F. Pesquerel, H. Saber, and O. Maillard. Stochastic bandits with groups of similar arms. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 19461–19472, 2021.

L. A. Prashanth, P. Krishna, Jagannathan, and R. K. Kolla. Concentration bounds for cvar estimation: The cases of light-tailed and heavy-tailed distributions. In *International Conference on Machine Learning*, 2020.

V. Raj and S. Kalyani. Taming non-stationary bandits: A bayesian approach. *arXiv preprint arXiv:1707.09727*, 2017.

B. Rennie and A. Dobson. On stirling numbers of the second kind. *Journal of Combinatorial Theory*, 7(2), 1969.

C. W. Richardson and D. A. Wright. Wgen: A model for generating daily weather variables. *ARS (USA)*, 1984.

C. Riou and J. Honda. Bandit algorithms based on thompson sampling for bounded reward distributions. In *Algorithmic Learning Theory - 31st International Conference (ALT) 2012*, 2020.

R. T. Rockafellar, S. Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

D. B. Rubin. The bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981.

A. Sani, A. Lazaric, and R. Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, 2012.

M. Seurin, F. Strub, P. Preux, and O. Pietquin. A Machine of Few Words Interactive Speaker Recognition with Reinforcement Learning. In *Conference of the International Speech Communication Association (INTERSPEECH)*, Shanghai, China, Oct. 2020.

J. Seznec, P. Menard, A. Lazaric, and M. Valko. A single algorithm for both restless and rested rotting bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3784–3794. PMLR, 2020.

W. F. Sharpe. The sharpe ratio. *The Journal of Portfolio Management*, 21(1):49–58, 1994.

B. Shiferaw, B. M. Prasanna, J. Hellin, and M. Bänziger. Crops that feed the world 6. past successes and future challenges to the role played by maize in global food security. *Food security*, 3(3):307–327, 2011.

I. Sobol. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4), 1967.

M. J. Streeter and S. F. Smith. An asymptotically optimal algorithm for the max k-armed bandit problem. In *AAAI*, pages 135–142, 2006a.

M. J. Streeter and S. F. Smith. A simple distribution-free approach to the max k-armed bandit problem. In *International Conference on Principles and Practice of Constraint Programming*, pages 560–574. Springer, 2006b.

J. Suk and S. Kpotufe. Tracking most significant arm switches in bandits. In P. Loh and M. Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 2160–2182. PMLR, 2022.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2018.

B. Szorenyi, R. Busa-Fekete, P. Weng, and E. Hüllermeier. Qualitative multi-armed bandits: A quantile-based approach. In *International Conference on Machine Learning*, 2015.

A. Tamkin, R. Keramati, C. Dann, and E. Brunskill. Distributionally-aware exploration for cvar bandits. In *NeurIPS 2019 Workshop on Safety and Robustness in Decision Making; RLDM 2019*, 2020.

P. Thomas and E. Learned-Miller. Concentration inequalities for conditional value at risk. In *International Conference on Machine Learning*, 2019.

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

M. Tollenaar and E. Lee. Yield potential, yield stability and stress tolerance in maize. *Field crops research*, 75(2-3):161–169, 2002.

F. Trovo, S. Paladino, M. Restelli, and N. Gatti. Sliding-window thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*, 68:311–364, 2020.

A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.

S. Vakili and Q. Zhao. Mean-variance and value at risk in multi-armed bandit problems. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2015.

S. Vakili and Q. Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE J. Sel. Top. Signal Process.*, 10:1093–1111, 2016.

C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

C. Wang, Y. Yu, B. Hao, and G. Cheng. Residual bootstrap exploration for bandit algorithms. *CoRR*, abs/2002.08436, 2020.

P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25: 287–298, 1988.

M. Zhang and C. S. Ong. Quantile bandits for best arms identification. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12513–12523. PMLR, 18–24 Jul 2021.

Q. Zhu and V. Tan. Thompson sampling algorithms for mean-variance bandits. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

A. Zimin, R. Ibsen-Jensen, and K. Chatterjee. Generalized risk-aversion in stochastic multi-armed bandits. *CoRR*, abs/1405.0833, 2014.

J. Zimmert and Y. Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *J. Mach. Learn. Res.*, 22:28:1–28:49, 2021.