Contributions to Multi-Armed Bandits : Risk-Awareness and Sub-Sampling for Linear Contextual Bandits

Nicolas Galichet

UNIVERSITÉ PARIS SUD

LABORATOIRE DE RECHERCHE EN INFORMATIQUE

UNIVERSITÉ PARIS-SUD

ECOLE DOCTORALE D'INFORMATIQUE
LABORATOIRE DE RECHERCHE EN INFORMATIQUE

DISCIPLINE : INFORMATIQUE

THÈSE DE DOCTORAT

Soutenue le 28 septembre 2015 par

Nicolas Galichet

Contributions to Multi-Armed Bandits: Risk-Awareness and Sub-Sampling for Linear Contextual Bandits

Directrice de thèse :	Michèle Sebag	Directrice de recherche CNRS
Co-directeur de thèse :	Odalric-Ambrym Maillard	Chargé de recherche INRIA Saclay

Composition du jury :

Président du jury :	Damien Ernst	Professeur à l'Université de Liège (Belgique)
Rapporteurs :	François Laviolette	Professeur titulaire à l'Université Laval (Québec)
	Olivier Pietquin	Professeur à l'Université Lille 1
Examinateurs :	Damien Ernst	Professeur à l'Université de Liège (Belgique)
	Yannis Manoussakis	Professeur à l'Université Paris-Sud

Contents

List of Algorithms

List of Figures

Abstract

This thesis focuses on sequential decision making in unknown environment, and more particularly on the Multi-Armed Bandit (MAB) setting, defined by Lai and Robbins in the 50s (Robbins, 1952; Lai and Robbins, 1985). During the last decade, many theoretical and algorithmic studies have been aimed at the exploration vs exploitation tradeoff at the core of MABs, where Exploitation is biased toward the best options visited so far while Exploration is biased toward options rarely visited, to enforce the discovery of the true best choices. MAB applications range from medicine (the elicitation of the best prescriptions) to e-commerce (recommendations, advertisements) and optimal policies (e.g., in the energy domain).

The contributions presented in this dissertation tackle the exploration vs exploitation dilemma under two angles.

The first contribution is centered on risk avoidance. Exploration in unknown environments often has adverse effects: for instance exploratory trajectories of a robot can entail physical damages for the robot or its environment. We thus define the exploration vs exploitation vs safety (EES) tradeoff, and propose three new algorithms addressing the EES dilemma. Firstly and under strong assumptions, the MIN algorithm provides a robust behavior with guarantees of logarithmic regret, matching the state of the art with a high robustness w.r.t. hyper-parameter setting (as opposed to, e.g. UCB (Auer et al., 2002)). Secondly, the MARAB algorithm aims at optimizing the cumulative "Conditional Value at Risk" (CVaR) rewards, originated from the economics domain, with excellent empirical performances compared to (Sani et al., 2012a), though without any theoretical guarantees. Finally, the MARABOUT algorithm modifies the CVaR estimation and yields both theoretical guarantees and a good empirical behavior.

The second contribution concerns the contextual bandit setting, where additional informations are provided to support the decision making, such as the user details in the content recommendation domain, or the patient history in the medical domain. The study focuses on how to make a choice between two arms with different numbers of samples. Traditionally, a confidence region is derived for each arm based on the associated samples, and the 'Optimism in front of the unknown' principle implements the choice of the arm with maximal upper confidence bound. An alternative, pio-

neered by (Baransi et al., 2014), and called BESA, proceeds instead by subsampling without replacement the larger sample set.
In this framework, we designed a contextual bandit algorithm based on sub-sampling without replacement, relaxing the (unrealistic) assumption that all arm reward distributions rely on the same parameter. The CL-BESA algorithm yields both theoretical guarantees of logarithmic regret and good empirical behavior.

Remerciements

La thèse est une longue épreuve aussi enrichissante que difficile et que l'on ne peut aborder seul. Je réalise au moment de finaliser ce manuscrit la chance que j'ai eu d'être bien entouré et j'aimerais remercier toutes les personnes sans lesquelles cette thèse n'aurait pu être menée à bien.

- Merci aux membres de mon jury, qui m'ont fait l'honneur de se déplacer pour examiner mon travail. Merci à Damien ERNST et Yannis MANNOUSSAKIS. Merci aux rapporteurs Olivier PIETQUIN et François LAVIOLETTE pour avoir pris le temps de réviser mon manuscrit.
- Merci à mes directeurs de thèse, Michèle et Odalric-Ambrym, pour leur disponibilité, leur patience, leurs conseils, leur aide et leur soutien indéfectibles.
- Merci à Olivier pour son aide précieuse à un moment clé de mon travail.
- Merci à Sophie LAPLANTE pour m'avoir orienté vers la recherche.
- Merci à tous les membres de l'équipe TAO d'hier et aujourd'hui pour les discussions (parfois scientifiques) et l'excellente ambiance. Merci à (par ordre alphabétique) : Adrien, Alexandre, Antoine, Asma, Aurélien, Basile, François, François-Michel, Gaétan, Jean-Baptiste, Jean-Joseph, Jean-Marc, Jérémie, Jérémy, Jialin, Karima, Lovro, Ludovic, Manuel, Marie-Liesse, Mouadh, Nacim, Olga, Riad, Sandra 1.0, Sandra 2.0, Sébastien, Simon, Sourava, Thomas, Vincent, Wassim, Weijia, Yoann. Merci à ceux que j'ai nécessairement oublié. Bonne chance et bon vent à ceux qui vont soutenir bientôt.
- Merci à FZ notre mascotte moustachue alsacienne et un clin d'oeil tout particulier aux membres de son bureau, d'hier et aujourd'hui. Que son esprit perdure pour des générations de scientifiques.
- Merci et longue vie au PBA Crew !
- Merci à ma famille pour son soutien sans faille. Merci à mes parents, mes grands-parents, mes oncles, à mes soeurs Alice et Elsa, à mon frère Louis et à Marine.

Chapter 1

Introduction

Contents

Multi-armed bandits (MAB) (Robbins, 1952; Lai and Robbins, 1985) is a simple, generic however rich framework constituting the theoretical and algorithmic background of the work exposed in the present document.

1.1 Motivations

Originally, the term *bandit* (Thompson, 1933) refers to the casino slot machine as the MAB problem can be interpreted as an optimal playing strategy problem: a player enters a casino and is proposed a set of options or *bandit arms* with unknow associated payoffs. Given a fixed number of trials or *time horizon*, his or her goal is to select arms in order to collect the highest amount of money possible.
While present under a wide variety of different settings, each of them emphazing specific learning aspects, the MAB framework is regarded as one of the most fundamental formalizations of the sequential decision making problem, and in particular an illustration of the *Exploration vs. Exploitation* dilemma:

- **Exploration** : The agent is assumed to have no prior information about the machine payoffs. This assumption implies the need for the agent to play, i.e. to *explore*, arms that were not or rarely tried.

- **Exploitation** : In order to gather the maximal amount of reward, the agent should play as often as possible, or *exploit*, the arms with estimated best payoffs.

In this respect, the MAB framework fundamentally differs from the statistical evaluation of the arm's reward distribution. The goal is to *discriminate* the best options along the play, evaluation and interaction being done *simultaneously*.
In particular, one focuses on the identification of the promising arms with the minimal number of samples, rather than on a precise estimation of every arm distributions. The latter approach indeed provides useless information (tight estimates of suboptimal arms) at the cost of a high number of trials of these arms and lower cumulative rewards.
As said, the generic MAB framework allows many practical problems to be formalized and addressed using solutions from the bandit literature. Formally, any situation involving an agent repeatedly facing a choice between a given number of options with *a priori* unknown returns can be seen as a MAB problem. As this chapter aims to demonstrate, such situations occur frequently, explaining the rapidly growing body of literature in the MAB field. The scientific interest is also motivated by an accurate tradeoff between the expressivity and simplicity of the model allowing both a realistic and practically useful formalization and an in-depth theoretical study.
This chapter will present a few examples of MAB applications, starting with some definitions.

1.2 First problem statement

Standard notations are defined to introduce more formally some bandit applications. The MAB setting considers a finite set of K actions or *bandit arms*. For $i \in \{1 \dots K\}$, the i-th arm is associated with a reward distribution ν_i with expectation $\mu_i \overset{\text{def}}{=} \mathbb{E}[\nu_i]$. At a given time step t, the agent choose, based on the previous observations, an arm $I_t \in \{1 \dots K\}$ and observes an instantaneous reward $Y_t \sim \nu_{I_t}$.
Letting T denote a finite time horizon ($T \in \mathbb{N}^\star$), the agent's goal is to maximize the cumulative sum of instantaneous rewards:

$$S = \sum_{t=1}^{T} Y_t \tag{1.1}$$

1.3 Applications

This section presents some applications presented in the multi-armed bandit literature. This (far from exhaustive) list aims to illustrate the wide variety of successful applications of the MAB framework and motivate its study in the present document. These example applications will also serve to illustrate our contributions, as discussed later on.

1.3.1 Clinical testing

Historically the introduction of the multi-armed bandit framework is due to Thompson, firstly considering a two-armed bandit problem (Thompson, 1933), then an arbitrary number of arms (Thompson, 1935). The studied problem considers a particular disease, with affected patients sequentially arriving. A (finite) set of drugs is available but their efficiency is unknown.
For each patient arriving (at time t), a drug is selected (I_t) and a reward (Y_t) is gathered depending on the efficiency of the treatment: $Y_t = 1$ if the treatment was effective, and 0 otherwise. In this clinical context, the relevance of the MAB setting is clear: as health and life are involved, it is vital to focus as quickly as possible on the best treatment rather than precisely determining the efficiency of every drug.

1.3.2 Algorithm selection

MAB has been used for Algorithm Selection in (Gagliolo and Schmidhuber, 2010). Let us consider a binary **decision** problem (e.g., SAT: for each problem instance, the decision problem is to determine whether this instance is satisfiable). Given a set $\mathscr{B} = \{b_1, \ldots, b_M\}$ of instances and a set $\mathscr{A} = \{a_1, \ldots, a_N\}$ of algorithms/solvers[1], the goal is to iteratively pick an algorithm in order to minimize the time-to-solution.
By denoting l_{I_t} the time spent (loss) by the algorithm a_{I_t} on the instance b_t presented at time t, a first formulation of this problem as a MAB problem is to consider the set of algorithms as the set of arms and to define the instantaneous reward as $Y_t = -l_{I_t}$. This approach aims at the "single-best" algorithm $a^\star \in \mathscr{A}$ getting the best performance for *every* instance b_m.
A refined setting considers a different set of arms, defined as K *time allocators* TA_j, mapping the history of collected performances of algorithms to a share $s \in [0,1]^N$ with $\sum_{i=1}^{N} s_i = 1$. The purpose of these time allocators is to run every algorithm in $\mathscr{A}$ in parallel on a single machine dividing the computational resource according to the

[1] The set $\mathscr{A}$ may include a same algorithm/solver with different hyper-parameter settings.

share s. On a given amount of time τ, $s_i\tau$ is dedicated to algorithm i and the runtime associated with the TA_j is $\min_i\{\frac{t_i}{s_i}\}$, with t_i runtime of algorithm i.

1.3.3 Motor-Task Selection for Brain Computer Interfaces

Brain-Computer Interfaces (BCI) are meant to provide a command law (e.g., of the computer or a wheeling chair) based the direct measurement of the brain activity. These measurements rely on electroencephalography (EEG) or magnetoencephalography (MEG) technologies. One of the most promising applications of the BCI interfaces is to provide severely handicapped patients with controllable tools (e.g. robotic arm (Hochberg et al., 2012)).

Sensori-motor rhythms (SMR) are expressed and captured (via EEG) during the execution or imagination of a specific movement. SMRs thereby define a efficient way of communicating through BCI. (Fruitet et al., 2012) considers the training of such a SMR-based BCI to brain-control a single button when a given (imagined) movement is detected.

The problem can be decomposed into two steps. First a motor task has to be selected. Based on this task, a classifier is learned in order to discriminate periods of (imaginative) execution of the task and resting periods. One of the main challenges of this setting is the high variability of best movements depending on the studied patient. The purpose of the approach thus is to discriminate the best task *while* constituting the classifier training sets.

The problem is defined by a number K of tasks and a total number of rounds T. Note that these two quantities must be small in order to propose reasonable training sessions: long (large N) or complex (large K) sessions are to be avoided. In a bandit context, task are arms (in practice images to be displayed to ask the user to imagine a given task) and T is the time horizon. Instantaneous rewards are Y_t are defined by the empirical classification rate associated to select task $I_t \in \{1,\ldots,K\}$.

The gain due to the MAB setting is explained by the early focus on *useful* discriminative tasks, through interaction with the user. Compared to the standard procedure of *uniform* exploration, each task being presented an equal number of times, the approach proposed by (Fruitet et al., 2012) and based on the MAB `UCB` algorithm (presented in chapter 4, (Auer et al., 2002)) showed benefits both in terms of classification rates (up to 18%) and training times (lowered by up to 50%).

1.3.4 Web applications

Document ranking

The MAB setting can be used to rank documents (Radlinski et al., 2008) and specifically web pages. The goal is to improve a web search engine performance by increasing the probability that users click on top-ranked results. This goal is formalized as follows. A fixed search query and a set of documents $\mathscr{D} = \{d_1 \dots d_n\}$ is considered. At time t, an user u_t is proposed an ordered set of documents $B_t = (b_1(t), \dots, b_k(t))$ and clicks on document d_i with probability p_{t_i}[2].
The goal is to avoid as much as possible the *abandonment* phenomenon: no relevant document is presented and the request is unsatisfied.The instantaneous reward $Y_t = 1$ if the user selects (clicks on) a document $d_i \in B_t$ and $Y_t = 0$ otherwise.
A first naive adaptation to the bandit framework would be to associate an arm to every possible ordered $k-$subset of $\mathscr{D}$, leading to MAB instance with $\frac{n!}{(n-k)!}$ number of arms. The authors propose a more efficient method termed `Ranked Bandits Algorithm` (RBA). The procedure relies on an preexisting MAB algorithm and runs k MAB instances (one per rank), each of these instances presenting n arms (one per document). The i-th MAB instance is in charge of choosing the document at rank i. In the case where the i-th MAB instance proposes an already selected document j for $j < i$, the choice is saved but another unpresented document is uniformly picked and presented. After user clicks on document $b_i(t)$, MAB instances receive a 1 (resp. 0) reward if $b_i(t)$ was selected by the instance (resp. dismissed).
A key feature of this conception is its ability to produce diverse document rankings. This property is desirable in such a context as methods assuming independent document relevances tend to output rankings with redundant documents, lowering the probability to present a relevant document to the user.

Content recommendation

MAB also is a natural framework for the recommendation and personalization of content (articles, advertisements, songs, videos ...) over Internet, which has been widely and intensively studed in the last decade (see for instance (Li et al., 2010; Pandey and Olston, 2006; Babaioff et al., 2009; Agarwal et al., 2009; Kohli et al., 2013)), boosted by the need of many companies to assess the quality of a web-service and/or estimate the amount of advertising revenue.
The news article recommendation can be stated as follows. Visitors access a website and are proposed a selection of articles. The goal is to select relevant articles in order to maximize the probability to display at least an article of interest to the visitor. A

[2]Probability p_{t_i} is conditioned by the user not clicking on documents ranked before document i.

standard measure of performance in such problem is the **Click-through rate** (CTR) defined as the ratio $\frac{\#clicks}{\#trials}$.
A straightforward bandit adaptation would be to consider each possible pool of articles as a bandit arm i and the reward $Y_t = 1$ if an article is clicked at time t and 0 otherwise, leading to a setting close to the clinical testing one; this formalization could also be easily adapted to any other recommendation problem as it does not require any specific assumption regarding the content.
Nevertheless, a key feature of this setting is the important quantity of available information which can be used to efficiently improve the recommendations. Indeed, Web services often benefit from a vast amount of data (regarding the user's interests or similar profiles) collected through over a long time and a large number of users.
This leads to the definition of the **Contextual bandit** framework, where side information is revealed to support the MAB decision process (Li et al., 2010). Contextual bandits will be further considered in Chapter 8, presenting the second contributions of this thesis.
Another MAB extension motivated by this application is proposed by Gentile et al. (2014), who maintain user clusters in parallel of the recommendations. The underlying intuition is that similar users would have similar preferences, and information of similarity should be shared among users to improve recommendation.

1.3.5 Monte-Carlo Tree Search

The MAB setting has been used to guide exploration in a tree structure, enabling to extend their application to the selection of a (quasi-)optimal action in the broader context of a Markov Decision Process (MDP).
MDP is the main formalization used in Reinforcement Learning (RL, (Sutton and Barto, 1998; Szepesvári, 2010)), where an agent evolving in an unknown environment learns how to perform well *while interacting* with it. MDP is formalized by a tuple $\left(\mathcal{S}, \mathcal{A}, P, R, \gamma\right)$ where:

1. $\mathcal{S}$ (resp. $\mathcal{A}$) is the **state** (resp. **action**) set.
2. $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is the transition probability
3. $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the instantaneous reward function
4. $\gamma \in [0,1]$ is the discount factor

The goal is to find a (deterministic) policy $\pi : \mathcal{S} \to \mathcal{A}$, maximizing the cumulative reward over T (possibly $T = \infty$) V^π defined by

$$V^\pi(s) \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R(s_t, a_t, a_{t+1}) | s_0 = s, a_t = \pi(s_t)\right]$$

Note that a MAB defines an MDP problem with a single state (thus without any dynamics), where the action set $\mathscr{A}$ is the (possibly large) set of arms.
The MAB setting was extended to the standard MDP setting, through a tree-structured exploration of the state space (Kocsis and Szepesvári, 2006b; Gelly et al., 2006). The approach makes the only assumption that a generative model of the MDP is available, allowing the sampling of states, actions and rewards; the best known algorithm to do so extends the Upper Confidence Bound (UCB, already mentioned (Auer et al., 2002)) to tree-structured search, defining the *Upper Confidence bound applied to Trees* (UCT) algorithm, where visited states are represented by tree nodes, and transitions by edges. Each node is associated with the cumulative rewards obtained by visiting the corresponding state.
Starting from the tree-root, UCT traverses the search tree in the following top-down way:

1. Considering the current state (tree node) s_i, the associated available actions and estimated rewards, the next action a_i is selected thanks to a bandit algorithm (UCB).

2. Next state s_{i+1} is determined from (s_i, a_i) using the generative model.

3. If s_{i+1} is a leaf, and the limit depth is not reached:

   - a random action a_{i+1} is selected and the arrival s_{i+2} state is provided by the generative model. The edge (a_{i+1}, s_{i+2}) is attached to s_{i+1};
   - a roll-out policy (e.g., uniformly random) selects next actions until a termination condition is met (usually, when reaching a fixed horizon T);
   - The cumulative reward R associated with the whole state-action sequence is computed;
   - Information associated to visited nodes is updated thanks to R.

4. Otherwise, goto 1.

This procedure is repeated until reaching the prescribed number N (budget) of time step; ultimately, it returns the action most visited at the top of the tree.
The MAB step achieves the iterative selection of actions (descendant nodes) at each stage of the search tree, leading to a search tree biased toward the most promising region of the search space. This feature offers a important double benefit, critical if the MDP problem involves large state and/or action spaces: it implies an efficient allocation of the computational resources and a reasonable memory usage.
(Kocsis and Szepesvári, 2006b) was the seminal work leading to the *Monte-Carlo Tree Search* (MCTS) family of algorithms. These algorithms received a considerable

scientific attention in the last decade due to their efficiency and genericity. Notable successes have been obtained in Computer Games, famous examples are Computer Go (Gelly et al., 2012) and Computer Hex (Hayward, 2009); in the latter game, artificial players now defeat human experts. The approach has also been applied to active learning (Rolet et al., 2009) or feature selection (Gaudel and Sebag, 2010) (among many others).
The interested reader is referred to (Browne et al., 2012) for an extensive survey of MCTS approaches and applications.

1.4 Risk consideration in Multi-Armed Bandits

As said, Multi-armed Bandits aim to discriminate and pull, *as fast as possible*, the arms with highest quality. This raises the central question of the relevant quality criterion. Most authors in the MAB community consider the maximization of the expected payoff, where the goal is to identify and pull the arm with highest $\mu_i = \mathbb{E}[\nu_i]$ value. In some situations however, this criterion appears to be inappropriate. For instance in the clinical testing problem, let us one consider the following situation:

- A treatment A providing good results on average with rare lethal side effects.
- A treatment B providing slightly lesser results on average without any known side effect.

In such a situation, one would like to take into account another quality criterion, enforcing a more cautious exploration, at the expense of a potentially lower cumulative reward.
The general situation is even more critical as a large number of MAB algorithms are **optimistic**: they maintain a confidence region for the arm expected rewards and select the one with highest possible reward according to the information gathered so far. In the case of high variance on the distribution of the best arms, this would lead to a high number of pulls of arm with potentially low rewards.
In essence, in such a situation the usual tradeoff between exploration and exploitation must be extended to consider also *decision safety*. The exploration vs exploitation vs safety tradeoff will be further considered in Chapter 7, presenting the first contributions of this thesis.

Chapter 2

Overview of the contributions

The presented thesis proposes two main contributions to the Multi-Armed Bandits state of the art, summarized in this chapter.

2.1 Risk-Aware Multi-armed Bandit

Considering K arms with unknown reward distributions ν_i, the mainstream MAB research is interested in maximizing the expected cumulative payoffs obtained after T trials. Equivalently, one is interested with minimizing the **cumulative regret**, sum of losses encountered comparatively to an oracle player. This problem has been extensively studied, both theoretically and empirically, leading to algorithms with proved optimal regret bound.
The limitation of the approach, in safety-critical situations, is to only consider the associated **expected** payoff μ_i of each arm. The associated theoretical guarantees require a sufficiently precise estimation of the arms' payoffs, implying exploration efforts. In the case of high variances associated with the (best) arms, this leads to occasionally obtain extremely low payoffs. In many critical situations, these low payoffs must be avoided as they mean for instance the loss of expensive gears, the deterioration of patient's health or even the loss of human lives.
The work presented in this manuscript propose mainly three contributions to the Risk-Aversion in the Multi-Armed Bandit setting, presented in Chapter 6:

- Two criteria have been proposed for taking risk into account: the minimal value associated with each arm, and its Conditionnal Value at Risk (CVaR). The first one is the lower bound of the distribution support, the second is defined as the reward obtained in the α% worst cases, α being an user-defined parameter.

  These criteria lead to designing three new algorithms: `MIN`, `MARAB` and `MARABOUT`.

- The regret rates of `MIN` (Section 6.2) and `MARABOUT` (Section 6.6) have been theoretically studied.
- An extensive empirical study based on artificial MAB instances has been proposed, showing the benefits of `MIN` and `MARAB` (Section 6.5)
- A real-world benchmark of energy allocation has been considered, confirming the gain obtained with risk aversion on moderate time horizons. (Section 6.5.4).

2.2 Sub-Sampling for Contextual Linear Bandits

The second contribution presented in the manuscript concerns the contextual linear bandit setting.
Generalizing the stochastic MAB setting, Contextual Bandits consider K arms with unknown reward distributions ν_i, where at each time t, a contextual information $X_t \in \mathbb{R}^d$ is revealed to the learner and contributes to the computation of the instantaneous reward. The Contextual Linear model further assumes a linear dependency between X_t and the instantaneous reward Y_t. In the so-called **disjoint model**, one assumes the existence of K vectors θ_i associated with each arm so that:

$$Y_t = \langle X_t, \theta_{I_t} \rangle + \eta_t$$

with I_t the chosen arm at time t and η_t an additive noise.
The contribution is to introduce the sub-sampling strategy first pioneered by (Baransi et al., 2014) in this setting. Sub-sampling bases the choice between two arms on equal quantities of information. Considering two arms denoted a and b with associated sampling sets (rewards and contexts) of respective sizes n_a and n_b, and assuming without loss of generality that $n_a \leqslant n_b$, n_a samples are uniformly drawn without replacement from the sampling set of arm b and the parameters θ_a and θ_b are estimated based each on a set of same size n_a.
Although counterintuitive, this scheme has shown striking properties in the stochastic case, motivating its extension. In this regard, the contributions are

- The redefinition of an algorithm, termed `CL-BESA`, for the contextual case (Section 7.2).
- The non-trivial derivation of a regret bound for a refined notion of regret (Section 7.3).
- An experimental validation on synthetic problem, illustrating the good empirical properties of the approach (Section 7.4).

The manuscript is organized as follows. The formal background and algorithms for the stochastic and contextual bandits are respectively presented in Chapters 3 and 4. Chapter 5 focuses on the notion of risk; it describes and discusses the state of the art concerned with risk-averse bandit algorithms. Chapter 6 presents our first contributions, and details algorithms based on the conditional value at risk, together with their theoretical and empirical analysis. Chapter 7 presents our second contributions, the extension of the sub-sampling technique to Contextual Linear Bandits. Finally, Chapter 8 concludes this manuscript by discussing the presented contributions, and describing future research directions.

Part I

State of the Art

Chapter 3

Multi-armed bandits: Formal background

Contents

This chapter introduces the formal background of the presented work, that is, the *stochastic multi-armed bandit problem* (MAB). After some definitions and notations, the different goals tackled by MAB settings are presented together with the associated evaluation criteria or loss functions known as *regret*. The chapter last presents the contextual MAB framework, which will be further considered in the manuscript. While there exists many other variants of the MAB setting, the chapter does not pretend to exhaustivity due to the fast growing literature devoted to the MAB settings and their extensions. The interested reader is referred to (Bubeck and Cesa-Bianchi, 2012) for a more comprehensive presentation of the related literature.

3.1 Position of the problem

Multi-armed Bandits (MAB), also referred to as K-armed bandit problem, consider an agent facing K independent actions, or bandit arms. Each arm is associated with an unknown, bounded, reward distribution. In each time step, the agent must select one of these arms on the basis of her current information about the arms and the associated distributions. Two main settings are studied:

- The best arm identification, where the agent is provided with a fixed budget (number T of time steps) and must after T trials decide for the arm that will be selected ever after. This setting is related to multipe hypothesis testing: the agent must decide for the best arm on the basis of the available evidence, and must select the evidence in order to do so with best confidence.

- The cumulative reward maximization, where in each time step the agent selects an arm and gathers the associated reward, with the goal of maximizing the sum of rewards gathered along time. In this second setting, the agent might know in advance the time budget (number T of time steps, also called *time horizon*), or might consider the anytime setting where the agent performance is assessed as a function of T, going to infinity.
  This setting models the situation of a gambler facing a row of slot machines (also known as one-armed bandit), and deciding which arm he should pull in each time step, in order to maximize his cumulative gains.

It is seen that both settings define a sequential decision problem[1] with finite or infinite time horizon, where the agent performance either is the sum of the rewards gathered in each time step (cumulative reward maximization), or the reward gathered in the final state (best arm identification). In particular, the Exploration vs Exploitation (EvE) trade-off is at the core of both MAB and RL settings:

- **Exploitation**: In order to maximize its performance, the agent should focus on the best options seen so far;

- **Exploration**: Still, focusing on the best options seen so far might miss some better options, possibly discarded due to unlucky trials[2].

[1] Accordingly, the MAB problem could be seen as a particular case of reinforcement learning (RL) problem (Sutton and Barto, 1998). A first difference is that MABs involve a single state whereas the RL decision takes into consideration the current state of the agent.

[2] A second difference between RL and MAB, and related to the fact that MABs involve a single state, is that the MAB agent must simply explore the other options while the RL agent must *plan to explore* (Roy and Wen, 2014). Typically, the RL agent must revisit known states as these can lead to other states, which themselves need be explored. A comparative discussion between the specifics of the RL and MAB settings is however outside the scope of this manuscript.

As mentioned in Chapter 1, both MAB settings are relevant to a vast number of applications, all related to optimal decision making under uncertainty and sequential design of experiments (Robbins, 1952), where the agent must make decisions with the antagonistic goals of optimizing its performance on the basis of the available evidence, and gathering more evidence.

3.1.1 Notations

Let us introduce the notations which will be used in the rest of the document.

K denotes the number of arms or options ($K \in \mathbb{N}^\star$);

ν_i denotes the (unknown) bounded reward distribution associated with the i-arm (for $i = 1 \dots K$);

μ_i denotes the expectation of ν_i ($\mu_i \stackrel{\text{def}}{=} \mathbb{E}[\nu_i]$);

$\mu^\star$ is the maximum expectation taken over all arms ($\mu^\star \stackrel{\text{def}}{=} \max_{i=1\dots K} \mu_i$);

$i^\star$ is an index (possibly not unique) such that $\mu_{i^\star} = \mu^\star$ ($i^\star$ in $1 \dots K$);

Δ_i is the optimality gap of the i-th arm ($\Delta_i \stackrel{\text{def}}{=} \mu^\star - \mu_i$);

T denotes the time horizon ($T \in \mathbb{N}^\star$), which might be finite or infinite;

t denotes the current time step;

I_t is the arm selected at time t;

Y_t is the reward obtained at time t;

$X_{i,t}$ is the t-th selected reward drawn from distribution ν_i;

$N_{i,t}$ is the number of times the i-th arm has been selected up to time t ($N_{i,t} = \sum_{s=1}^{t} \mathbb{1}_{I_t = i}$).

Without loss of generality, it is assumed that each reward distribution ν_i has its support in $[0,1]$; it belongs to the set $\mathcal{M}([0,1])$ of probability measures on $[0,1]$. The main three MAB settings involve:

- Stochastic MAB, where all distributions ν_i are stationary and the reward Y_t gathered at time step t upon selecting arm $i = I_t$ is independently drawn from distribution ν_i.

- Adversarial MAB, where reward Y_t is decided prior to the agent's decision by the adversary (Bubeck and Cesa-Bianchi, 2012). One distinguishes the case where the adversary is independent from the agent's past actions (oblivious adversary), and the case where it depends on the past (non-oblivious adversary). The study of the adversarial setting provides worst-case guarantees about the pessimistic case where the environment plays against the agent (e.g. when considering financial engineering problems[3]);

- Markovian MAB, where arm distributions ν_i evolve according to a on/off Markov process. In Gittins et al. (2011), the distribution only evolves when the arm is selected, and it is frozen otherwise; in restless bandits (Guha et al., 2010), the state of the underlying Markov process controlling the evolution of the arm distributions is only sparsely revealed to the agent. This setting, related to the Partially Observable Markov Decision Process setting, is relevant for applications in wireless scheduling and unmanned aerial vehicle (UAV) routing.

Only the stochastic and contextual MAB frameworks will be considered in the following of the manuscript.

3.2 The stochastic MAB framework

The stochastic MAB framework considers stationary reward distributions. At each time step t, the agent selects an arm $I_t \in \{1 \dots K\}$; the environment draws reward Y_t according to distribution ν_{I_t}, independently from the past, and reveals it to the agent. As said, the MAB setting tackles one out of two goals:

- Maximizing the cumulative gain gathered along time, defined as:

$$\sum_t Y_t$$

- Identify with a budget of T trials, the best arm. In the best arm identification case, the agent selects arm I_T based on the sample $Y_1 \dots Y_{T-1}$ (with $Y_t \sim \nu_i, i = I_t$), thus with the goal of maximizing μ_{I_t}.

The agent strategy (selecting an arm in each time step on the basis of the available evidence) is assessed in terms of *regret* compared to the optimal strategy. In the cumulative gain maximization (respectively in the best arm identification) problem, the optimal strategy, also called oracle strategy, selects in each time step (resp. at time

[3] Although empirical results suggest that the MAB strategies are too conservative, or equivalently, that the environment might be only moderately adversarial in such contexts.

T) the best arm $i^\star$. It is clear that maximizing the agent performance is equivalent to minimizing its regret. One reason why the theoretical analysis considers the regret minimization (as opposed to, e.g., the cumulative gain maximization) in the literature is that the regret in some sense abstracts the difficulty of the problem *per se* and focuses on the quality of the strategy. For more difficult problems, even the oracle performance will be degraded; but the regret analysis only considers how much worse the strategy does, comparatively to the oracle.

3.2.1 Regrets

Three definitions of regret are introduced in the cumulative gain maximization case, respectively referred to as regret, pseudo-regret and empirical regret. In the last two cases, the rewards associated to the oracle strategy are simply set to their expectation, $\mu^\star$.

Definition 3.2.1 (Cumulative regret)**.** *With same notations as above, the **cumulative regret** of the agent at time t is defined as:*

$$R_t \overset{\text{def}}{=} \max_{i\in\{1\dots K\}} \sum_{s=1}^{t} X_{i,s} - \sum_{s=1}^{t} Y_s \tag{3.1}$$

*and the **expected cumulative regret**:*

$$\mathbb{E}_{\nu_1,\dots,\nu_K}[R_t] \overset{\text{def}}{=} \mathbb{E}\left[\max_{i\in\{1\dots K\}} \sum_{s=1}^{t} X_{i,s} - \sum_{s=1}^{t} Y_s\right] \tag{3.2}$$

Definition 3.2.2 (Cumulative pseudo-regret)**.** *With same notations as above, the **pseudo-cumulative regret** of the agent at time t is defined as:*

$$\overline{R}_t \overset{\text{def}}{=} \sum_{s=1}^{t} \left(\mu^\star - \mu_{I_s}\right) = t\mu^\star - \sum_{s=1}^{t} \mu_{I_s} = \sum_{i=1}^{K} \Delta_i N_{i,t} \tag{3.3}$$

Definition 3.2.3 (Empirical cumulative regret)**.** *With same notations as above, the **empirical cumulative regret** of the agent at time t is defined as:*

$$\widehat{R}_t \overset{\text{def}}{=} \sum_{s=1}^{t} \left(\mu^\star - Y_s\right) \tag{3.4}$$

Likewise, the regret defined for the best arm identification problem, called simple regret, measures the difference between the expectation of the true best arm, and the recommended arm:

Definition 3.2.4 (Simple regret)**.** *Letting J_t denote the recommended arm after t trials,*

*the **simple regret** r_t is defined as:*

$$r_t \stackrel{\text{def}}{=} \mu^\star - \mu_{J_t} = \Delta_{J_t} \tag{3.5}$$

The significant difference between both goals, the cumulative gain maximization and the best arm identification, concerns the related EvE tradeoffs, and the price of exploration as manifested in the different regrets (see Stoltz et al. (2011)):

- When considering the maximization of the cumulative gain, the exploration and the exploitation both take place *during the T fixed trials*; the role of exploration is to discriminate the non-optimal arms. The optimal regret rate, that is logarithmic in the budget T after Lai and Robbins (1985), requires that non optimal arms are played at most a logarithmic number of times.

- On the other hand, the best arm identification involves two separated exploration and exploitation phases. The simple regret (eq. 3.5) minimization involves the pure exploitation of a single arm *after* a pure T-step exploration phase, and no loss is encountered during the exploration phase. It is said that the *pure* exploration phase is followed by a single exploitation phase, also called *recommendation.* (Stoltz et al., 2011) show that optimal rates are here conversely obtained for a linear number of suboptimal arm trials.

The regret is studied in the literature considering two main settings: the distribution-free (where no assumption is done about e.g. the moments of the underlying distributions) and the distribution-dependent settings. Typically, in the distribution-free setting, the loss incurred by selecting n times the i-th arm cannot be higher than $n\Delta_i$, whereas the (distribution-dependent) bounds on the regret depend on $1/\Delta_i$ (since, the lower Δ_i the more easily one can mistake the i-th arm for the optimal arm).

3.2.2 Lower bounds

This section presents the main results on lower bound for the distribution-free and the distribution-dependent cases.

Theorem 1 (Auer et al. (1995))**.** *Let* sup *be the supremum over all stochastic bandit bounded in* $[0,1]$ *and* inf *the infimum over all forecasters, the following holds true:*

$$\text{inf}\,\text{sup}\ \mathbb{E}[R_t] \geqslant \frac{1}{20}\sqrt{tK} \tag{3.6}$$

Definition 3.2.5 (Kullback and Leibler (1951))**.** *Let $\mathscr{P}([0,1])$ be the set of probability distribution over* $[0,1]$*. The Kullback-Leibler divergence between two distributions P*

and Q in $\mathscr{P}([0,1])$ is defined as:

$$KL(P,Q) = \begin{cases} \int_{[0,1]} \frac{dP}{dQ} \log \frac{dP}{dQ} dQ & \text{if } P \ll Q \\ +\infty & \text{otherwise} \end{cases} \tag{3.7}$$

A first version of the following theorem has been proposed by (Lai and Robbins, 1985) and then extended by (Burnetas and Katehakis, 1996).

Theorem 2 (Burnetas and Katehakis (1996))**.** *Let $\mathscr{P} \subset \mathscr{M}([0,1)$ and a forecaster consistent with $\mathscr{P}$, i.e. for any stochastic bandit, for any suboptimal arm i and any $\beta > 0$, $\mathbb{E}[N_{i,t}] = o(T^{\beta})$. For any stochastic bandit with distribution in $\mathscr{P}$, the following holds true:*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R_t]}{\log t} \geqslant \sum_{i:\Delta_i>0} \frac{\Delta_i}{\mathscr{K}_{\inf}(\nu_i, \mu^\star)} \tag{3.8}$$

with $\mathscr{K}_{\inf}(\nu_i, \mu^\star) = \inf\{KL(\nu_a, \nu) : \nu \in \mathscr{P} \text{ and } \mathbb{E}[\nu] > \mu^\star\}$

3.3 The contextual bandit case

3.3.1 Position of the problem

The contextual bandit differs from the canonical MAB setting as it is assumed that at each time step a *context* is revealed to the agent. The agent strategy thus becomes to associate with each context an arm, in order to minimize the so-called contextual regret (below). The context, or side information, most naturally arises in some standard applications of MABs. For instance in the ads placement problem, the customer comes with some characteristic features (e.g., gender, age, location). One could see the contextual bandit as closer to the reinforcement learning setting, than the canonical MAB, by considering that the *context* corresponds to some extent to the state of the agent. However the notion of dynamics in the state space, in relation with the action selection of the agent, is not present in the contextual bandit setting; still, the contextual bandit is a setting where the exploration / exploitation dilemma of the agent must take into account the current context.
In practice, contextual bandits are most often handled as a classification problem, associating with each context the arm yielding best reward for this context. Based on the information gathered during an exploration phase, one constructs a set of triplets (x_t, I_t, Y_t) which is used to learn how reward Y_t depends on the context x_t for the I_t-th arm; this prediction is thereafter used to seek the arm yielding the maximum reward for a given context.
With same MAB notations as in Section 3.1, let us introduce the notations related to Contextual Bandits. The contextual bandit adds side information in the form of a

context vector from a given set $\mathcal{X} \subset \mathbb{R}^d$. At each time step $t \in \mathbb{N}$, a context $X_t \in \mathcal{X}$ is drawn and revealed (e.g., in the case of a recommender system, X_t may be a vector summarizing the known information about the incoming user). Based on the information provided by X_t, the learner chooses an arm I_t and observes an instantaneous reward Y_t (e.g., in the recommender system case, the arm set represents the different available contents, and Y_t indicates whether the user clicks on the displayed contents).

3.3.2 Linear contextual bandit

In the linear contextual bandit case, one assumes that the i-th arm is associated with a (hidden) weight vector θ_i, such that the instantaneous reward associated to the i-th arm is the scalar product of the context and the weight vector, augmented with some centered scalar noise η_t:

$$Y_t = \langle X_t, \theta_{I_t} \rangle + \eta_t \tag{3.9}$$

with $\{\theta_1 \dots \theta_K\}$ a set of unknown vector parameters in $\Theta \subset \mathbb{R}^d$.
The following additional assumptions are made:

- Firstly, it is assumed that contexts X_t are independently drawn by Nature.
- Secondly, it is assumed that the parameters θ_i associated with each arm (to be estimated) are independent from each other. This scheme is called **disjoint linear model** (see (Li et al., 2010)); it differs from the `OFUL` (Abbasi-Yadkori et al., 2011) and `Thompson sampling` (Agrawal and Goyal, 2012a) setting, where a common parameter is shared among the arms.

An algorithm based on sub-sampling for Linear Contextual bandit in the disjoint linear model will be proposed in Chapter 7.

Chapter 4

Multi-Armed Bandit Algorithms

Contents

4.1 Introduction

The many algorithms designed in the past decade to tackle the multi-armed bandit problem, specifically addressing the stochastic and the contextual MAB settings, can be divided into four categories:

1. ***Greedy algorithms*** are mostly based on the exploitation of the best arm evaluated so far, with little to no exploration.

2. ***Optimistic algorithms***: these algorithms rely on the *optimism in face of uncertainty* principle. They use the reward samples to maintain a confidence region for every arm reward expectation, where the mean lies with a high probability. The optimistic strategy consists of selecting the arm with highest possible mean reward based on the current knowledge.

3. ***Bayesian algorithms***: These algorithms, mainly variants of `Thompson sampling`, maintain an estimated payoff distribution for each arm. Given a prior distribution on each arm, and the reward samples gathered every time the arm is selected, the posterior arm distribution is computed using Bayesian inference and the arm with maximal estimated expectation is selected.

4. ***Sub-Sampling based algorithm***: Recently introduced in (Baransi et al., 2014), this algorithm selects the arm with best average reward, where the average estimate is based on an equal quantity of samples for each arm.

4.2 Greedy algorithms

This section is devoted to the presentation of some greedy MAB algorithms. After introducing a pure greedy method, some methods achieving an exploitation / exploration trade-off are presented and discussed. As will be seen in Chapters 5 and 6, the simplicity of the greedy algorithms does not prevent them from being competitive in some settings.

4.2.1 Pure Greedy

The simplest strategy for the MAB problem consists of trying once each arm and greedily selecting the arm with best average reward ever after (Algorithm 1).
The greedy, exploitation-only, approach is likely to miss the best arm. Consider for instance the 2-arm problem with Bernoulli distributions $\nu_1 = \mathscr{B}(0.2)$ and $\nu_2 = \mathscr{B}(0.8)$. It is clear than $i^\star = 2$. However, if unfortunately the first trials yield $Y_1 = 1$ and $Y_2 = 0$, the pure greedy algorithm is bound to select the 1st arm ever after, yielding a linear

Algorithm 1 Pure Greedy for K arms

Require: Time horizon T

1: **for** $t = 1 \dots K$ **do**
2: Select $I_t = t$ and receive reward Y_t
3: $\widehat{\mu_t} = Y_t$
4: **end for**
5: **for** $t = K+1 \dots T$ **do**
6: Select

$$I_t \in \underset{i \in \{1,\dots,K\}}{\operatorname{argmax}} \{\widehat{\mu_i}\} \tag{4.1}$$

7: **end for**

(and maximal) pseudo-regret: $\overline{R_t} = (T-1)\Delta$. This simple example shows the failure of the pure greedy approach, and the fact that exploration is definitely required to achieve logarithmic (pseudo-)regret rates guarantees.

4.2.2 ε- greedy algorithm

A simple way to achieve exploration is by mixing the above pure greedy strategy with a uniformly random exploration strategy, where the exploration strategy is launched with probability $\varepsilon \in (0,1)$. This mixed strategy is referred to as ε-greedy algorithm (Sutton and Barto, 1998; Watkins, 1989) (Algorithm 2).

Algorithm 2 ε-greedy for K arms

Require: Time horizon T, Parameter $\varepsilon \in (0,1)$

1: **for** $t = 1 \dots K$ **do**
2: Select $I_t = t$ and receive reward Y_t
3: $\widehat{\mu_t} = Y_t$
4: **end for**
5: **for** $t = K+1 \dots T$ **do**
6: Select

$$I_t = \begin{cases} \underset{i \in \{1 \dots K\}}{\operatorname{argmax}} \{\widehat{\mu_i}\} & \text{with probability } 1-\varepsilon. \\ \mathbb{U}(\{1,\dots,K\}) & \text{with probability } \varepsilon. \end{cases} \tag{4.2}$$

7: **end for**

While the exploration makes it more unlikely to miss the best arm, the ε-greedy algorithm clearly yields a linear regret with constant ε.

4.2.3 ε_t-greedy

As shown by Auer et al. (2002), the ε-greedy approach can be improved by allowing the ε parameter to decrease with t, defining the ε_t-greedy algorithm (Algorithm 3), and yielding a sublinear regret rate.

Algorithm 3 ε_t-greedy for K arms

Require: Time horizon T, Non-increasing parameter $\varepsilon_t \in (0,1)$

1: **for** $t = 1 \dots K$ **do**
2: Select $I_t = t$ and receive reward Y_t
3: $\widehat{\mu_t} = Y_t$
4: **end for**
5: **for** $t = K+1 \dots T$ **do**
6: Select

$$I_t = \begin{cases} \underset{i\in\{1\dots K\}}{\operatorname{argmax}}\{\widehat{\mu_i}\} & \text{with probability } 1-\varepsilon_t. \\ \cup(\{1,\dots,K\}) & \text{with probability } \varepsilon_t. \end{cases} \tag{4.3}$$

7: **end for**

Proposition 4.2.1 (Cesa-Bianchi and Fischer (1998); Auer et al. (2002))**.** *Let $c > 0$ and $0 < d < 1$ be two parameters.*
By setting $\varepsilon_t = \min\left\{1, \frac{cK}{d^2 t}\right\}$, the pseudo-regret of ε_t-greedy is upper-bounded, for any $t > \frac{cK}{d}$, as follows:

$$\mathbb{E}[\overline{R_t}] \leqslant \sum_{i:\Delta_i>0} \Delta_i \left\{ \frac{c}{d^2 t} + 2\left(\frac{c}{d^2}\log\left(\frac{(t-1)\,d^2 e^{1/2}}{cK}\right)\right)\left(\frac{cK}{(t-1)\,d^2 e^{1/2}}\right)^{\frac{c}{5d^2}} + \frac{4e}{d}\left(\frac{cK}{(t-1)\,d^2 e^{1/2}}\right)^{\frac{c}{2}} \right\} \tag{4.4}$$

Remark 4.2.1. *The empirical study proposed by Vermorel and Mohri (2005) did not show any practical gain of the ε_t-greedy over the ε-greedy algorithm, despite the better guarantees on the regret of the former algorithm.*

4.2.4 ε-first strategy

The ε-first strategy is another simple greedy algorithm (Algorithm 4).
This algorithm is a basis for the study of the Multi-Armed Bandit in the probably approximately correct (PAC) framework (Even-dar et al., 2002; Mannor and Tsitsiklis, 2004).
Let an ε-optimal arm be defined as an arm with gap less than ε (a is ε-optimal iff $\mu^\star \geqslant \mu_a - \varepsilon$). Then, with probability $\delta \in (0,1)$, it requires $O(\frac{K}{\varepsilon^2}\log\frac{1}{\delta})$ to discover an ε-optimal arm.

Algorithm 4 ε-first strategy for K arms

Require: Time horizon T, Parameter ε

1: **for** $t = 1 \dots \lceil \varepsilon T \rceil$ **do**

2: Select (exploration phase)

$$I_t = \mathbb{U}([1, \dots, K]) \tag{4.5}$$

3: **end for**

4: **for** $t = \lceil \varepsilon T \rceil + 1 \dots T$ **do**

5: Select (exploitation phase)

$$I_t \in \underset{i \in \{1,\dots,K\}}{\operatorname{argmax}} \{\widehat{\mu_i}\} \tag{4.6}$$

6: **end for**

Note that the idea of decoupling the bandit algorithm into two successive exploration and exploitation phases will be used in the ExpExp algorithm (section 5.2.1).

4.3 Optimistic algorithms

This section is devoted to the presentation of optimistic algorithms, which proceed by maintaining a confidence region for each arm expected payoff. Key ingredients in the construction of these algorithms are the *concentration inequalities*, upper bounding the error of estimation, that is the difference between the estimator and the quantity of interest (here the reward expectation).
Let us first recall the Hoeffding inequality (Hoeffding, 1963), which plays a key role in the optimistic algorithm design.

Theorem 3 (Hoeffding (1963))**.** *Let ν be a probability distribution with $\mathbb{E}[\nu] = \mu$ and $x_1, \dots, x_n$ an i.i.d. sample of ν.*
Let us further suppose that distribution ν is bounded, i.e. there exists $(a_i, b_i)_{i=1}^n$ such that $\mathbb{P}(a_i \leqslant x_i \leqslant b_i) = 1$.
Then, by denoting $\widehat{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$ the empirical mean and for $\varepsilon > 0$, it comes:

$$\mathbb{P}\left(\widehat{x}_n - \mu \geqslant \varepsilon\right) \leqslant \exp\left(-\frac{2\varepsilon^2 n^2}{\sum_{i=1}^n (b_i - a_i)}\right) \tag{4.7}$$

$$\mathbb{P}\left(\widehat{x}_n - \mu \leqslant -\varepsilon\right) \leqslant \exp\left(-\frac{2\varepsilon^2 n^2}{\sum_{i=1}^n (b_i - a_i)}\right) \tag{4.8}$$

And, by combining Equation 4.7 and 4.8, it comes:

$$\mathbb{P}\left(|\widehat{x}_n - \mu| \geqslant \varepsilon\right) \leqslant 2\exp\left(-\frac{2\varepsilon^2 n^2}{\sum_{i=1}^n (b_i - a_i)}\right) \tag{4.9}$$

4.3.1 Upper Confidence Bound and variants

Upper Confidence Bound

Introduced by Auer et al. (2002), the Upper Confidence Bound algorithm essentially relies on the Hoeffding bound in Theorem 3 to derive an upper bound on the regret (Algorithm 5). It derives for each arm a confidence interval on the associated expected reward, and selects the arm with the highest confidence bound [1].

Algorithm 5 UCB for K arms (Auer et al., 2002)

Require: Time horizon T, parameter $C > 1$

1: **for** $t = 1 \dots T$ **do**

2: Select

$$I_t \in \underset{i \in \{1,\dots,K\}}{\operatorname{argmax}} \left\{ \widehat{\mu_i} + \sqrt{\frac{C \log t}{N_{i,t}}} \right\} \tag{4.10}$$

3: **end for**

Theorem 4 (Auer et al. (2002))**.** *For $C = 2$, the pseudo-regret of UCB is upper-bounded by:*

$$\mathbb{E}\left[\overline{R_t}\right] \leqslant \sum_{i:\mu_i<\mu^\star} \left\{ \frac{8 \log t}{\Delta_i} + 1 + \frac{\pi^2}{3} \right\} \tag{4.11}$$

Sketch of proof. Let us present the main ideas from the proof, as they will serve in Chapter 6.

As the equality $\mathbb{E}\left[\overline{\mathscr{R}_t}\right] = \sum_{i=1}^{K} \Delta_i \mathbb{E}\left[N_{i,t}\right]$ suggests, upper-bounding the expected number $\mathbb{E}\left[N_{i,t}\right]$ of sub-optimal arm pulls permits a direct pseudo-regret bounding.

Let us denote $c_{t,s} \overset{\text{def}}{=} \sqrt{\frac{2 \log t}{s}}$ the exploratory term. Then, when a sub-optimal arm i is selected at time t, one of the following conditions is verified:

$$\widehat{X_{i^\star,N_{i^\star,t}}} \leqslant \mu^\star - c_{t,N_{i^\star,t}} \tag{4.12}$$

$$\widehat{X_{i,N_{i,t}}} \geqslant \mu_i + c_{t,N_{i,t}} \tag{4.13}$$

$$\mu^\star < \mu_i + 2c_{t,N_{i,t}} \tag{4.14}$$

Informally, the selection of an under-optimal arm is explained as: i) the best arm payoff is largely understimated (Equation 4.12); or ii) the suboptimal arm i is largely overestimated (Equation 4.13); or iii) the gap Δ_i is too small for the i-th arm to be distinguishable from the best arm with only $N_{i,t}$ samples (Equation 4.14)

[1] In the original definition (Auer et al., 2002), the constant $C = 2$.

To conclude the proof, it suffices to assume that $N_{i,t}$ is large enough so that Equation 4.14 does not hold. In such condition, Equation 4.12 or 4.13 necessarily hold, and the probability of the associated events are precisely controlled by the Hoeffding bound. □

Variance estimates for Multi-Armed Bandit

Introduced by (Auer et al., 2002), the idea of exploiting the empirical variance of the arm has been firstly used in the `UCB-Tuned` algorithm. Informally, everything else being equal, an arm with large variance should be explored more often than an arm with small variance. In the general case, `UCB-Tuned` has been shown to empirically outperform `UCB`, encouraging the formal study of the variance estimation in `UCB`-like bandits in (Audibert et al., 2009). The resulting algorithm, termed Upper confidence bound with Variance (`UCB-V`), is described in Algorithm 6 with $V_i = \frac{1}{N_{i,t}} \sum_{s=1}^{t} \left(X_{i,s} - \widehat{\mu_i}\right)^2$ the empirical variance of arm i.

Algorithm 6 `UCB-V` for K arms (Audibert et al., 2009)

Require: Time horizon T, exploration parameter $C > 0$

1: **for** $t = 1 \dots K$ **do**
2: Select $I_t = t$
3: **end for**
4: **for** $t = K+1 \dots T$ **do**
5: Select (arbitrary tie break)

$$I_t \in \underset{i \in \{1,\dots,K\}}{\operatorname{argmax}} \left\{ \widehat{\mu_i} + \sqrt{\frac{2CV_i \log(t)}{N_{i,t}}} + 3C\frac{\log(t)}{N_{i,t}} \right\} \tag{4.15}$$

6: **end for**

Theorem 5 ((Audibert et al., 2009, 2010))**.** *For $\alpha > 1$ the expected regret of `UCB-V` is upper bounded as follows:*

$$\mathbb{E}[R_t] \leqslant 8\alpha \sum_{i:\Delta_i > 0} \left(\frac{\sigma_i^2}{\Delta_i} + 2 \right) \log(t) + \Delta_i \left(2 + \frac{12}{\log(\alpha+1)} \left(\frac{\alpha+1}{\alpha-1} \right)^2 \right) \tag{4.16}$$

Remark 4.3.1. *The bound shows that the regret might be lower than the one of `UCB`, especially in the case where the suboptimal arm has a small variance σ^2.*

Optimistic algorithm in the distribution-free case

As stated in the previous chapter, two types of bound are considered for the regret: distribution-free or distribution-dependent. Most algorithms are concerned by the

distribution-dependent setting, with regret bound depending of distribution parameters such as Δ_i.
In the distribution-free setting, Audibert and Bubeck (2010) derive an algorithm termed (*Minimax Optimal Strategy in the Stochastic case* MOSS) matching the lower-bound from (Auer et al., 1995). The algorithm involves a modified penalization associated to each arm (Algorithm 7).

Algorithm 7 MOSS for K arms (Audibert and Bubeck, 2010)

Require: Time horizon T

1: **for** $t = 1 \dots K$ **do**
2: Select $I_t = t$
3: **end for**
4: **for** $t = K+1 \dots T$ **do**
5: Select

$$I_t \in \underset{i \in \{1,\dots,K\}}{\operatorname{argmax}} \left\{ \widehat{\mu_i} + \sqrt{\frac{\max\left(\log\left(\frac{n}{KN_{i,t}}\right), 0\right)}{N_{i,t}}} \right\} \tag{4.17}$$

6: **end for**

Theorem 6 (Audibert and Bubeck (2010))**.** *MOSS satisfies:*

$$\sup \mathbb{E}[R_t] \leqslant 49\sqrt{tK} \tag{4.18}$$

with the supremum taken over all K arm distributions in $[0,1]$.
The following problem-dependent bound also holds:

$$\mathbb{E}[R_t] \leqslant 23K \sum_{i:\Delta_i>0} \frac{\max\left(\log\left(\frac{n\Delta_i^2}{K}\right), 1\right)}{\Delta_i} \tag{4.19}$$

4.3.2 Kullback-Leibler based algorithms

Inspired by the seminal papers of (Lai and Robbins, 1985; Burnetas and Katehakis, 1996), new algorithms have been proposed based on the estimation of the whole reward distribution, as opposed to estimating only its first (UCB) or first and second (UCB-V) moments.
As stated by Maillard et al. (2011), the goal is to reach optimality guarantees in the sense of (Burnetas and Katehakis, 1996), i.e., deriving algorithms such that:

$$\mathbb{E}\left[N_{i,t}\right] \leqslant \left(\frac{1}{\mathcal{K}_{inf}(\nu_i, \mu^\star)} + o(1) \right) \log T \tag{4.20}$$

While UCB achieves the logarithmic regret rate, its constant depends on the quantity Δ_i^2; the above constant, in $1/\mathcal{K}_{inf}$ thus is improved as $\mathcal{K}_{inf}$ might be greater than Δ_i^2 due to Pinkser's inequality (Maillard et al., 2011).

Deterministic Minimum Empirical Divergence

The *Minimum Empirical Divergence* (MED) and *Deterministic Minimum Empirical Divergence* (DMED) algorithms are proposed by (Honda and Takemura, 2010, 2011). Given a distribution ν and $\mu \in \mathbb{R}$ its expectation, let D_{min} be defined as:

$$D_{min}(\nu, \mu) \stackrel{\text{def}}{=} \inf\left\{KL(\nu, \nu') : \nu' \in \mathscr{A}, E[\nu'] \geqslant \mu\right\} \tag{4.21}$$

with $\mathscr{A}$ the set of distribution ν with $Supp(\nu) \subset [0,1]$.
Let $\widehat{\nu_i, t} = \frac{1}{N_{i,t}} \sum_{s=1}^{N_{i,t}} \delta_{X_{i,s}}$ denote the empirical distribution of ν_i at time t, and $\widehat{\mu^\star(t)} = \max_i \left\{\widehat{\mu_i(t)}\right\}$. The authors are interested in the set of quasi-optimal arms, i.e. such that:

$$N_{i,t} D_{min}(\widehat{\nu_{i,t}}, \widehat{\mu^\star(t)})) \leqslant \log n - \log N_{i,t} \tag{4.22}$$

The algorithm divides the time horizon in a sequence of loops. In each loop, arms satisfying Equation 4.22 are played. This leads to an asymptotically optimal algorithm, satisfying:

$$\limsup_{t\to\infty} \frac{\mathbb{E}[N_{i,t}]}{\log t} \leqslant \frac{1}{\mathcal{K}_{inf}(\nu_i, \mu^\star)}$$

The algorithm requires to compute the quantity D_{min} (Equation 4.21) for each arm. This computation can be done explicitly via a dual formulation. The interested reader is referred to (Honda and Takemura, 2010) for further details.

Kullback-Leibler upper confidence bound

Kullback-Leibler upper confidence bound (KL-UCB) refers to a family of index-based optimistic algorithms relying on the Kullback-Leibler divergence and inspired by Lai and Robbins (1985). The idea and analysis simultaneously appeared in (Garivier, 2011) and (Maillard et al., 2011) and are presented in an extended and unified way in (Cappé et al., 2013). As for DMED, the aim is to provide optimal solutions, matching the lower regret bound from (Burnetas and Katehakis, 1996). However, and in opposition to DMED, these algorithms benefit from a finite-time analysis showing their optimality even in the non-asymptotic case.
Following Cappé et al. (2013), two classes of algorithms can be considered:

1. kl-UCB in the case of one-parameter exponential reward distributions.

2. `Empirical KL-UCB` in the case of bounded distributions with finite support.

`kl-UCB` and `Empirical KL-UCB` are respectively described in Algorithms 8 and 9. By considering I the open interval of all possible values of μ, `kl-UCB` selects the arm maximizing the upper-bound U_i defined by Equation 4.23. In the case of `Empirical KL-UCB`, the expectation is maximized on the space $\mathfrak{M}_1\left(\text{Supp}\left(\widehat{\nu_i(t)}\right)\cup\{1\}\right)$ defined as the set of distributions with support $\text{Supp}\left(\widehat{\nu_i(t)}\right)\cup\{1\}$ (Equation 4.24).
As shown by Cappé et al. (2013), both algorithms can be written in a generic form. The Kullback-Leibler divergence of Eq. 4.24 is simplified in the exponential case, as these distributions are characterized by their expected value and can be explicitly given by a closed-form function d, depending on the considered family of reward distributions. For instance, for Bernoulli arms, $I=(01)$ and $d_{Ber}(\mu,\mu')=\mu\log\frac{\mu}{\mu'}+\left(1-\mu\right)\log\frac{1-\mu}{1-\mu'}$.
For the considered cases, a finite-time analysis of $\mathbb{E}\left[N_{i,t}\right]$ shows the optimality of the approaches. In the more general bounded cases, `Empirical KL-UCB` is experimentally validated on synthetic problems but deriving theoretical guarantees remains an open problem.

Algorithm 8 `kl-UCB` for K arms (Cappé et al., 2013)

Require: Time horizon T, non-decreasing function $f:\mathbb{N}\to\mathbb{R}$

for $t=1\dots K$ **do**
  Select $I_t=t$
end for
for $t=K+1\dots T$ **do**
  Select

$$I_t\in\underset{i\in\{1,\dots,K\}}{\text{argmax}}\{U_i(t)\}$$

  with

$$U_i(t)\overset{\text{def}}{=}\sup\left\{\mu\in\overline{I}:d(\widehat{\mu}_i(t),\mu)\leqslant\frac{f(t)}{N_{i,t}}\right\}\tag{4.23}$$

end for

4.4 Bayesian algorithms

Bayesian algorithms include the `Thompson sampling` algorithms, introduced by Thompson (1933, 1935) simultaneously to the Multi-Armed Bandit problem. They have since been extensively studied (see (Chapelle and Li, 2011; Agrawal and Goyal, 2012b; Kaufmann et al., 2012)).

4.4.1 Algorithm description

The algorithms are described in (Agrawal and Goyal, 2012b) (Algorithms 10 and 11).

Algorithm 9 `Empirical KL-UCB` for K arms (Cappé et al., 2013)

Require: Time horizon T, non-decreasing function $f : \mathbb{N} \to \mathbb{R}$

1: **for** $t = 1 \dots K$ **do**
2: Select $I_t = t$
3: **end for**
4: **for** $t = K+1 \dots T$ **do**
5: Select

$$I_t \in \underset{i \in \{1,\dots,K\}}{\operatorname{argmax}} \{U_i(t)\}$$

with

$$U_i(t) \overset{\text{def}}{=} \sup \left\{ \mathbb{E}(\nu) : \nu \in \mathfrak{M}_1 \left(\operatorname{Supp}\left(\widehat{\nu_i(t)}\right) \cup \{1\} \right) \text{ and } KL(\widehat{\nu_i(t)}, \nu) \leqslant \frac{f(t)}{N_{i,t}} \right\} \quad (4.24)$$

6: **end for**

Algorithm 10 `Thompson sampling` for K Bernoulli arms (Agrawal and Goyal, 2012b)

Require: Time horizon T

1: **for** $i = 1 \dots K$ **do**
2: $S_i(1) = F_i(1) = 0$
3: **end for**
4: **for** $t = 1 \dots T$ **do**
5: For each arm i, sample $\theta_i(t) \sim \beta(S_i(t)+1, F_i(t)+1)$
6: Play

$$I_t \in \underset{i \in \{1 \dots K\}}{\operatorname{argmax}} \{\theta_i(t)\} \quad (4.25)$$

7: Observe Y_t, if $Y_t = 1$, $S_{I_t} = S_{I_t} + 1$, otherwise $F_{I_t} = F_{I_t} + 1$
8: **end for**

Algorithm 11 `Thompson sampling` for K arms (Agrawal and Goyal, 2012b)

Require: Time horizon T

1: **for** $i = 1 \dots K$ **do**
2: $S_i(1) = F_i(1) = 0$
3: **end for**
4: **for** $t = 1 \dots T$ **do**
5: For each arm i, sample $\theta_i(t) \sim \beta(S_i(t)+1, F_i(t)+1)$
6: Play

$$I_t \in \underset{i \in \{1 \dots K\}}{\operatorname{argmax}} \{\theta_i(t)\} \quad (4.26)$$

7: Observe Y_t and sample $X \sim Ber(\widehat{Y_{I_t}})$; if $X = 1$, $S_{I_t} = S_{I_t} + 1$, otherwise $F_{I_t} = F_{I_t} + 1$
8: **end for**

`Thompson sampling` algorithms rely on a randomized strategy sampling. At each time step t, after having observed $S_i(t)$ successes and $F_i(t)$ failures, the distribution attached to the i-th arm is estimated by a posterior Beta distribution (Line 5 in Algorithm 10 and 11).
The selected arm is (in expectation) the one with maximal expected reward respectively to these distribution estimates (Line 6 in Algorithm 10 and Algorithm 11). The posterior distributions are then updated, depending on the algorithm.
Algorithm 11 extends Algorithm 10 to the general case of any distribution, by sampling an external binary random variable X in such a way that its expectation is the same as the empirical average of the selected distribution; it thereafter updates S_i and F_i as done in Algorithm 10 for Bernoulli variables.

4.4.2 Discussion

The use of Bernoulli prior distribution is convenient as it is a conjugate distribution of the Beta distribution. This allows easy computations of the posteriors distribution after the observation of a Bernoulli realization. Also, $\beta(1,1)$ is the uniform distribution; initializing the estimated distributions to a uniform distribution seems reasonable.
Note that `Thompson sampling` defines a family of algorithms as the choice of the priors is left to the user/designer (it is not necessarily uniform).

4.4.3 Historical study of `Thompson sampling`

`Thompson sampling` is the first algorithm proposed for the multi-armed bandit problem, as Thompson (1933) defined both the problem and the algorithm. However, the interest around `Thompson sampling` has been revived only recently. First, Chapelle and Li (2011) provided an extended empirical evaluation of the algorithm, emphazing its advantages. This paper concludes with the need of a theoretical study of the algorithm to improve its popularity and usability.
An important first theoretical result is due to Agrawal and Goyal (2012b), who established a logarithmic expected regret for the stochastic multi-armed bandit. More recently, (Kaufmann et al., 2012) improved this result and showed than an optimal finite time analysis can be achieved, matching the Lai and Robbins lower bound:

Theorem 7 (Kaufmann et al. (2012))**.** *For every $\varepsilon > 0, \exists C = C(\varepsilon, \mu_1, \ldots, \mu_K)$, such that the pseudo-regret of Thompson Sampling is upper-bounded as:*

$$\mathbb{E}[\overline{R_t}] \leqslant (1+\varepsilon) \sum_{i \in \{1,\ldots,K\}: \mu_i < \mu^\star} \frac{\Delta_i \left(\log(i) + \log\left(\log(t)\right)\right)}{K\left(\mu_i, \mu\star\right)} + C \tag{4.27}$$

with $K(p,q) \overset{\text{def}}{=} p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ *the Kullback-Leibler Divergence between* $\mathscr{B}(p)$ *and* $\mathscr{B}(q)$.

4.5 The subsampling strategy

4.5.1 Introduction

Recently, (Baransi et al., 2014) introduced an algorithm with a different approach, based on a sub-sampling technique. Intuitively, comparisons between two arms should be based on an equal quantity of information. To do so, the algorithm sub-samples from the most played arm a number of samples equal to the number of samples of the least-sampled arm (without replacement). The arm with maximal empirical mean (taken over the sub-sample) is selected.

4.5.2 Definition

Two arm case

The standard notations are as follows:

- $Wr(n,m)$ is the subsampling distribution without replacement, with the convention $Wr(n,m) = \{1,\dots,n\}$ if $m \geqslant n$.
- For $X_{1:n} \overset{\text{def}}{=} (X_1,\dots,X_n)$ an i.i.d sample of a random distribution, and $I = \{i_1,\dots,i_m\} \subset \{1,\dots,n\}$, $X_{1:n}(I) = \{X_{i_1},\dots,X_{i_m}\}$.
- For a and b two arms, $X^a_{1:N_a(t)}$ and $X^b_{1:N_b(t)}$ respectively denote the sample for arm a and b up to time t.

With these notations, the *Best Empirical Sampled Average* (`BESA`) algorithm is described in the two-arm case as follows (Alg. 12).

Algorithm 12 `BESA` (a,b) for two arms

Require: Current time t, Two arms a and b, Time horizon T

1: Sample $I^a_t \sim Wr(N_a(t), N_b(t))$ and $I^a_t \sim Wr(N_b(t), N_a(t))$
2: Compute $\widehat{\mu_{t,a}} = \widehat{\mu}\left(X_{1:N_t(a)^a}\left(I^a_t\right)\right)$ and $\widehat{\mu_{t,b}} = \widehat{\mu}\left(X_{1:N_t(b)^b}\left(I^b_t\right)\right)$
3: Select (break ties by selecting the less played arm)

$$I_t = \underset{i\in\{a,b\}}{\operatorname{argmax}}\left\{\widehat{\mu}_{t,i}\right\} \tag{4.28}$$

K-arm case

The extension to any finite number K of arms is achieved using a tournament between arms (Alg. 13). To avoid a bias due to a fixed comparison order among the arms, the arm set $\mathscr{A}$ is shuffled in each time step.

Algorithm 13 BESA ($\mathscr{A}$)

Require: Current time t, Arm set $\mathscr{A}$ of size K

if $\mathscr{A} = \{a\}$ **then**
  $I_t = a$
else
  $I_t = \texttt{BESA}\left(\texttt{BESA}\left(\left\{1,\dots,\lceil\frac{K}{2}\rceil\right\}\right),\texttt{BESA}\left(\left\{\lfloor\frac{K}{2}\rfloor,\dots,K\right\}\right)\right)$
end if

4.5.3 Regret bound

Definition 4.5.1. *For integers M, n and for $\lambda \in [0,1]$, the **balance function** of the distributions $(\nu_a, \nu_\star)$ is defined as:*

$$\alpha_\lambda(M,n) = \mathbb{E}_{Z\sim\nu_\star,n}\left[\left(1 - F_{\nu_{a,n}}(Z) + \lambda\nu_{a,n}(Z)\right)^M\right]$$

with $\nu_{a,n}$ is the distribution of $\sum_{i=1}^{n} X_i^a$ with $X_i^a \sim \nu_a$ and F_ν the cumulative distribution function of ν (i.e. $F_\nu(x) = \mathbb{P}(X \leqslant x)$ with $X \sim \nu$).

Theorem 8 (Baransi et al. (2014))**.** *Let $\mathscr{A} = \{\star, a\}$ be a two arm set with bounded rewards in $[0,1]$ and $\Delta = \mu^\star - \mu_a$ be the mean gap. If there exist two constants $\alpha \in (0,1)$ and $c > 0$, such that $\alpha_{1/2}(M,1) < c\alpha^M$, then the regret of BESA at time t is upper-bounded as:*

$$\mathbb{E}[R_t] \leqslant \frac{11\log t}{\Delta} + C_{\nu_a,\nu_\star} + O(1) \tag{4.29}$$

Remark 4.5.1. *This bound establishes the relevance of the sub-sampling idea for the multi-armed bandit setting, although one might have thought intuitively that sub-sampling entails a loss of information.*

Remark 4.5.2. *Together with this theoretical result, empirical results confirm the BESA efficiency, besides other desirable properties such as implementation simplicity and flexibility. The latter property is particularly appreciated as one can apply the same algorithm for different distribution families while preserving the strong results, as opposed to the fact that KL-UCB or Thompson sampling for instance require one to know the family of reward distributions.*

4.6 Contextual MAB Algorithms

This section describes some contextual MAB algorithms.

4.6.1 OFUL

The *Optimism in the Face of Uncertainty Linear* (`OFUL`) bandit algorithm is due to Abbasi-Yadkori et al. (2011). `OFUL` does not consider any contextual information X_t, nor does it looks for an unknown parameter per arm. Instead, it assumes that a (possibly continuous) arm set $\mathscr{A}_t$ is provided at each time step and supposed perfectly known.
The goal is to learn a shared unknown parameter $\beta^\star$ such that

$$Y_t = \langle \beta^\star, \theta_{I_t} \rangle + \eta_t$$

with η_t a sub-Gaussian additive noise (Equation (7.2)).
Its pseudo-code for the K-armed bandit is presented in Algorithm 14. For clarity, the original notations have been changed to correspond to the one provided in Chapter 3.

Algorithm 14 `OFUL` for K arms (Abbasi-Yadkori et al., 2011)

Require: Current time t, current confidence ellipsoid C_{t-1}, parameters R_{OFUL}, S_{OFUL}, λ, and confidence δ.

1: Compute

$$(I_t, \widetilde{\beta_t}) = \underset{(i,\beta)\in\{1,\dots,K\}\times C_{t-1}}{\operatorname{argmax}} \langle \theta_i, \beta \rangle \tag{4.30}$$

2: Play arm I_t and observe reward Y_t
3: Update the confidence ellipsoid C_t thanks to Equation (4.31)

The confidence ellipsoid C_t is maintained, such that $\beta^\star$ provably belongs to C_t with probability at least $1-\delta$:

$$C_t \overset{\text{def}}{=} \left\{ \beta \in \mathbb{R}^d : \|\beta - \widehat{\beta_t}\|_{\overline{V_t}} \leqslant R_{OFUL} \sqrt{2\log\left(\frac{\det \overline{V_t}^{1/2} \det \lambda I_d{}^{-1/2}}{\delta} \right)} + \lambda^{1/2} S_{OFUL} \right\} \tag{4.31}$$

with:

- $\delta \in (0,1)$ a user-defined confidence level.
- $\overline{V_t} \overset{\text{def}}{=} \lambda I_d + \sum_{s=1}^{t} \theta_{I_s} \theta_{I_s}^T$.

- $\widehat{\beta_t} \overset{\text{def}}{=} \left(B_{[t]}^T B_{[t]} + \lambda I_d\right)^{-1} B_{[t]}^T Y_{[t]}$ with $\lambda > 0$ the least-squares estimate of $\beta^\star$ with $B_{[t]} \overset{\text{def}}{=} (\theta_{I_1}^T, \dots, \theta_{I_t}^T)^T \in \mathbb{R}^{td}$ and $Y_{[t]} \overset{\text{def}}{=} (Y_1, \dots, Y_t)^T$,.
- R_{OFUL} a positive parameter such that η_t is R_{OFUL} sub-Gaussian.
- S_{OFUL} a positive parameter such that $\|\beta^\star\|_2 \leqslant S_{OFUL}$.

In the general case (continuous bandit set), solving the optimization problem defined in Equation (4.30) requires the use of a iterative procedure (for instance a Newton method) at each time step, implying significant computational overhead in high dimension.
However, in the case of a finite arm set, (Rusmevichientong and Tsitsiklis, 2010, Equations 6) provides a closed-form solution to this optimization problem, yielding faster computations.
Here, for $i \in \{\{1, \dots, K\}$:

$$\max_{\beta \in C_t} \langle \theta_i, \beta \rangle = \left(R_{OFUL} \sqrt{2 \log \left(\frac{\det \overline{V_t}^{1/2} \det \lambda I_d^{-1/2}}{\delta} \right)} + \lambda^{1/2} S_{OFUL} \right) \|\theta_i\|_{\overline{V_t}^{-1}} + \widehat{\beta_t}^T \theta_i ,$$

and the optimization problem in Equation(4.30) is solved by:

$$I_t = \max_{i \in \{1, \dots, K\}} \left\{ \max_{\beta \in C_t} \langle \theta_i, \beta \rangle \right\}.$$

Finally, the authors proposed an alternative version of `OFUL` where $\widehat{\beta}$ is recomputed only in the case where $\det \overline{V_t}$ is increased by a factor $(1 + C)$ with C a user-defined parameter. As showed by Abbasi-Yadkori et al. (2011), this variant exhibited both theoretical and empirical benefits in addition to the reduction of the computational cost. Note however that it requires the challenging tuning of the extra parameter C.

4.6.2 Thompson Sampling for contextual linear bandits

`Thompson sampling` is a family of (pseudo-)Bayesian algorithms. This section considers the `Thompson sampling` as defined in (Agrawal and Goyal, 2012a) (Algorithm 15).
`Thompson sampling` and `OFUL` share the same setting; in particular they do not consider any context X_t, the vector θ_i associated to the i-th arm is assumed to be known, and both algorithms aim at learning a single shared parameter β[2].

[2]While Agrawal and Goyal (2012a) refer to θ_i as "context vectors", it must be noted that the notion of context vector differs with the one considered in this manuscript.

Algorithm 15 Contextual linear `Thompson sampling` for K arms . (Agrawal and Goyal, 2012a)

Require: Current time t, parameter $v \in \mathbb{R}$
1: **if** $t = 1$ **then**
2: Initialize $B \leftarrow I_d$ and $\widehat{\beta} \leftarrow 0_d$, $f \leftarrow 0_d$
3: **end if**
4: Sample $\widehat{\beta_t} \sim \mathcal{N}(\widehat{\beta}, v^2 B^{-1})$
5: Play $I_t = \text{argmax}_{i \in \{1,\dots,K\}} \theta_i^T \widehat{\beta_t}$ (arbitrary tie break) and observe reward Y_t
6: Update $B \leftarrow B + \theta_{I_t}\theta_{I_t}^T$, $f \leftarrow f + Y_t\theta_{I_t}$, $\widehat{\mu} \leftarrow B^{-1}f$.

`Thompson sampling` is built around the assumptions of Gaussian likelihood and Gaussian prior:

- The **likelihood** of the instantaneous reward Y_t is supposed drawn with respect to $\mathcal{N}(\theta_{I_t}^T\beta^\star, v^2)$ with $v \overset{\text{def}}{=} R_{TS}\sqrt{\frac{24}{\varepsilon}d\log\frac{1}{\delta}}$.

  ε is a confidence parameter, δ a confidence level and R_{TS} a positive value such that η_t is a R_{TS} sub-Gaussian noise.

- The **prior** distribution for $\beta^\star$ is assumed to be $\mathcal{N}(\widehat{\beta}(t), v^2 B(t)^{-1})$ with $B(t) \overset{\text{def}}{=} I_d + \sum_{s=1}^{t-1}\theta_{I_s}\theta_{I_s}^T$ and $\widehat{\beta}(t) \overset{\text{def}}{=} B(t)^{-1}\left(\sum_{s=1}^{t-1}\theta_{I_s}Y_s\right)$

The assumptions are required for the computation of the posterior distribution and the theoretical analysis. However, and as emphasized in the article, these assumptions do *not* restrict the applicability of the approach. No further assumption on the reward distribution is required apart of the sub-Gaussian noise.

4.6.3 LinUCB

Like `OFUL`, `LinUCB` is an optimistic algorithm inspired from `UCB` (Li et al., 2010) (Algorithm 16).
Contrarily to `OFUL` and `Thompson sampling`, `LinUCB` (with disjoint linear models, see Section 3.1 in (Li et al., 2010)) considers an unknown parameter θ_i per arm, which is to be learned using the contextual information. Actually, Li et al. (2010) consider that context vector $X_{i,t}$ can depend on the arm.
By denoting $\mathcal{S} = \left(\left(X_{i_1}, Y_{i_1}\right)\dots\left(X_{i_S}, Y_{i_S}\right)\right)$ a context-reward sample of size S, $\mathbf{X}(\mathcal{S}) \overset{\text{def}}{=} (X_{i_1},\dots,X_{i_S})^T$ the $S \times d$ context matrix, $\mathbf{Y}(\mathcal{S}) \overset{\text{def}}{=} (Y_{i_1},\dots,Y_{i_S})^T$ the corresponding reward vector and $\mathcal{S}_{i,t} \overset{\text{def}}{=} \left\{(X_{t'}, Y_{t'}), t' \leqslant t, I_t = i\right\}$ the set of observations where i is played, `LinUCB` estimates at time t the parameter θ_i by :

$$\widehat{\theta}_i(\mathcal{S}_{i,t}) \overset{\text{def}}{=} \left(\mathbf{X}(\mathcal{S}_{i,t})^T\mathbf{X}(\mathcal{S}_{i,t}) + I_d\right)^{-1}\mathbf{X}(\mathcal{S}_{i,t})^T\mathbf{Y}(\mathcal{S}_{i,t})$$

Algorithm 16 `LinUCB` for K arms. (Li et al., 2010)

Require: Current time t, context X_t, parameter $\delta \in (0,1)$

1: **if** $t = 1$ **then**
2: Initialize $B \leftarrow I_d$, $\widehat{\mu} \leftarrow 0_d$, $i \in \{1,\dots,K\}$
3: **else**
4: $\widehat{\theta}_i \leftarrow A_i^{-1} b_i$, $i \in \{1,\dots,K\}$
5: $p_{t,i} \leftarrow \widehat{\theta}_i^{\,T} X_t + \alpha\sqrt{X_t^T A_i^{-1} X_t}$, $i \in \{1,\dots,K\}$ (α defined in Equation (4.32))
6: **end if**
7: Play $I_t = \text{argmax}_{i\in\{1,\dots,K\}}\{p_{i,t}\}$ (arbitrary tie break) and observe reward Y_t.
8: Update $A_{I_t} \leftarrow A_{I_t} + X_t X_t^T$ and $b_{I_t} \leftarrow b_{I_t} + Y_t X_t$

(Line 4).

By Theorem 2.1 of (Walsh et al., 2009), it holds that, with probability at least $1-\delta$, for every context vector X_t and every arm i:

$$|X_t^T \widehat{\theta}_i - X_t^T \theta_i| \leqslant \alpha\sqrt{X_t^T \left(\mathbf{X}(\mathscr{S}_{i,t-1})^T \mathbf{X}(\mathscr{S}_{i,t-1}) + I_d\right)^{-1} X_t},$$

with

$$\alpha \stackrel{\text{def}}{=} 1 + \sqrt{\frac{\log(2/\delta)}{2}} \tag{4.32}$$

a constant parametrized by δ. By defining $A_{i,t} \stackrel{\text{def}}{=} \left(\mathbf{X}(\mathscr{S}_{i,t-1})^T \mathbf{X}(\mathscr{S}_{i,t-1}) + I_d\right)$, it implies that $p_{i,t} \stackrel{\text{def}}{=} \widehat{\theta}_i^{\,T} X_t + \alpha\sqrt{X_t^T A_{i,t}^{-1} X_t}$ is an upper confidence bound on the expected payoff $\theta_i^T X_t$ of the arm i. In a `UCB` fashion, the arm with highest confidence bound is selected (Line 7).

Finally, note that the authors do not provide a regret bound for `LinUCB`. However a variant termed `SupLinUCB` is studied in (Chu et al., 2011) with a $O\left(\sqrt{Td\log^3\left(KT\log T/\delta\right)}\right)$ high-probability problem-dependent regret bound.

Chapter 5

Risk-Aversion

Contents

5.1 Introduction

The term *risk* is defined by the Oxford English dictionary as "the possibility of loss, injury, or other adverse or unwelcome circumstance". Regarding this definition, *risk-aversion* would design any method including a protecting mechanism against unwelcome outcomes. This definition hardly enables a statistical or computational analysis, as the notion of "unwanted outcomes" does not accept a unique definition.
As firstly introduced in Section 1.4, there are situations where the standard expectation maximization seems inappropriate. For instance, in the clinical testing example, one can think of:

- A treatment A with constant and moderate expected efficiency $\mu_A = 0.4$ and small variability $\sigma_A = 0.1$.
- A treatment B with better expected efficiency $\mu_B = 0.5$ and high variability $\sigma_B = 0.2$.
- A treatment C with $\mathbb{P}\left(eff = 0.4\right) = 0.2$ and $\mathbb{P}\left(eff = 0.5\right) = 0.8$

Depending on the medical context and physician's preferences, one of the three treatments can be favored. The usual player, interested in the best average outcome would favor arm A. On the other hand, a player reluctant to high performance variability may prefer the treatment B even at the cost of a slightly lower expected payoffs. Finally, a player concerned with the performances in the worst case scenario would favor the treatment C.
The purpose of this section is to present a few definitions of statistical tools to take risk into account, notably the coherency notion in Section 5.1.1, formalizing the above intuitions. Based on such risk measures, risk-averse algorithms will be presented in the context of multi-armed bandit in Section 5.2. One of our contributions will be new bandit algorithms based on a distinct coherent risk measure (Chapter 6).

5.1.1 Coherent risk measure

As illustrated in the previous section, the notion of risk is ambiguous and should be adapted to the specific problem studied. However, certain properties of measures are considered desirable and are the object of a consensus leading to the definition of *coherent risk measures*. This definition is extracted from the financial literature where the notion of risk emerges naturally.

Definition 5.1.1 (Rockafellar (2007))**.** *Let* $\mathscr{L}^2 \overset{\text{def}}{=} \{X : \mathbb{E}[X^2] < \infty\}$ *be the set of random variables with finite moments of order* 2.
A function $\mathscr{R} : \mathscr{L}^2 \to (-\infty, \infty]$ *is a* ***coherent risk measure*** *if it satisfies the following axioms:*

- *(A1)* $\mathscr{R}(C) = C$ *for all constant* C
- *(A2)* ***Convexity****:* $\forall \lambda \in (0,1), \mathscr{R}((1-\lambda)X + \lambda Y) \leqslant (1-\lambda)\mathscr{R}(X) + \lambda \mathscr{R}(Y)$
- *(A3)* ***Monotonicity****:* $\mathscr{R}(X) \leqslant \mathscr{R}(Y)$ *for* $X \leqslant Y$
- *(A4)* ***Closedness****:* $\mathscr{R}(X) \leqslant 0$ *when* $\|X_k - X\|_2 \to 0$ *with* $\mathscr{R}(X_k) \leqslant 0 \forall k$.

Remark 5.1.1. *Artzner et al. (1997, 1999) add the fifth axiom:*

- *(A5)* ***Positive homogeneity****:* $\forall \lambda > 0, \mathscr{R}(\lambda X) = \lambda \mathscr{R}(X)$.

The notion of coherency is coming from the financial literature, and the above axioms have a intuitive interpretation within this realm. Notably:

- the monotonicity assumption states that, a portfolio Y with higher values on almost every scenario than a portfolio X has a greater risk value.
- the combination of (A2) and (A5) leads to the **subadditivity**: $\mathscr{R}(X+Y) \leqslant \mathscr{R}(X) + \mathscr{R}(Y)$. This corresponds to the *diversity* principle in investment: the risk is lowered when the portfolio is diversified.

Examples

- The conditional Value-at-Risk at level α, informally defined as the average over the α% worst cases, is a coherent measure of risk (see (Acerbi and Tasche, 2002)). It will serve as the basis of the proposed algorithms for risk-awareness (Chapter 6).
- The Value at risk at level α defined by $VaR_\alpha(X) = \inf\{s, \mathbb{P}(X \leqslant s) \geqslant \alpha\}$ is not a coherent risk measure as it is not sub-additive. Indeed, consider X and Y i.i.d variables with $X \sim Ber(0.5)$. For $\alpha = 0.3$, $VaR_\alpha(X+Y) = 1 > 0 = VaR(X) + VaR(Y)$.

Risk-Aversion for the multi-armed bandit can then be defined, by replacing the classical expectation with a risk measure.

5.2 Risk-Aversion for the multi-armed bandit

The purpose of this section is to present two families of risk-averse bandit algorithms. First the `MVLCB` and `ExpExp` algorithms based on the Mean-Variance measure will be presented in Section 5.2.1, then `RA-UCB` based on the log-Laplace measure will be exposed in Section 5.2.2.
More remotely related is the work presented by Yu and Nikolova (2013), deriving PAC lower bounds for three distinct risk measures, with the goal of identifying the arm with lowest risk (pure exploration), as opposed to, minimizing a suitable regret.

5.2.1 Algorithms for the Mean-Variance

This section describes two algorithms proposed in (Sani et al., 2012a,b) based on a standard risk measure introduced by (Markowitz, 1952).

Mean-Variance

Definition 5.2.1 (Mean-Variance Markowitz (1952))**.** *Let $X \sim \nu$ be a random variable and $\rho > 0$ be a risk tolerance. By denoting, μ and σ^2 respectively the expected value and variance of X, the **Mean-Variance** of X with respect to ρ is defined as:*

$$MV_\rho(X) \stackrel{\text{def}}{=} \sigma^2 - \rho\mu \tag{5.1}$$

Remark 5.2.1. *Mean-Variance is not a coherent risk measure. Indeed, by considering a constant C and $\rho \notin \{0,1\}$, it comes $MV(C) = \rho C \neq C$.*

Remark 5.2.2. *In a risk-averse setting, MV is to be minimized.*

Remark 5.2.3. *This definition is an intuitive introduction to the risk management: in this setting, the risk of a variable is defined by its variance. Between two random variables X and Y with close expected payoffs, the one with smallest variance is favored.*

Remark 5.2.4. *Parameter ρ controls the trade-off between the risk-minimization (small ρ) and reward maximization (large ρ). In the extreme cases, MV minimization boils down to variance minimization for $\rho = 0$ and to the classical expected payoff maximization for $\rho \to \infty$.*

In a standard way, for $\rho > 0$: i) MV_i denotes the mean-variance of the i-th arm with distribution ν_i; ii) the best arm is denoted $i^\star \overset{\text{def}}{=} \underset{i\in\{1\dots K\}}{\operatorname{argmin}}\{MV_i\}$; and iii) the margin $\Delta_{MV,i} = MV_i - MV_{i^\star}$.

Definition 5.2.2 (Estimator)**.** *Let $x_1,\dots,x_n$ be a i.i.d. sample of a distribution ν and $X \sim \nu$. The estimator of $MV_\rho(X)$ is defined by:*

$$\widehat{MV_\rho}(X) \overset{\text{def}}{=} \widehat{\sigma_n}^2 - \rho\widehat{\mu_n} \tag{5.2}$$

with

$$\widehat{\mu_n} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} x_i \quad \textit{and} \quad \widehat{\sigma_n}^2 \overset{\text{def}}{=} \sum_{i=1}^{n}\left(x_i - \mu_n\right)^2$$

For a fixed $\rho > 0$ (implicit in the following), the mean-variance of the rewards $Y_1,\dots,Y_t$ gathered by an algorithm $\mathscr{A}$ up to time t is denoted $\widehat{MV}_t(\mathscr{A})$. Similarly, $\widehat{MV}_{i,t}$ denotes the mean-variance estimate built on $X_{i,1},\dots,X_{i,t}$.
The authors define in a classical way the notions of regret and pseudo-regret with respect to Mean-Variance.

Definition 5.2.3 (Regret)**.** *The regret of an algorithm $\mathscr{A}$ at time t is defined by:*

$$R_{\rho,t} \overset{\text{def}}{=} \widehat{MV}_t(\mathscr{A}) - \widehat{MV}_{i^\star,t} \tag{5.3}$$

Remark 5.2.5. *Two independent and distinct sets of samples are used for the computation of $\widehat{MV}_t(\mathscr{A})$ and $\widehat{MV}_{i^\star,t}$.*

Remark 5.2.6. *Contrarily to the usual cumulative regret, the mapping $t \mapsto \mathscr{R}_t$ is not necessarily a non-decreasing mapping. Actually, any reasonable Mean-Variance sensitive algorithm should present a decreasing regret with t.*

Definition 5.2.4 (Pseudo-regret)**.** *The **Mean-Variance pseudo-regret** after t time step is defined as:*

$$\overline{R}_{\rho,t} \overset{\text{def}}{=} \frac{1}{t}\sum_{i=1}^{K} N_{i,t}\Delta_i + \frac{2}{t^2}\sum_{i=1}^{K}\sum_{j\neq i} N_{i,t}N_{j,t}\Gamma_{i,j}^2 \tag{5.4}$$

with $\Gamma_{i,j} \overset{\text{def}}{=} \mu_i - \mu_j$

Sani et al. (2012a) propose the following notations for the two parts of the pseudo-regret:

$$\overline{R}^{\Delta}_{\rho,t} \overset{\text{def}}{=} \frac{1}{t}\sum_{i=1}^{K} N_{i,t}\Delta_i \tag{5.5}$$

$$\overline{R}^{\Gamma}_{\rho,t} \overset{\text{def}}{=} \frac{2}{t^2}\sum_{i=1}^{K}\sum_{j!=i} N_{i,t}N_{j,t}\Gamma^2_{i,j} \tag{5.6}$$

The following proposition establishes a relation between the regret upper bound, and the pseudo-regret upper bound, as follows:

Proposition 5.2.1 (Sani et al. (2012a))**.** *For $\rho > 0$ and with probability at least $1-\delta$,*

$$R_{\rho,t} \leqslant \overline{R}_{\rho,t} + (5+\rho)\sqrt{\frac{2K\log(6nK/\delta)}{t}} + 4\sqrt{2}\frac{K\log(6nK/\delta)}{t}$$

Mean-Variance Lower Confidence Bound

A first algorithm, termed Mean-Variance Lower Confidence Bound (`MVLCB`), is proposed to minimize $\overline{R}_t$ and its pseudo-code is presented in Algorithm 17.

Algorithm 17 K-armed `MVLCB`

Require: Time horizon T, confidence $\delta \in (0,1)$

1: **for** $t = 1 \dots K$ **do**

2: $I_t = t$; gather Y_t; initialize $\widehat{MV_i}(X_{i,1}) = -\rho Y_t$

3: **end for**

4: **for** $t = K+1 \dots T$ **do**

5: Choose (arbitrary tie break)

$$I_t \in \underset{i\in\{1,\dots,K\}}{\operatorname{argmin}} \left\{ \widehat{MV_{i,N_{i,t}}} - (5+\rho)\sqrt{\frac{\log(1/d)}{2N_{i,t}}} \right\}. \tag{5.7}$$

6: **end for**

The algorithm is designed in a classical fashion as an optimistic bandit approach. After initialization (Line 2), a lower bound $B_{i,N_{i,t}} \overset{\text{def}}{=} \widehat{MV_{i,N_{i,t}}} - (5+\rho)\sqrt{\frac{\log(1/d)}{2N_{i,t}}}$ is updated for each arm i and the arm with lowest lower bound is selected.

Note that `MVLCB` remains an optimistic algorithm as *minimizing* a *lower* bound is equivalent to the *maximization* of an *upper* bound.

Thanks to that design, and to the applicability of the Hoeffding inequality (Hoeffding, 1963) to derive a high-probability confidence interval for the MV_i, the following

theorem ensures a logarithmic regret both in expectation and with high probability in the case of a unique optimal arm.

Theorem 9 ((Sani et al., 2012a,b)). *Under the assumption of a unique optimal arm denoted $i^\star$, with $b = 2(5+\rho)$, MVLCB pseudo regret is upper-bounded with probability at least $1-6tK\delta$ as:*

$$\overline{R_{\rho,t}} \leqslant \frac{b^2 \log(1/\delta)}{n}\left(\sum_{i\neq i^\star}\frac{1}{\Delta_{MV,i}} + 4\sum_{i\neq i^\star}\frac{\Gamma^2_{i^\star,i}}{\Delta^2_{MV,i}} + \frac{2b^2\log(1/\delta)}{t}\sum_{i\neq i^\star}\sum_{j\notin\{i,i^\star\}}\frac{\Gamma^2_{i,j}}{\Delta^2_{MV,i}\Delta^2_{MV,j}}\right) + \frac{5K}{t} \tag{5.8}$$

For $\delta = \frac{1}{t^2}$, the pseudo-regret is also bounded in expectation by:

$$\mathbb{E}\left[\overline{R_{\rho,t}}\right] \leqslant \frac{2b^2\log t}{t}\left(\sum_{i\neq i^\star}\frac{1}{\Delta_{MV,i}} + 4\sum_{i\neq i^\star}\frac{\Gamma^2_{i^\star,i}}{\Delta_{MV,i}} + \frac{4b^2\log t}{t}\sum_{i\neq i^\star}\sum_{j\notin\{i,i^\star\}}\frac{\Gamma^2_{i,j}}{\Delta^2_{MV,i}\Delta^2_{MV,j}}\right) + (17+6\rho)\frac{K}{t} \tag{5.9}$$

This result proves that $\mathscr{R}_{\rho,t} = O(\frac{\log^2 t}{t})$ and the consistency of the approach. However, and as pointed by (Sani et al., 2012a,b; Maillard, 2013) and by the Equations (5.5) (5.6), the regret definition adds an extra penalization for the exploration of arms with distinct means. In a worst case scenario ($K=2$, $\rho=0$, $\mu_1 \neq \mu_2$), this would lead to a constant regret. To overcome this difficulty, the authors propose a new algorithm, decoupling exploration and exploitation and presented in the next section. Another approach is described in 5.2.2).

Remark 5.2.7. *The unicity assumption is essential. As shown in (Sani et al., 2012a, Theorem 1, remark 2), if this hypothesis does not hold true, it is possible to exhibit an environment where MVLCB suffers a constant regret as the regret increases by switching from one arm to another.*

Exploration-Exploitation algorithm

The Exploration-Exploitation (ExpExp) algorithm is described in Algorithm 18.
The scheme is extremely simple. For a given and parametrized budget $\tau \in \mathbb{N}$, the algorithm performs an uniform exploration (Equation 5.10). In a second phase, starting at time $t = \tau+1$ until the end, the best estimated arm is always selected.
The idea of stopping the exploration after a given budget overcomes the previously identified problem in the worst case. More precisely, it yields the following, problem independent, regret bound:

Theorem 10 (Sani et al. (2012a,b)). *Let ExpExp be run with $\tau = K(t/14)^{2/3}$. Then for any choice of distribution ν_i, the regret of ExpExp is bounded by:*

$$\mathbb{E}\left[\overline{R_{\rho,t}}\right] \leqslant 2\frac{K}{t^{1/3}} \tag{5.12}$$

Algorithm 18 K-armed ExpExp

Require: Time horizon T, $\tau \in \mathbb{N}$

1: **for** $t = 1 \dots \tau$ **do**

2: Choose (Exploration)

$$I_t = (t-1) \,\%K + 1 \tag{5.10}$$

3: **end for**

4: **for** $t = \tau + 1 \dots T$ **do**

5: Choose (Exploitation)

$$I_t = \underset{i \in \{1 \dots K\}}{\operatorname{argmin}} \left\{ \widehat{MV}_{i,t} \right\} \tag{5.11}$$

6: **end for**

Also, experiments show (see Section 6.5 and (Sani et al., 2012b)) the excellent performances of this approach. It is, however, important to emphasize the fact that the knowledge of the time horizon T is critical, although it might be unavailable in quite a few real-world applications.

5.2.2 Risk-Averse Upper Confidence Bound Algorithm

This section presents the algorithm *Risk-Averse Upper Confidence Bound* (RA-UCB) algorithm due to Maillard (2013).
This algorithm takes its roots in the KL-UCB (Cappé et al., 2013) algorithm. The main modification is the switch from a mean maximization framework to the maximization of a coherent risk measure, able to control the lower tail (mass below the mean) of the distributions.

log-Laplace

Definition 5.2.5 (Cumulant generating function)**.** *Let X be a real random variable. The **cumulant generating function** of X, g is defined as the logarithm of the moment-generating function of X:*

$$g(t) \overset{\text{def}}{=} \log \left\{ \mathbb{E}\left[\exp(tX) \right] \right\}$$

Proposition 5.2.2 (Maillard (2013))**.** *Let X be a random variable with a finite cumulant-generating function around 0 and $\delta \in (0,1)$ be a confidence level.*
Then, the following inequality holds:

$$\mathbb{P}\left(X \leqslant \sup_{\lambda > 0} \left\{ -\frac{1}{\lambda} \log\left(\mathbb{E}\left[\exp(-\lambda X) \right] \right) - \frac{\log(1/\delta)}{\lambda} \right\} \right) \leqslant \delta \tag{5.13}$$

Proof. The result follows from applications of the Markov and Jensen inequalities. □

This result demonstrates the key role played by the quantity $-\frac{1}{\lambda}\log\left(\exp\left[-\lambda X\right]\right)$ as it controls the probability that X is small. This quantity has thus been thoroughly studied.

Definition 5.2.6 (log-Laplace Transform)**.** *Let $X \sim \nu$ be a arbitrary random variable. The **(rescaled by λ) log-Laplace of** $\mathbf{X}$ is defined, for $\lambda \neq 0$ as:*

$$\kappa_{\lambda,\nu} \stackrel{\text{def}}{=} \frac{1}{\lambda}\log\left(\mathbb{E}\left[\exp\left(\lambda X\right)\right]\right) \tag{5.14}$$

Proposition 5.2.3 (Rockafellar (2007))**.** *For any distribution ν, $\kappa_{-\lambda,\nu}$ is a coherent risk measure (in the extended sense).*

Remark 5.2.8. *In the stricter and more classical coherency definition from (Artzner et al., 1997, 1999), $\kappa_{-\lambda,\nu}$ cannot be considered a coherent risk measure as it violates the positive homogeneity: $\exists\gamma > 0$ and ν s.t. $\kappa_{-\lambda,\gamma\nu} \neq \gamma\kappa_{-\lambda,\nu}$.*

Remark 5.2.9. *By supposing $\nu = \mathcal{N}(\mu,\sigma^2)$, it comes $\kappa_{-\lambda,\nu} = \mu - \frac{\lambda\sigma^2}{2}$. In particular, in the Gaussian case, the log-Laplace and the mean-variance criteria coincide.*

The quantity $\kappa_{-\lambda,\nu}$ satisfies by definition $\kappa_{-\lambda,\nu} \leqslant \mathbb{E}_\nu[X]$. Equation (5.13) indicates that $\kappa_{-\lambda,\nu}$ controls the probability that the random variable X is small. Therefore, it has a specific interest in a risk-averse scenario. As a consequence, the design of the proposed algorithm is to select the optimal arm(s) with maximal $\kappa_{-\lambda,\nu_i}$:

$$i^\star \in \underset{i\in\{1\dots K\}}{\operatorname{argmax}}\left\{\kappa_{-\lambda,\nu_i}\right\} \tag{5.15}$$

Algorithm

Let us denote:

- $\widehat{\nu_{i,t}} \stackrel{\text{def}}{=} \frac{1}{N_{i,t}}\sum_{s=1}^{t}\delta(X_{i,s})\mathbb{I}\{I_t = i\}$ the empirical cumulative distributive function of the i-th arm;
- $KL\left(\nu,\nu'\right)$ the Kullback-Leibler divergence between the distributions ν and ν'.
- $\mathfrak{M}_1^+\left(\mathbb{R}_B\right)$ the set of distribution defined on $\mathbb{R}_B \stackrel{\text{def}}{=} (-\infty, B)$.

Then the *Risk-Averse Upper Confidence Bound Algorithm* proposed by Maillard (2013) is defined as follows (Alg. 19).
An optimistic algorithm maximizing a log-Laplace criterion, `RA-UCB` works in two phases. First, an upper bound is associated with each arm (Line 1), then the arm with a maximal upper bound is selected (Line 2). The algorithm also requires two

Algorithm 19 `RA-UCB` (Maillard, 2013)

Require: Current time t, parameters λ, f function.

1: Compute

$$U(i) \overset{\text{def}}{=} \sup\left\{\kappa_{-\lambda,\nu} : \mathbf{K}\left(\widehat{\nu_{i,t}}, \kappa_{-\lambda,\nu}\right) \leqslant \frac{f(t)}{N_{i,t}}\right\} \tag{5.16}$$

with

$$\mathbf{K}\left(\widehat{\nu_{i,t}}, \kappa_{-\lambda,\nu}\right) \overset{\text{def}}{=} \inf\left\{KL\left(\widehat{\nu_{i,t}}, \nu\right) : \nu \in \mathfrak{M}_1^+(\mathbb{R}_B, \kappa_{-\lambda,\nu} \geqslant r\right\} \tag{5.17}$$

2: Select (arbitrary tie break)

$$I_t = \underset{i\in\{1\dots K\}}{\operatorname{argmax}}\{U(i)\} \tag{5.18}$$

parameters: first a function f generally chosen so that $f = O(\log(t))$ and $\lambda > 0$ a confidence parameter.
For its successful application, `RA-UCB` relies on two important properties:

1. A concentration inequality can be derived for the estimator $\mathbf{K}$ (see Proposition 2 in Appendix of (Maillard, 2013)):
   For a given $\lambda > 0$ and $\varepsilon > 0$:

$$\mathbb{P}\left(\mathbf{K}\left(\nu_t, \kappa_{-\lambda,nu}\right) > \varepsilon\right) \leqslant \varepsilon\,(t+2)\exp\left(-t\varepsilon\right)$$

   As already stated, this is a critical point to upper-bound the number of sub-optimal arm pulls and derive regret bounds.

2. On a practical side, the estimators $\mathbf{K}(\widehat{\nu}_t, r)$ are practically computable by a dual formulation of the optimization program of Equation (5.17) thanks to the Karush-Kuhn-Tucker optimality conditions (Boyd and Vandenberghe, 2004).

Regret analysis

Maillard (2013) naturally extends the pseudo-regret definition (Equation (3.3)) to its risk-averse setting.

Definition 5.2.7. *With same notations as in Section 3.1.1, denoting $i^\star$ an optimal arm defined in Equation (5.15), and considering a risk parameter $\lambda > 0$, the **risk-averse pseudo-regret** $\overline{R_t}(\lambda)$ is defined by*

$$\overline{R}_t(\lambda) \overset{\text{def}}{=} \sum_{i=1}^{K}\left(\kappa_{-\lambda,i^\star} - \kappa_{-\lambda,i}\right)\mathbb{E}\left[N_{i,t}\right] \tag{5.19}$$

Under assumptions regarding `RA-UCB` parametrization, the following proposition demonstrates the logarithmic regret of the strategy.

Proposition 5.2.4 (Maillard (2013))**.** *By defining:*

- $f(t) = \log\left(2e(t+e)t^2/\gamma\right)$ *with* $\gamma = \Theta(T^{-1})$
- $\varepsilon_i > 0$ *a problem dependent constant characterizing the difficulty to assess if* i *is close to optimal (see the paper for further details).*
- $\mathbf{K}_i \overset{\text{def}}{=} \inf\left\{KL(\nu_i||\nu) : \nu \in \mathfrak{M}_1^+(\mathbb{R}_B)\kappa_{-\lambda,\nu} > \kappa_{-\lambda,\nu_i}\right\}$
- $\Delta_{\kappa,i} \overset{\text{def}}{=} \kappa_{-\lambda,i^\star} - \kappa_{-\lambda,i}$ *the quality gaps.*

Then,

$$\overline{R_t}(\lambda) \leqslant 5 \sum_{i \neq i^\star} \frac{(1+\varepsilon_i)\,\Delta_{\kappa,i}}{\mathbf{K}_i} \log(t) + O(1) \tag{5.20}$$

This result confirms the ability of `RA-UCB` to reach a logarithmic regret rate even in the challenging risk-averse setting. Moreover, Equation (5.20) extends the known expected regret bound to the risk-averse case as the constant $\frac{\Delta_{\kappa,i}}{\mathbf{K}_i}$ in front of the logarithmic factor is closely related to the theoretical constants (Lai and Robbins, 1985; Burnetas and Katehakis, 1996) or to `KL-UCB` (Maillard et al., 2011; Cappé et al., 2013) inequalities.

Part II

Contributions

Chapter 6

Risk-Awareness for Multi-Armed Bandits

Contents

This chapter presents our first contributions, focused on risk-awareness in multi-armed bandits. The limit case, where the learner aims at maximizing its minimal reward, is first introduced and it is shown that under mild hypotheses, an algorithm with good guarantees can be obtained. The general case is then studied, where the goal is to find the best arm in terms of *conditional value at risk*, and the `MARAB` algorithm tackling this goal is presented. Its empirical behavior is comparatively assessed to the state of the art, the `MVLCB` and `ExpExp` algorithms presented in chapter 5. These contributions were first presented in (Chou et al., 2014) and (Galichet et al., 2013).
A refinement of `MARAB` called `MARABOUT` is finally described, together with its performance guarantees.

6.1 Motivations

As said, the multi-armed bandit is a widely studied problem. This in-depth examination has led to the fruitful development of numerous algorithms able to handle the exploration *vs.* exploitation dilemma efficiently, often associated with guarantees of performance. However, the vast majority of these approaches assess the quality of an arm with respect to its payoff value in expectation. Even though this criterion is natural and successfully applicable to real world problems, it is inappropriate in quite a few situations. A first, already discussed example is the medical testing. Other examples come from real-world problems involving an environment with hazards and risks. For instance, roboticists often face a problem termed *Reality Gap* (see (Nolfi and Floreano, 2000)): controllers learned by simulation *in silico* are not able to reproduce the desired behavior *in situ*. A common approach to solve this issue is to directly learn a controller on a physical robot. Whenever the real-world environment is not considered to be safe, and hazards may eventually lead to the destruction of the robot, the need for a risk-aware learning process naturally emerges.
In such contexts, the focus shifts from the average reward associated to an arm, to the average reward taken over the α% worst cases. The (domain and user-dependent) parameter α is referred to as risk parameter in the rest of the chapter. We shall first consider the limit case where α goes to 0 (section 6.2), before considering the general case (section 6.3).

6.2 The max-min approach

This section presents an overview of the `MIN` algorithm, addressing the limit case when risk parameter α goes to 0.

6.2.1 Algorithm definition

Let $m_{i,t}$ denote the min empirical reward associated to the i-th arm up to the t-th time step:

$$m_{i,t} \overset{\text{def}}{=} \min\{Y_u \; s.t. \; I_u = i, \; u = 1 \dots t\}$$

The `MIN` algorithm (Algorithm 20) proceeds by initializing the empirical minimum of all arms (Line 1); thereafter, it systematically selects the arm with maximal $m_{i,t}$ (Line 2).

Algorithm 20 `MIN` for K arms

Require: Time horizon T

1: **for** $i = 1 \dots K$ **do**
2: Select the i-th arm, and observe the reward Y_i;
3: $m_{i,i} = Y_i$
4: **end for**
5: **for** $t = K+1 \dots T$ **do**
6: Select (arbitrary tie break)

$$I_t \in \underset{i \in \{1,\dots,K\}}{\operatorname{argmax}} \{m_{i,t}\} \tag{6.1}$$

7: Update $m_{I_t,t}$
8: **end for**

The most simple `MIN` algorithm allows for an online update of the $m_{i,t}$, which are the only quantities that need to be stored.

6.2.2 Analysis

The goal of `MIN` is to find the best essential infimum of the arms, where the essential infimum is defined as follows.

Definition 6.2.1. *Let ν be a probability distribution and $X \sim \nu$ a real random variable. The **essential infimum** a_ν of ν is defined by*

$$a_\nu \overset{\text{def}}{=} \max_{a \in \mathbb{R}} \{\mathbb{P}(X < a) = 0\}$$

Let us make the mild assumption that distribution ν satisfies Equation 6.2, as illustrated in Fig. 6.1. Then, the empirical min taken over a uniform sampling according to ν, converges exponentially fast toward the essential infimum.

Lemma 6.2.1. *Let ν be a bounded distribution with support in $[0,1]$, with a its essential infimum, and assume that ν satisfies:*

$$\exists A > 0, \forall \varepsilon > 0, \mathbb{P}(X \leqslant a + \varepsilon) \geqslant A\varepsilon \quad \textit{with } X \sim \nu \tag{6.2}$$

Let $x_1 \dots x_t$ be a t-sample independently drawn after ν. Then, the minimum value over $x_u, u = 1 \dots t$ goes exponentially fast to a:

$$\mathbb{P}(\min_{1 \leqslant u \leqslant t} x_u \geqslant a + \varepsilon) \leqslant \exp(-tA\varepsilon) \tag{6.3}$$

Proof. As the x_u are iid, it comes:

$$\begin{aligned} \mathbb{P}(\min_{1 \leqslant u \leqslant t} x_u \geqslant a + \varepsilon) &= \mathbb{P}(\forall u \in \{1, \dots, t\}, x_u \geqslant a + \varepsilon) \\ &= \prod_{u=1}^{t} \mathbb{P}(x_u \geqslant a + \varepsilon) \leqslant (1 - A\varepsilon)^t \leqslant exp(-tA\varepsilon) \end{aligned}$$

where the last inequality follows from $(1 - z) \leqslant exp(-z)$. □

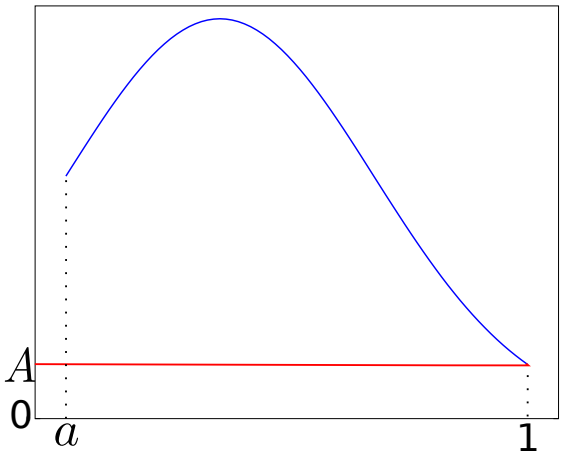


Figure 6.1: Illustration of an example of distribution satisfying the assumption of Equation 6.2.

The assumption supporting the above result (Equation 6.2, illustrated in Figure 6.1) does not require the positive constant A to be known; it only requires that there is enough probability mass in the neighborhood of a. Equation 6.3 confirms the exponential convergence toward a as a function of A.
A surprising result is that, under this assumption, the convergence toward the minimum might be faster than the convergence toward the mean. Specifically, the Hoeffding bound on the convergence toward the mean decreases exponentially like $-t\varepsilon^2$, whereas after Equation 6.3 the convergence toward the min decreases exponentially

like $-tA\varepsilon$ (as ε is going to 0, $A\varepsilon >> \varepsilon^2$).

Under this assumption, it follows without difficulty that with high probability the empirical min of each arm is exponentially close to its essential infimum after each arm has been tried t times.

Lemma 6.2.2. *Let $\nu_1 \dots \nu_K$ denote K distributions with bounded support in $[0,1]$ with a_i their essential infimum.*
Assume that ν_i satisfies Equation 6.2 for some constant A for $i = 1 \dots K$.
Denoting $x_{i,u}$, $u = 1 \dots t$, $i = 1 \dots K$, t samples independently drawn after ν_i, one has:

$$\mathbb{P}(\exists i \in \{1,\dots,K\}, \min_{1 \leqslant u \leqslant t} x_{i,u} \geqslant a_i + \varepsilon) \leqslant K \exp(-tA\varepsilon) \tag{6.4}$$

Proof. After Lemma 6.2.1,

$$\begin{aligned} \mathbb{P}(\exists i \in \{1,\dots,K\}, \min_{1 \leqslant u \leqslant t} x_{i,u} \geqslant a_i + \varepsilon) &\leqslant 1 - (1 - (1 - A\varepsilon)^t)^K \\ &\leqslant K(1 - A\varepsilon)^t \\ &\leqslant K \exp(-tA\varepsilon) \end{aligned}$$

Where the first inequality follows from $(1-z)^y \geq 1 - y.z$ and the second inequality from $(1-z) \leq exp(-z)$, which concludes the proof. □

Let us consider the two distinct goals of finding the arm with best expectation, and the arm with best essential infimum. If these goals are compatible (that is, the optimal arm in terms of min value also is the optimal arm in terms of mean value), then the MIN algorithm achieves a logarithmic regret under the above assumptions.

Proposition 6.2.1. *Let $\nu_1 \dots \nu_K$ denote K distributions with bounded support in $[0,1]$ with μ_i (resp. a_i) their mean (resp. their essential infimum). Further assume that ν_i satisfies Equation 6.2 for some constant A for $i = 1 \dots K$, and that the arm with best mean value $\mu^\star$ also is the arm with best min value $a^\star$. Let $\Delta_{\mu,i} = \mu^\star - \mu_i$ (resp. $\Delta_{a,i} = a^\star - a_i$) denote the mean-related (resp. essential infimum-related) margins.*
Then, with probability at least $1-\delta$, the cumulative pseudo-regret is upper bounded as follows:

$$\overline{R_t} \leqslant \frac{K-1}{A} \frac{\Delta_{\mu,\max}}{\Delta_{a,\min}} \log\left(\frac{tK}{\delta}\right) + (K-1)\Delta_{\mu,\max} \tag{6.5}$$

with $\Delta_{a,\min} = \min_{i:\Delta_{a,i}>0} \Delta_{a,i}$ and $\Delta_{\mu,\max} = \max_{i:\Delta_{\mu,i}>0} \Delta_{\mu,i}$.
Furthermore, the expectation of the cumulative pseudo-regret is upper-bounded as follows for t sufficiently large ($t \geqslant \frac{K-1}{A}\frac{\Delta_{a,\min}}{\Delta_{\mu,\max}}$):

$$\mathbb{E}[\overline{R_t}] \leqslant \frac{K-1}{A} \frac{\Delta_{\mu,\max}}{\Delta_{a,\min}} \left(\log\left(\frac{t^2 KA}{K-1} \frac{\Delta_{a,\min}}{\Delta_{\mu,\max}}\right) + 1\right) + (K-1)\Delta_{\mu,\max} \tag{6.6}$$

Proof. Suppose that there exists a single optimal arm (this point will be discussed below). Taking inspiration from (Sani et al., 2012a), let $x_{i,u}$ be independent samples drawn after ν_i, and define the event $\mathscr{E}$ as follows:

$$\mathscr{E} = \left\{ \forall i \in \{1,\dots,K\}, \forall s \in \{1,\dots u\} \min x_{i,s} - a_i \leqslant \frac{\varepsilon}{u} \right\} \tag{6.7}$$

The probability of the complementary event $\mathscr{E}^c$ is bounded after Lemma 6.2.2:

$$\begin{aligned} \mathbb{P}(\mathscr{E}^c) &= \mathbb{P}(\exists i \in \{1,\dots,K\}, \exists u \in \{1,\dots,t\}, \min_{1 \leqslant s \leqslant u} x_{i,s} - a_i > \frac{\varepsilon}{u}) \\ &\leqslant \sum_{u=1}^{t} \mathbb{P}(\exists i \in \{1,\dots,K\}, \min_{1 \leqslant s \leqslant u} x_{i,s} - a_i > \frac{\varepsilon}{u}) \\ &\leqslant \min(1, tK\exp(-A\varepsilon)) \end{aligned}$$

Let $t > 1$ be an iteration where a sub-optimal arm i is selected; this implies that the empirical min of the i-th arm is higher than that of the best arm $i^\star$:

$$\min_{1 \leqslant u \leqslant N_{i^\star,t-1}} x_{i^\star,u} < \min_{1 \leqslant u \leqslant N_{i,t-1}} x_{i,u} \quad \Leftrightarrow \quad \underbrace{\min_{1 \leqslant u \leqslant N_{i^\star,t-1}} x_{i^\star,u} - a_i}_{\geqslant a_{i^\star} - a_i = \Delta_{a,i}} < \underbrace{\min_{1 \leqslant u \leqslant N_{i,t-1}} x_{i,u} - a_i}_{\leqslant \frac{\varepsilon}{N_{i,t-1}} (\star)}$$

where $(\star)$ holds if t belongs to the event set $\mathscr{E}$, thus with probability at least $1 - tKexp(-A\varepsilon)$ after Lemma 6.2.2.
It follows that with probability at least $1 - tKexp(-A\varepsilon)$

$$\frac{\varepsilon}{N_{i,t-1}} \geqslant \Delta_{a,i} \text{ hence } N_{i,t} \leqslant \frac{\varepsilon}{\Delta_{a,i}} + 1$$

since $N_{i,t} \leq N_{i,t-1} + 1$.
With probability at least $1 - tKexp(-A\varepsilon)$, the cumulative regret $\overline{R_t}$ can thus be upper-bounded:

$$\begin{aligned} \overline{R_t} &= \sum_{i=1}^{K} N_{i,t}\Delta_{\mu,i} \leqslant \sum_{i=1}^{K} (\frac{\varepsilon}{\Delta_{a,i}} + 1)\Delta_{\mu,i} \qquad (6.8) \\ &\leqslant (K-1)\left(\frac{\Delta_{\mu,max}}{\Delta_{a,min}}\varepsilon + \Delta_{\mu,\max}\right) \text{ with } \Delta_{\mu,\max} = \max_{1 \leqslant i \leqslant K} \Delta_{\mu,i} \text{ and } \Delta_{a,\min} = \min_{1 \leqslant i \leqslant K} \Delta_{a,i} \end{aligned}$$

Finally, by setting $\delta = \min(1, tK\exp(-A\varepsilon))$, it follows that with probability $1 - \delta$,

$$\overline{R_t} \leqslant \frac{K-1}{A} \frac{\Delta_{\mu,max}}{\Delta_{a,min}} \log(\frac{tK}{\delta}) + (K-1)\Delta_{\mu,\max} \tag{6.9}$$

In the case where there exists $k > 1$ optimal arms, Eq. 6.9 still holds, by replacing $K - 1$

factor with $K-k$.
The expectation of the cumulative regret is similarly upper-bounded:

$$\begin{aligned}\mathbb{E}[\overline{R_t}] &= \mathbb{E}[\overline{R_t}\mathbb{I}_{\mathscr{E}}] + \mathbb{E}[\overline{R_t}\mathbb{I}_{\mathscr{E}^c}] \\ &\leqslant \frac{K-1}{A}\frac{\Delta_{\mu,max}}{\Delta_{a,min}}\log(\frac{tK}{\delta}) + (K-1)\Delta_{\mu,\max} + \delta t \text{ by bounding } \overline{R_t} \text{ by } t \text{ over } \mathscr{E}^C.\end{aligned}$$

For t sufficiently large ($t \geqslant \frac{K-1}{A}\frac{\Delta_{\mu,max}}{\Delta_{a,min}}$), by setting $\delta = \frac{K-1}{tA}\frac{\Delta_{\mu,max}}{\Delta_{a,min}}$, it comes:

$$\mathbb{E}[\overline{R_t}] \leqslant \frac{K-1}{A}\frac{\Delta_{\mu,max}}{\Delta_{a,min}}\left(\log\left(\frac{t^2KA}{(K-1)}\frac{\Delta_{a,\min}}{\Delta_{\mu,\max}}\right)+1\right) + (K-1)\Delta_{\mu,\max} \tag{6.10}$$

which concludes the proof. □

Remark 6.2.1. *This result can be compared to the regret bound derived for the* `UCB` *algorithm, similarly achieving a logarithmic regret (Auer et al., 2002):*

$$\mathbb{E}[\overline{R_t}] \leqslant 8\sum_{i\neq i^\star}\frac{\log t}{\Delta_{\mu,i}} + (1+\frac{\pi^2}{3})\sum_{i=1}^{K}\Delta_{\mu,i} \tag{6.11}$$

where $i^\star$ stands for the index of the optimal arm. `MIN` *and* `UCB` *thus both achieve a logarithmic regret uniformly over t, where the regret rate involves the mean-related margin in* `UCB` *(resp. the min-related margin in* `MIN`*, multiplied by the constant A).*

A stronger result can be obtained for `MIN`, under an additional assumption on the lower tails of the arm distributions.

Proposition 6.2.2. *With same notations and assumptions as in Prop. 6.2.1, let us further assume that for every $i = 1\dots K, \Delta_{\mu,i} = \mu^\star - \mu_i \leqslant a^\star - a_i = \Delta_{a,i}$.*

Then, with probability at least $1-\delta$,

$$\overline{R_t} \leqslant \frac{K-1}{A}\log(\frac{tK}{\delta}) + (K-1)\Delta_{\mu,\max}$$

with $\Delta_{\mu,\max} = \max_i \Delta_{\mu,i}$.
Furthermore, if $t > \frac{K-1}{A}$, the expectation of $\overline{R_t}$ is upper-bounded as follows:

$$\mathbb{E}[\overline{R_t}] \leqslant \frac{K-1}{A}\left(\log\left(\frac{t^2KA}{K-1}\right)+1\right) + (K-1)\Delta_{\mu,\max} \tag{6.12}$$

Proof. The proof closely follows the one of Prop. 6.2.1, noting that in Eq. 6.8 $\Delta_{a,i}$ is now greater than $\Delta_{\mu,i}$. Setting $\delta = \frac{(K-1)}{tA}$ concludes the proof of Eq. 6.12. □

Discussion. The comparison of UCB and MIN only makes sense when the two goals are the same, naturally, that is, the same arm is optimal in terms of expectation and in terms of essential infimum. When it is the case, Eq. 6.12 and Eq. 6.11 suggests that MIN might outperform UCB when: i) margins $\Delta_{\mu,i}$ are small, ii) distributions ν_i are not too thin in the neighborhood of the essential infimum (that is, A is not too small), and iii) the assumption $\Delta_{a,i} \geqslant \Delta_{\mu,i}$ holds.
Note that the latter assumption boils down to considering that better arms (in the sense of their mean) also have a narrower support for their lower tail, thus a lower risk. If this assumption does not hold however, then risk minimization and regret minimization are likely to be conflicting objectives.
A last remark is that the assumptions done (lower bounded distribution density in the neighborhood of the essential minimum and mean-related margin greater than the minimum-related margin) yield a significant improvement compared to the continuous distribution-free case, where the optimal regret is known to be $O(\sqrt{t})$ (Audibert and Bubeck, 2009, 2010).

6.3 Conditional Value At Risk: formal background

In the general case, the goal is, informally speaking, to maximize the average taken over the α% worst . Let us first introduce some notations and definitions.

6.3.1 Definitions

Definition 6.3.1 (Value at risk)**.** *Let X be a real random variable and $\alpha \in (0,1]$.*

*The **Value at Risk at level** α of X (or α-quantile) is defined as:*

$$VaR_\alpha(X) \stackrel{\text{def}}{=} \inf_{\xi \in \mathbb{R}} \{P(X \leqslant \xi) \geqslant \alpha\} \tag{6.13}$$

Definition 6.3.2 (Conditional Value at Risk (Rockafellar and Uryasev, 2002; Pflug, 2000))**.** *Let X be a real random variable and $\alpha \in (0,1]$.*

*The **Conditional Value at Risk at level** α of X is defined as:*

$$CVaR_\alpha(X) \stackrel{\text{def}}{=} \inf_{\xi \in \mathbb{R}} \left\{ \xi + \frac{1}{1-\alpha} \mathbb{E}\left[(X-\xi)_+\right] \right\} \tag{6.14}$$

where $(x)_+$ denotes the positive part of x defined by:

$$(x)_+ = \begin{cases} x & x > 0 \\ 0 & x \leqslant 0 \end{cases}$$

Proposition 6.3.1. *(Acerbi and Tasche (2002)) If there is no probability atom at $VaR_\alpha(X)$ (in particular, if X is a continuous random variable), it holds:*

$$CVaR_\alpha(X) = \mathbb{E}[X|X > VaR_\alpha(X)] \tag{6.15}$$

Remark 6.3.1. *Intuitively, X is viewed as a loss, with $CVaR_\alpha(X)$ (the expected losses in the α percent worst cases) to be* minimized.

Remark 6.3.2. *The* Conditional Value at Risk *terminology is justified by Eq. 6.15. CVaR is also termed* Average value-at-risk. *Under restrictive assumptions (e.g. distribution continuity), CVaR coincides with several other risk criteria like* Expected shortfall *(see Acerbi and Tasche (2002)).*

Remark 6.3.3. *Definitions of VaR and $CVaR$ may vary depending of the authors; this might be confusing. For instance, (Acerbi and Tasche, 2002) denotes*

$$\begin{aligned} CVaR_\beta(Y) &\overset{\text{def}}{=} \inf_\xi \left\{ \frac{\mathbb{E}[(Y-\xi)_-]}{\beta} - \xi \right\} \\ &= -\mathbb{E}\left[Y | Y \leqslant VaR_\beta(Y) \right] \text{ (if } Y \text{ is continuous)} \end{aligned} \tag{6.16}$$

with $Y \overset{\text{def}}{=} -X$, $\beta \overset{\text{def}}{=} 1-\alpha$ and $(x)_- \overset{\text{def}}{=} -(-x)_+$

These two definitions remain equivalent but a particular attention should be given to determine if the "worst case" are represented by the left or right tail of the distribution.

As the bandit literature commonly considers the rewards as payoff to be *maximized*, and by taking inspiration from Equation (6.16), the approach proposed is to *maximize*, for a random variable X the quantity $-CVaR_\alpha(-X) (= \mathbb{E}[X|X < VaR_\alpha(X)])$. This is an equivalent procedure to the *minimization* of the conditional Value at Risk of $-X$. Hence, the next definition provides a name and notation for this new quantity.

Definition 6.3.3 (Modified Conditional Value at Risk)**.** *Let X be a random variable and $\alpha \in (0,1]$ a confidence parameter.*

The modified Conditional Value at Risk at level α of X is defined by:

$$mCVaR \overset{\text{def}}{=} -CVaR_\alpha(-X) \tag{6.17}$$

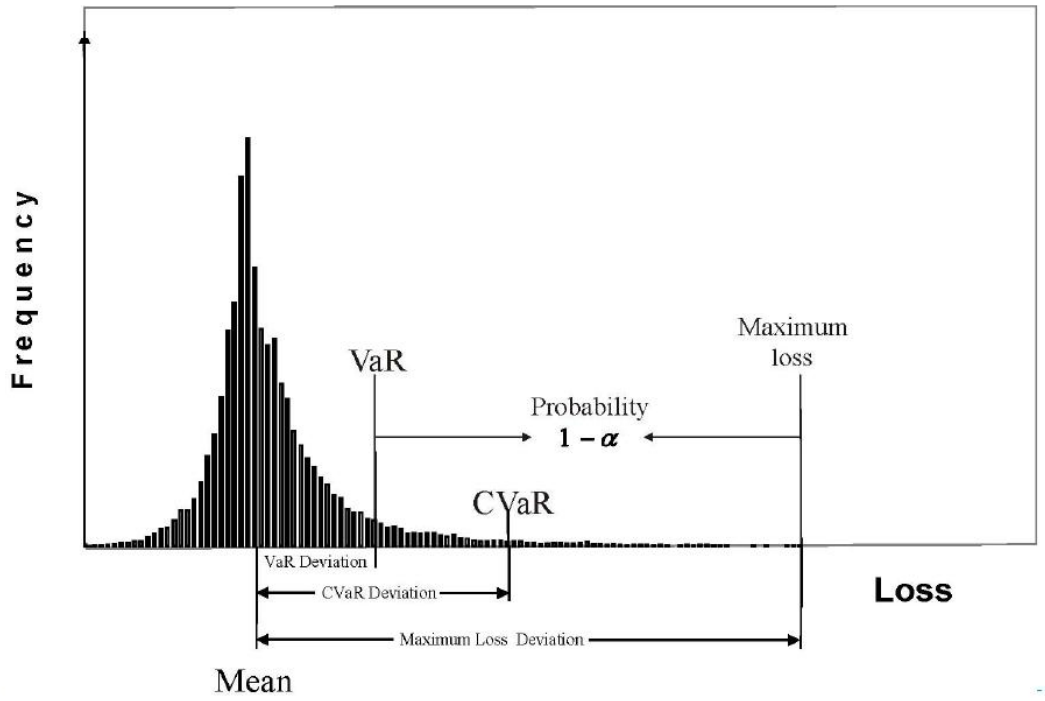


Figure 6.2: Value at risk and Conditional Value at Risk (from Rockafellar and Uryasev (2002)).

6.3.2 Estimation of the Conditional Value at Risk

Definition 6.3.4. *Let $\alpha \in (0,1]$ be a confidence level, and let $x_1, \dots, x_n$ be a sample of n i.i.d realizations of a distribution ν. Assuming without loss of generality that $x_1 \leqslant x_2 \leqslant \dots \leqslant x_n$, an estimator of $mCVaR_\alpha(X)$ with $X \sim \nu$ is defined by:*

$$\widehat{mCVaR}_\alpha(x_1, \dots, x_n) \stackrel{\text{def}}{=} \frac{1}{\lceil n\alpha \rceil} \sum_{i=1}^{\lceil n\alpha \rceil} x_i \tag{6.18}$$

where $\lceil n\alpha \rceil$ denotes the ceil integer of $n\alpha$.

Remark 6.3.4. *For $\alpha = 1$, the estimator boils down to the empirical average of x_i.*

Remark 6.3.5. *Under technical assumptions (weak dependency and bounded second derivative of the density), (Chen, 2008) shows that:*

$$\sqrt{\alpha n}\sigma_0^{-1}(\alpha, n)(\widehat{mCVaR}_\alpha(x_1, \dots, x_n) - mCVaR_\alpha(-\nu)) \xrightarrow{d} \mathcal{N}(0,1) \tag{6.19}$$

In particular, this result states that $\widehat{mCVaR}_\alpha(x_1, \dots, x_n)$ is a consistent estimator and converges to $mCVaR_\alpha(X)$ as $\sqrt{\alpha n}$.

6.4 The Multi-Armed Risk-Aware Bandit Algorithm

This section presents the MARAB algorithm, first described in (Galichet et al., 2013). Letting $N_{i,t}$ be the number of pulls of arm i, we denote $\widehat{mCVaR}_{\alpha,i}(X_{i,1},\ldots,X_{i,N_{i,t}}) = \widehat{mCVaR}_{\alpha}(X_{i,1},\ldots,X_{i,N_{i,t}})$

6.4.1 Description

The MARAB algorithm (Algorithm 21) starts by visiting once each arm, thus getting an estimate of the payoff in the α% worst cases thanks to Equation (6.18) (Line 2). A lower confidence bound associated with this estimate is derived, and MARAB selects the arm with highest lower confidence bound (Line 5).

Algorithm 21 K-armed MARAB

Require: Time horizon T; risk level α; exploration parameter $C > 0$.

1: **for** $t = 1 \ldots K$ **do**
2: $I_t = t$; gather Y_t; initialize $\widehat{mCVaR}_{\alpha,t}(X_{t,1}) = X_{t,1} = Y_t$
3: **end for**
4: **for** $t = K+1 \ldots T$ **do**
5: Choose (arbitrary tie break)

$$I_t \in \underset{i\in\{1,\ldots,K\}}{\operatorname{argmax}} \left\{ \widehat{mCVaR}_{\alpha,i}(X_{i,1},\ldots,X_{i,N_{i,t}}) - \sqrt{\frac{C\log\lceil t\alpha\rceil}{\lceil \alpha N_{i,t}\rceil}} \right\}. \tag{6.20}$$

6: **end for**

6.4.2 Discussion

A key difference compared to the *Optimism in front of the unknown* motto at the core of the UCB algorithm is that MARAB features a risk-averse or pessimistic behavior. The exploratory term in Equation (6.20) comes with a negative coefficient: a larger C value corresponds to higher bias toward exploitation, i.e. a more conservative behavior. In other terms, when facing two arms with same $\widehat{mCVaR}$ value, MARAB will select the *most visited* arm.

Note that such a behavior is representative of real-world behaviors, e.g. in the economic realm, where a bias toward known partners is at the core of economic exchanges. This bias is justified by the risk involved by choosing unknown partners.

According to the $mCVaR$ estimate defined in Equation (6.18), MARAB involves two successive phases:

1. In a first **initialization phase** ($N_{i,t} < \frac{1}{\alpha}$ and $\lceil N_{i,t} \rceil = 1$), one has

$$\widehat{mCVaR}_{\alpha,i}(X_{i,1},\ldots,X_{i,N_{i,t}}) = X_{(1)} = \min_{s\in\{1,\ldots,N_{i,t}\}} \{X_{i,s}\}$$

   the quality of the arm is assessed from its minimal value (thus monotonically decreasing along time).
   Indeed, the duration of the initial phase increases as α decreases toward 0. In this phase, the maximization of the $\widehat{mCVaR}_{\alpha,i}(X_{i,1},\ldots,X_{i,N_{i,t}})$ boils down to a standard max-min optimization problem and the `MARAB` behavior must resembles that of `MIN` (section 6.2, except for the negative exploratory term of `MARAB`). In these early iterations, the only exploration achieved by `MARAB` is due to the fact that $\widehat{mCVaR}_{\alpha,i}(X_{i,1},\ldots,X_{i,N_{i,t}})$ monotonically decreases with the number of trials $N_{i,t}$, possibly leading to revisit less visited arms.
   However, the pessimistic nature of the approach prevents the algorithm from visiting again an arm that provided poor rewards in the first trials.

2. In a second **stabilization phase**, the estimate $\widehat{mCVaR}_{\alpha,i}(X_{i,1},\ldots,X_{i,N_{i,t}})$ is computed with an increasing precision, the approximation error converging to 0 as $\sqrt{N_{i,t}}$ (Chen, 2008).
   Only the most played arms enter this second phase ($N_{i,t} \geq 1/\alpha$) and their empirical $\widehat{mCVaR}_{\alpha,i}$ tends to stabilize with time. Note however that, due to the lack of any exploration bonus, there is no guarantee that each arm is visited an infinite number of times as t goes to infinity.

6.5 Experimental validation

This section presents the empirical validation of `MIN` and `MARAB`, comparatively to `UCB` and to the risk-aware `MVLCB` and `ExpExp` algorithms (Sani et al., 2012a).

6.5.1 Experimental setting

The empirical validation considers three settings:

- `MIN` and `MARAB` are compared to `UCB` in the favorable case, satisfying the assumptions done in Prop 6.2.2 (Equation 6.2 satisfied, same arm ordering for the mean and for the essential infimum values with decreasing margin).

- A relaxed problem generator, only satisfying the assumption of Equation 6.2, is then considered to extensively compare `MIN` and `MARAB` to `UCB`, `MVLCB` and `ExpExp`.

- Last, a simplified real-world problem pertaining to the domain of energy management is considered (first presented in Chou et al. (2014)).

In all problems, the number K of arms is set to 20. The time horizon is set to $T = K \times 100$ and $T = K \times 200$. For all problems, all results over (respectively the average result out of) 40 runs are displayed.
The goal of experiments is to answer three questions:

- The first one regards the price to pay in terms of performance loss for a risk-aware behavior (in the favorable case where there is same arm ordering w.r.t. the mean and to the essential infimum criteria), and how the cumulative regret increases with the number of iterations, specifically focussing on short time horizons. Unless otherwise specified, only the empirical cumulative regret is considered.

- The second question regards the robustness of the algorithms, and their sensitivity w.r.t. parameters.

- The third question is whether `MARAB`, `MVLCB` and `ExpExp` do avoid exploring risky arms; this question is investigated by inspecting the low tail of the gathered rewards.

6.5.2 Proof of concept

The generator used in this experiment is meant to exactly satisfy the assumptions in Prop. 6.2.2.

- The i-th arm distribution ν_i is uniform on a segment in $[0,1]$, centered on its mean μ_i with radius r_i ($\nu_i = \mathcal{U}([\mu_i - r_i, \mu_i + r_i])$).

- Mean μ_i decreases with i ($\mu^\star = \mu_1 > \mu_2 \ldots > \mu_K$):

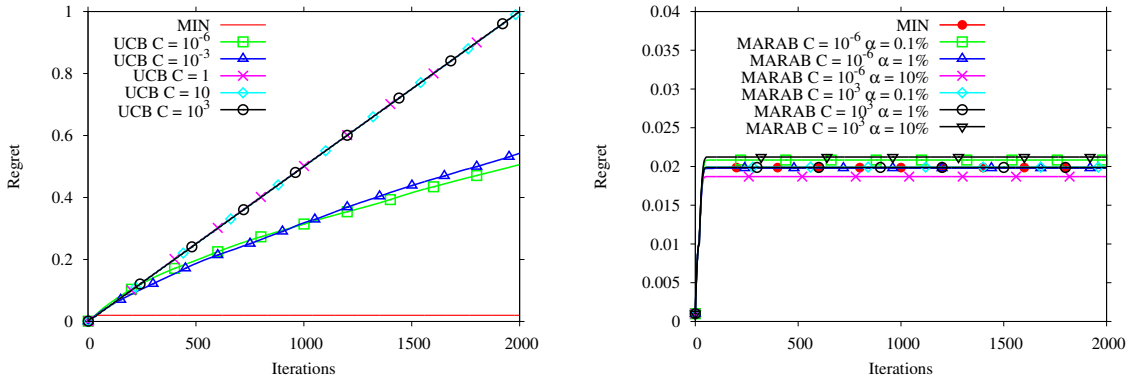
$$\mu_i = \mu^\star - \frac{i-1}{(K-1)} \Delta_{\max}$$

- Radius r_i increases with i: denoting $r_1 = \mu^\star - a^\star$,

$$r_i = r_1 + \frac{i-1}{K-1} r_{\max}$$

The generator, controlled from hyper-parameters $\Delta_{\max}$ and $r_{\max}$, thus enforces that the mean-related margin $\Delta_{\mu,i}$ is an increasing function of Δ_{max}; furthermore, the min-related margin $\Delta_{a,i}$ is controlled from Δ_{max} and r_{max}, in such a way that $\Delta_{a,i} > \Delta_{\mu,i}$.

The experimental comparison of MIN, UCB and MARAB in this favorable case is displayed on Fig. 6.3 (cumulative pseudo-regret averaged on 40 independent runs with $\mu^\star = 0.5$, $a^\star = \mu^\star - 10^{-3}$ and maximal radius 0.5). The risk parameter α ranges from .01 to .1. The exploration coefficient C in MARAB ranges in $\{10^{-6}, \ldots, 10^3\}$.



(a) MIN and UCB
$C = 10^i,\ i = -6 \ldots 3$

(b) MIN and MARAB
$C = 10^i,\ i = -6 \ldots 3,\ \alpha = .1\%, 1\%, 10\%)$

Figure 6.3: Cumulative pseudo-regret of UCB, MIN and MARAB under the assumptions of Prop. 6.2.2, averaged out of 40 runs. Parameter C ranges in $\{10^i, i = -6 \ldots 3\}$. Risk quantile level α ranges from .1% to 10%. Left: UCB regret increases logarithmically with the number of iterations for well-tuned C; MIN identifies the best arm after 50 iterations and its regret is constant thereafter. Right: zoom on the lower region of Left, with MIN and MARAB regrets; MARAB regret is close to that of MIN, irrespective of the C and α values in the considered ranges.

Indeed by construction (Prop. 6.2.2 and discussion), MIN is better suited to this artificial problem than UCB. Note that since the ν_is are uniform, constant A is set to 1. In this most favorable case, the lessons learned are the following:

- UCB yields a logarithmic regret for a well tuned C. Its disappointing performance comparatively to MIN and MARAB is blamed on the high variance of the worse arms, slowing down their estimation (Fig. 6.3, left);
- MIN catches the best arm after 50 iterations, and its regret stops increasing at this point due to the lack of exploration;
- MARAB interestingly yields the same behavior as MIN for a wide range of risk values α (in $[.01, 1]$), with almost no sensitivity with respect to the exploration coefficient C (Fig. 6.3, right). Complementary experiments show that the MARAB sensitivity w.r.t. C increases for higher values of α ($\alpha > .2$).

6.5.3 Artificial problems

A relaxed setting is considered, where the generator only satisfies the assumption of Equation 6.2. The i-th arm distribution ν_i is a mixture of n_i truncated Gaussians:

- n_i is uniformly drawn in $1 \ldots 4$;
- for $j = 1 \ldots n_i$ the j-th Gaussian $\mathcal{N}(\mu_{i,j}, \sigma_{i,j})$, is defined by uniformly sampling $\mu_{i,j}$ in $[0,1]$ and $\sigma_{i,j}$ in $[.12,.5]$;
- probabilities $p_{i,j}$, with $1 \leq j \leq n_i$ are drawn such that $\sum_j p_{i,j} = 1$;
- the minimum a_i of the i-th arm is uniformly drawn in $[0,.05]$.

Upon selecting the i-th arm, the reward is drawn by: i) selecting the j-th Gaussian with probability $p_{i,j}$; ii) drawing a reward r from $\mathcal{N}(\mu_{i,j}, \sigma_{i,j})$; iii) going to i) if $r < a_i$ or $r > 1$ (rejection-based truncation).
1,000 independent problem instances are generated. On these problems, `UCB`, `MARAB`, `MVLCB` and `ExpExp` are launched, recording their empirical cumulative regrets for time horizon $T = 2,000$ and $T = 4,000$, on the one hand, and the distribution of the gathered rewards, on the other hand. These results are inspected to examine i) how the cumulative regret is deteriorated by risk-awareness; and ii) whether the risk-aware algorithms manage to avoid triggering risky arms.

Cumulative regrets: the cost of risk-awareness

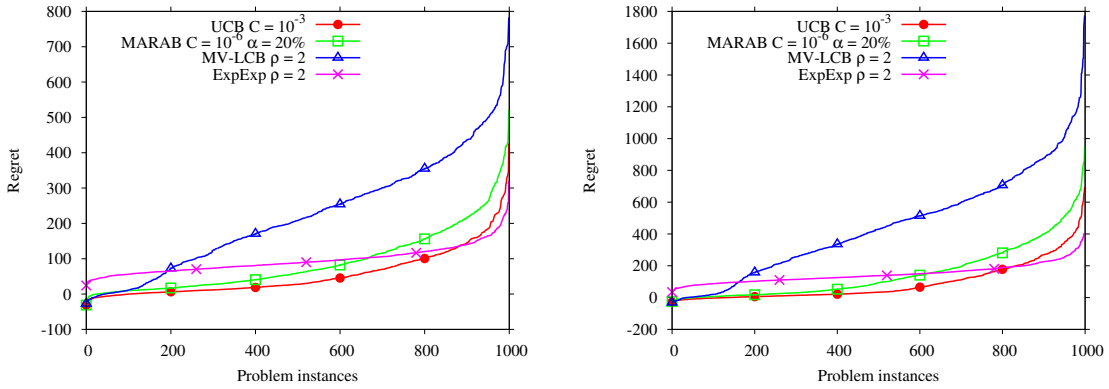
Fig. 6.4 reports the cdf of the empirical cumulative regrets of `UCB`, `MARAB`, `MVLCB` and `ExpExp` for time horizon $T = 2,000$ (Fig. 6.4, left) and $T = 4,000$ (Fig. 6.4, right), plotting for each $x = 1 \ldots 1000$ the value y such that x of the problem instances have cumulative regret less than y. All algorithm hyper-parameters are set to their optimal value after preliminary experiments. These experiments show that:

- As could have been expected, `UCB` yields the best cumulative regret overall whenever C is well tuned.
- `MARAB` suffers an extra regret compared to `UCB`; this extra regret is bounded in the considered experimental setting, and it seemingly does not increase as the time horizon increases. As could have been expected this extra regret decreases as α increases and the selection rule involves a better estimation of the empirical means. Interestingly, `MARAB` shows a very low sensitivity w.r.t. C.
- `MVLCB` yields the worst regret of all strategies, with a very low sensitivity w.r.t. parameter ρ on the considered problems.

- ExpExp yields very good results; the fact that it does never get very low cumulative regret is explained from its initial exploratory phase; a caveat is that its optimal setting used in the experiments requires the time horizon to be known in advance.

Comparatively,

- ExpExp significantly improves on MVLCB with probability circa 90%; it even improves on UCB with probability 10% (circa 20% for medium time horizon).
- MARAB improves on ExpExp with probability 70%, albeit with maximal cumulative regrets (over the problem instances) higher than for ExpExp.
- Overall, MARAB with risk level $\alpha = 20\%$ and untuned C value yields results slightly less than UCB with tuned C, for both short and medium time horizons. The risk-aware MARAB suffers a low regret increase compared to risk-neutral UCB, with a very low sensitivity w.r.t. C.
- Interestingly, a twice longer time horizon does not modify the performance order of the algorithms.



(a) Time horizon 2,000

(b) Time horizon 4,000

Figure 6.4: Distribution of empirical cumulative regret of UCB, MARAB, MVLCB and ExpExp on 1,000 problem instances (independently sorted for each algorithm) for time horizons $T = 2,000$ and $T = 4,000$. All algorithm parameters are optimally tuned ($C = 10^{-3}$ for UCB, $C = 10^{-3}$ for MARAB, $\alpha = 20\%, \rho = 2, \delta = \frac{1}{T^2}, \tau = K(\frac{T}{14})^{2/3}$).

Actual Risk Avoidance

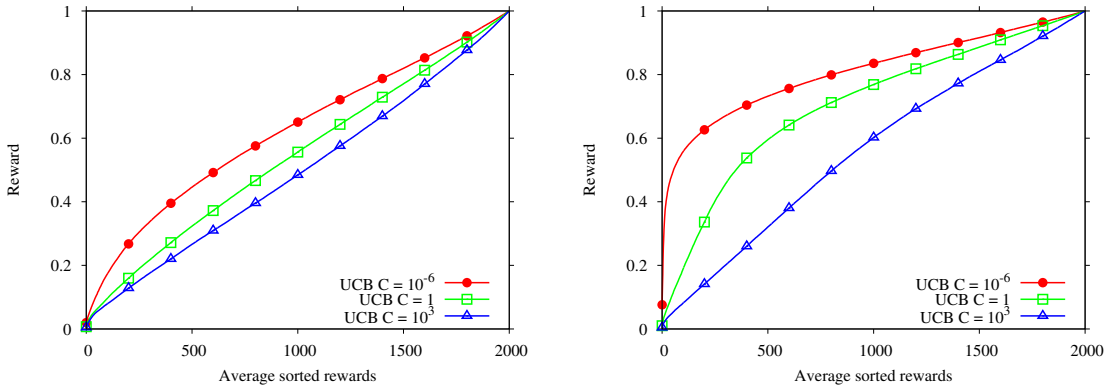
The effective risk avoidance of UCB, MVLCB, ExpExp and MARAB is investigated by inspecting the distribution of the gathered rewards. For clarity, Fig. 6.5 reports the cdf

of the rewards on two problem instances, with respectively lowest (left) and highest (right) variance of the best arm, out of the 1,000 artificial problems. For each problem instance, 40 runs are launched with time horizon $T = 2,000$. For each run, instantaneous rewards are sorted in increasing order defining a cdf. Fig 6.5 plots the average cdf : $\bar{r}_i$ denotes the average over the 40 runs of the i-th ranked reward and for each $x = 1 \ldots T$ the value $y = \overline{r}_x$ is plotted.
An efficient risk-avoiding behavior is seen as the reward cdf rises abruptly on the left, indicating that the corresponding algorithm hardly tried poor arms. Fig. 6.5 confirms previous results:

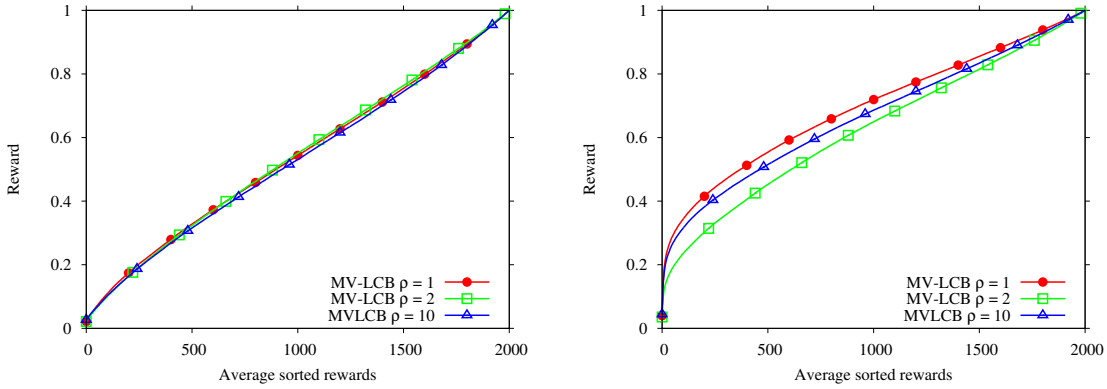
- `UCB` shows a high sensitivity w.r.t. parameter C, all the more so as the variance of the best arm is high (Fig. 6.5, top row).
- The bad performance of `MVLCB` is confirmed; its sensitivity w.r.t. ρ increases with the variance of the best arm (Fig. 6.5, second row), with a best performance for medium values of ρ.
- `ExpExp` features an excellent risk avoidance as the risky trials only take place during the exploratory phase (Fig. 6.5, third row).
- The general robustness of `MARAB` w.r.t. C is confirmed; moreover, its robustness w.r.t. the risk level α on high variance problems is empirically shown (Fig. 6.5, bottom row). It is seen that for low to medium risk ($\alpha < 20\%$), the empirical distribution of the rewards rises faster for `MARAB` than for `ExpExp`, which is explained again from the systematic exploratory phase in `ExpExp`.

6.5.4 Optimal energy management

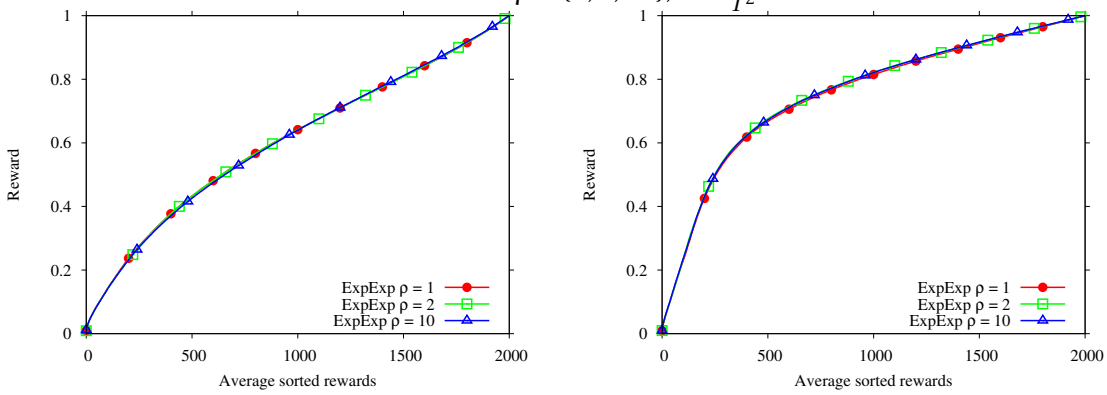
The real-world problem motivating the presented approach is a battery management problem, where the environment is described by the energy demand and the energy cost in each time step. The decision to be taken in each time step is a real-value x, determining how much energy is either used from the battery (if $x > 0$) or stored in the battery (if $x < 0$). In each time step, one must meet the demand by buying $\min(0, \text{demand} - x)$ energy; the instant reward is the opposite of the cost of the bought energy if the demand exceeds the available energy. Additionally, the battery loses some energy in each time step. A simplified setting is considered, where i) the energy cost is constant, the random process only dictates the energy demand in each time step; ii) 20 arms, corresponding to pre-defined strategies are considered. The strategy reward is drawn by uniform sampling with replacement from the 117 available realizations of the strategy.
Same general trends as for the artificial problems are observed on this real-world problem (Fig. 6.6): i) The cumulative regret is minimal for `UCB` with optimally tuned



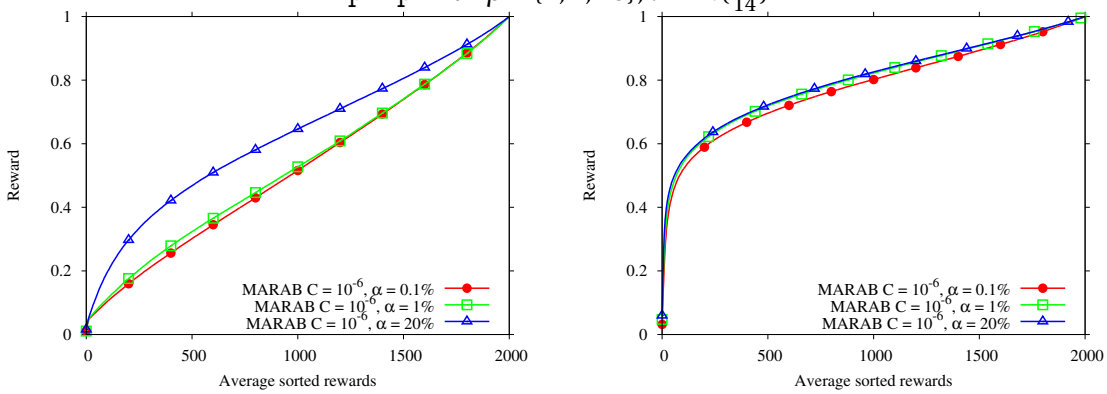
UCB with $C = 10^i, i = -6, 0, 3$



MVLCB with $\rho \in \{1, 2, 10\}, \delta = \frac{1}{T^2}$



ExpExp with $\rho \in \{1, 2, 10\}, \tau = K(\frac{T}{14})^{2/3}$



MARAB with $C = 10^{-6}, \alpha \in \{0.1\%, 1\%, 20\%\}$

Figure 6.5: Comparative risk avoidance for time horizon $T = 2,000$ for two artificial problems with low (left column) and high (right column) variance of the optimal arm. Top: UCB; second row: MVLCB; third row: ExpExp; bottom: MARAB.

C; ii) MVLCB is dominated by all other algorithms w.r.t. both risk avoidance and cumulative regret; iii) the ExpExp regret increases linearly during the exploration phase and then reaches a plateau; iv) MARAB shows its good risk-avoidance ability regardless of the C value, and MIN yields same results. Overall, MARAB offers a slightly better reward cdf in the region of low rewards, at the expense of a slight regret increase compared to UCB at its best.

6.6 MARABOUT: The Multi-Armed Risk-Aware Bandit OUThandled Algorithm

This section presents a modified version of MARAB, with a provably controlled regret. This result relies on concentration inequalities for a new mCVaR estimator, presented in Section 6.6.1.

6.6.1 Concentration inequalities

As stated in chapter 4, a key ingredient of UCB regret proof is the Hoeffding inequality, allowing the control of the probability of approximation error of the mean. In a similar fashion, the design and proof of the enhanced version of MARAB derived in this section is based on deviations inequalities firstly proposed by (Brown, 2007) and further improved in (Wang and Gao, 2010).

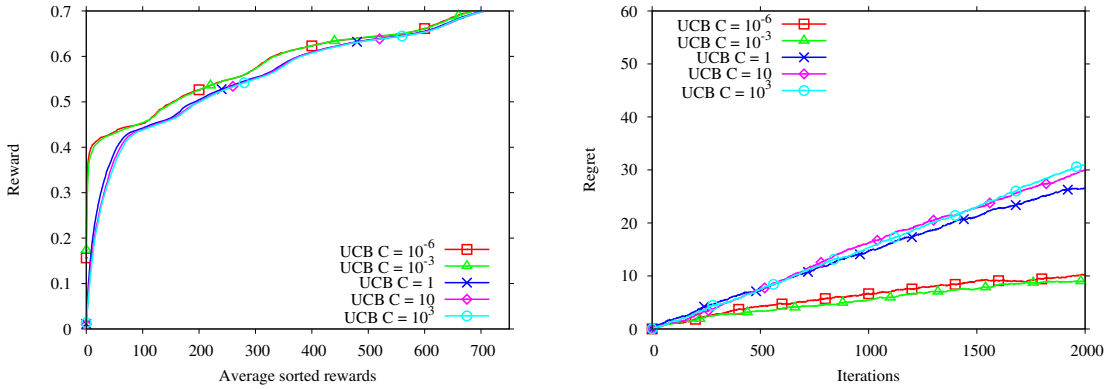
Definition 6.6.1. *Let X be a real random variable with distribution ν and let $\alpha \in (0,1]$ denote a risk level.*
Let $x_1,\dots,x_n$ be an i.i.d n-sample drawn according to ν. Assuming without loss of generality that $x_1 \leqslant x_2 \leqslant \dots \leqslant x_n$, the estimator derived by the method of moments of $mCVaR(X)$ is defined as:

$$\widetilde{mCVaR}_\alpha(x_1,\dots,x_n) \overset{\text{def}}{=} x_{(\lceil n\alpha \rceil)} + \frac{1}{n\alpha}\sum_{i=1}^{\lfloor n\alpha \rfloor}\left(x_{(i)} - x_{(\lceil n\alpha \rceil)}\right) \tag{6.21}$$

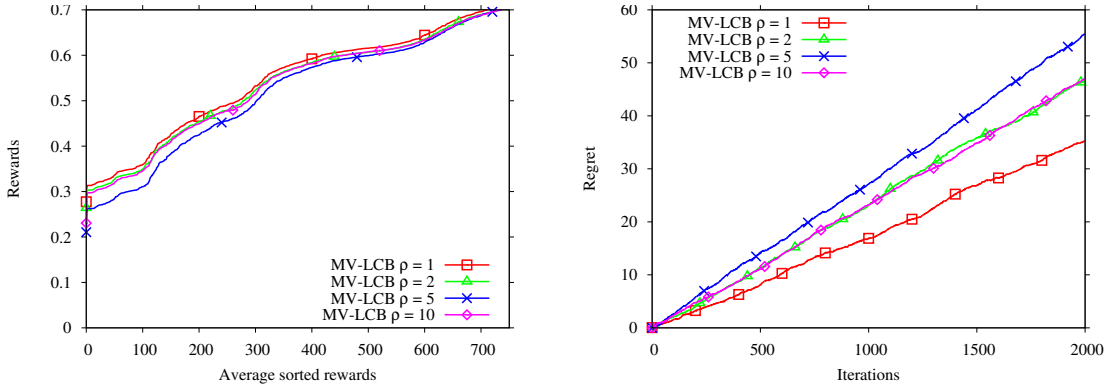
Proposition 6.6.1 (Wang and Gao (2010))**.** *Let X be a random variable with $Supp(X) \subset [a,b]$. For any $\varepsilon > 0$*

$$\mathbb{P}\left(\widetilde{mCVaR}_\alpha(x_1,\dots,x_n) \leqslant mCVaR_\alpha(X) - \varepsilon\right) \leqslant 3\exp\left(-\frac{1}{11}\alpha\left(\frac{\varepsilon}{b-a}\right)^2 n\right) \tag{6.22}$$

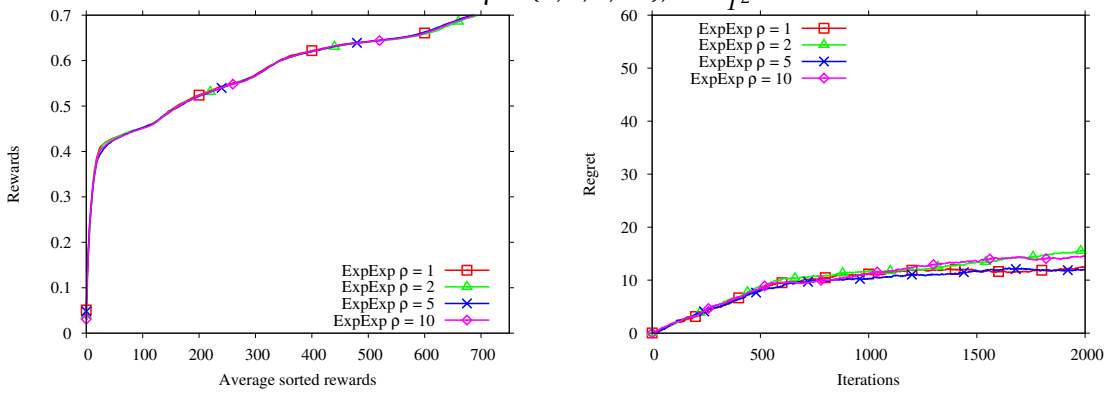
$$\begin{aligned}\mathbb{P}\left(\widetilde{mCVaR}_\alpha(x_1,\dots,x_n) \geqslant mCVaR_\alpha(X) + \varepsilon\right) &\leqslant 3\exp\left(-\frac{1}{5}\alpha\left(\frac{\varepsilon}{b-a}\right)^2 n\right) \\ &\leqslant 3\exp\left(-\frac{1}{11}\alpha\left(\frac{\varepsilon}{b-a}\right)^2 n\right)\end{aligned} \tag{6.23}$$



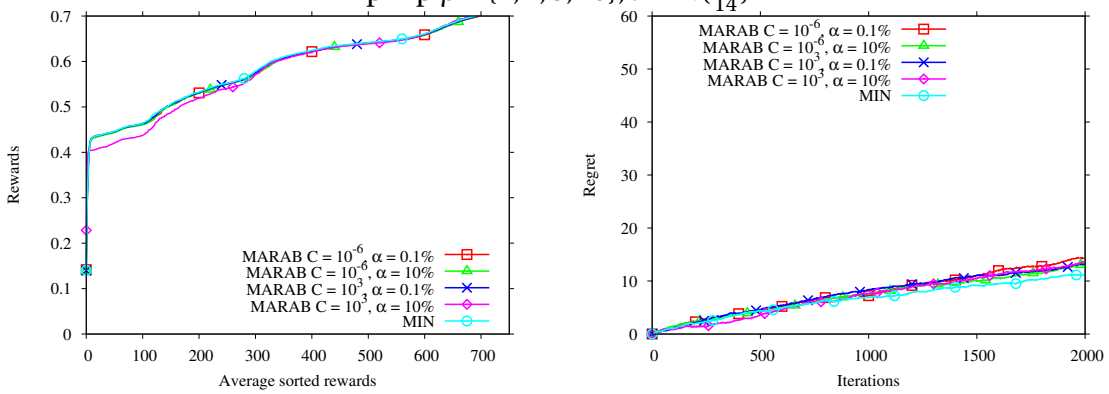
UCB. $C \in \{10^{-6}, 10^{-3}, 1, 10, 10^3\}$



MV-LCB $\rho \in \{1, 2, 5, 10\}, \delta = \frac{1}{T^2}$



ExpExp $\rho \in \{1, 2, 5, 10\}, \tau = K(\frac{T}{14})^{2/3}$



MARAB $C \in \{10^{-6}, 10^3\}, \alpha \in \{0.1\%, 10\%\}$

Figure 6.6: Comparative performance of UCB, MVLCB, ExpExp and MARAB on a real-world energy management problem. Left: sorted instant rewards (truncated to the 37.5% worst cases for readability). Right: empirical cumulative regret with time horizon $T = 100K$, averaged out of 40 runs.

6.6.2 The MARABOUT Algorithm

Based on the above proposition, a refined version of the MARAB algorithm is defined, called *Multi-Armed Risk-Aware Bandit OUThandled* (MARABOUT. Algorithm 22).

Algorithm 22 K-armed MARABOUT

Require: Time horizon T, risk level α, exploration coefficients $C > 2$ and $\beta \in [0,1]$.

1: **for** $t = 1 \dots K$ **do**
2: $I_t = t$; gather Y_t; initialize $\widetilde{mCVaR}_{\alpha,i}(X_{i,1}) = Y_t$
3: **end for**
4: **for** $t = K+1 \dots T$ **do**
5: Choose (arbitrary tie break)

$$I_t \in \underset{i\in\{1,\dots,K\}}{\operatorname{argmax}} \left\{ \widetilde{mCVaR}_{\alpha,i}(X_{i,1},\dots,X_{i,N_{i,t}}) + \sqrt{\frac{11(C\log t + \beta\log 3)}{\alpha N_{i,t}}} \right\}. \tag{6.24}$$

6: **end for**

MARABOUT assesses the quality of each arm thanks to the estimator defined in Equation (6.21) (Line 2). For each arm, a upper confidence bound is derived and the algorithm selects an arm I_t maximizing this bound (Equation (6.24), Line 5)
A key difference between MARAB and MARABOUT is that the latter does involve some exploration, the strength of which is controlled by parameters C and β (the larger these parameters, the more exploration the algorithm achieves). In particular, the optimistic assessment of the mCVaR (Equation (6.24)) leads to visit infinitely often every arm as t goes to ∞.
The merit of this new algorithm is to allow for bounding its mCVaR-related pseudo-regret in expectation. Let us first define this pseudo-regret.

Definition 6.6.2. *Let a K-armed MAB problem with reward distributions ν_i, with $\alpha \in (0,1]$ a risk level.*
Let for brevity $mCVaR_i$ denote the $mCVaR_\alpha(X_i)$ with $X_i \sim \nu_i$, with $mCVaR^\star$ the optimal $mCVaR_i$ for i ranging in $1 \dots K$. Let $\Delta_{mCVaR,i}$ denote the associated margin ($\Delta_{mCVaR,i} \overset{\text{def}}{=} mCVaR^\star - mCVaR_i$).
The mCVaR-related pseudo-regret of a MAB algorithm at time t, with same notations as in section 3.1.1 is defined by:

$$\overline{R_{mCVaR,t}} \overset{\text{def}}{=} t \times mCVaR^\star - \sum_{s=1}^{t} mCVaR_{I_t} \tag{6.25}$$

$$= \sum_{i=1}^{K} \Delta_{mCVaR,i} N_{i,t} \tag{6.26}$$

Proposition 6.6.2. *With same notations as above, assume that the support of distributions ν_i is in $[0,1]$ for i ranging in $1\dots K$.*
Then for $C>2$ and $\beta\in[0,1]$, the expected pseudo-regret of `MARABOUT` *is upper-bounded as:*

$$\mathbb{E}\left[\overline{R_{mCVaR,t}}\right] \leqslant \sum_{i:\Delta_{mCVaR,i}>0}\left\{\frac{44(C\log(t)+\beta\log(3))}{\alpha\Delta_{mCVaR,i}}+\Delta_{mCVaR,i}\left(1+\frac{2\times 3^{1-\beta}}{C-2}\right)\right\} \quad (6.27)$$

Proof. The proof is inspired from the `UCB` proof (Auer et al., 2002; Bubeck and Cesa-Bianchi, 2012).
Let $i^\star$ in $1\dots K$ be such that $mCVaR_{i^\star}=mCVaR^\star$ and $\widetilde{mCVaR_{i,N_{i,t}}}\overset{\text{def}}{=}\widetilde{mCVaR}_{i,s}\left(X_{i,1},\dots,X_{i,N_{i,t}}\right)$

Considering a time step τ where a suboptimal arm is selected ($I_\tau=i\notin\underset{k\in\{1,\dots,K\}}{\operatorname{argmax}}\, mCVaR_k$), we claim that at least one of the three following conditions below is true:

$$\widetilde{mCVaR}_{i^\star,N_{i^\star,\tau-1}}+\sqrt{\frac{11(C\log(\tau)+\beta\log(3)}{\alpha N_{i^\star,\tau-1}}}\leqslant mCVaR^\star \quad (6.28)$$

$$\widetilde{mCVaR}_{i,N_{i,\tau-1}}>mCVaR_i+\sqrt{\frac{11(C\log(\tau)+\beta\log(3))}{\alpha N_{i,\tau-1}}} \quad (6.29)$$

$$N_{i,\tau-1}<\frac{44(C\log(t)+\beta\log(3))}{\alpha\Delta^2_{mCVaR,i}} \quad (6.30)$$

By way of contradiction, assume that these three inequalities are false, it comes:

$$\begin{aligned}
\widetilde{mCVaR}_{i^\star,T(i,\tau-1)}+\sqrt{\frac{11(C\log(\tau)+\beta\log(3))}{\alpha N_{i^\star,\tau-1}}} &> mCVaR^\star \text{ (Eq. 6.28)}\\
&= mCVaR_i+\Delta_{mCVaR,i}\\
&\geqslant mCVaR_i+2\sqrt{\frac{11(C\log(\tau)+\beta\log(3))}{\alpha N_{i,\tau-1}}} \text{ (Eq. 6.30)}\\
&\geqslant \widetilde{mCVaR}_{i,N_{i,t-1}}+\sqrt{\frac{11(C\log(\tau)+\beta\log(3))}{\alpha N_{i,\tau-1}}}\\
&\quad \text{(Eq. 6.29)}
\end{aligned}$$

which in turn implies $I_t\neq i$, a contradiction. The claim is therefore proven.

One can write for $u = \lceil \frac{44(C\log(t)+\beta\log(3))}{\alpha\Delta^2_{CVaR,i}} \rceil$:

$$\begin{aligned}
\mathbb{E}[N_{i,t}] = \mathbb{E}\left[\sum_{\tau=1}^{t} \mathbb{I}\{I_\tau = i\}\right] &\leqslant u + \mathbb{E}\left[\sum_{\tau=u+1}^{t} \mathbb{I}\{I_\tau = i \text{ and Eq. 6.30 is false}\}\right] \\
&\leqslant u + \mathbb{E}\left[\sum_{\tau=u+1}^{t} \mathbb{I}\{\text{Eq. 6.28 or Eq. 6.29 is true}\}\right] \\
&\leqslant u + \sum_{\tau=u+1}^{t} \{\mathbb{P}(\text{Eq. 6.28 is true}) + \mathbb{P}(\text{Eq. 6.29 is true})\}
\end{aligned}$$

Probability of the events of Equations 6.28 and 6.29 can be upper-bounded:

$$\begin{aligned}
\mathbb{P}(\text{Eq. 6.28 is true}) &\leqslant \mathbb{P}(\exists s \in \{1\dots\tau\} : \widetilde{mCVaR}_{i^\star,s} + \sqrt{\frac{11(C\log(\tau)+\beta\log(3))}{\alpha s}} \leqslant mCVaR^\star) \\
&\leqslant \sum_{s=1}^{\tau} \mathbb{P}(\widetilde{mCVaR}_{i^\star,s} + \sqrt{\frac{11(C\log(\tau)+\beta\log(3))}{\alpha s}} \leqslant mCVaR^\star) \\
&\leqslant \sum_{s=1}^{\tau} 3^{1-\beta}\tau^{-C} = 3^{1-\beta}\tau^{1-C} \text{ by Prop. 6.6.1}
\end{aligned}$$

The same upper bound holds for the event of Equation 6.29.
Finally,

$$\begin{aligned}
\mathbb{E}[N_{i,t}] &\leqslant u + \sum_{\tau=u+1}^{t} \mathbb{P}(\text{Eq. 6.28 is true}) + \mathbb{P}(\text{Eq. 6.29 is true}) \\
&\leqslant \frac{44(C\log(t)+\beta\log(3))}{\alpha\Delta^2_{mCVaR,i}} + 1 + 2\times 3^{1-\beta} \sum_{\tau=u+1}^{t} \frac{1}{\tau^{C-1}} \\
&\leqslant \frac{44(C\log(t)+\beta\log(3))}{\alpha\Delta^2_{mCVaR,i}} + 2\times 3^{1-\beta} \underbrace{\left\{\sum_{\tau=u+1}^{t} \frac{1}{\tau^{C-1}} + 1\right\}}_{\leqslant \frac{C-1}{C-2} \text{ by } \sum_{k=1}^{\infty} \frac{1}{k^\gamma} < \frac{\gamma}{\gamma-1} \text{ for } \gamma>1} + 1 - 2\times 3^{1-\beta} \\
&\leqslant \frac{44(C\log(t)+\beta\log(3))}{\alpha\Delta^2_{mCVaR,i}} + 1 + \frac{2\times 3^{1-\beta}}{C-2}
\end{aligned}$$

By using the regret definition, the announced result is proved.

□

Remark 6.6.1. *As said, the exploration strength of MARABOUT increases with β and C. After the regret bound, it is seen that increasing β reduces the multiplicative factor of term $\Delta_{mCVaR,i}$; this extra-exploration thus might result in lowering the upper-bound*

on the regret. In practice however, it will be seen that $\beta = 0$ is a recommandable value. This empirical finding is in agreement with the fact that risk-aversion is better served by a conservative algorithmic behavior.

6.6.3 Experimental validation

Synthetic problems

MARABOUT is experimentally validated on three 2-armed artificial MAB problems, reflecting the fact that the MAB difficulty in a risk-aware setting depends on three factors. The most obvious one is the mCVaR margin Δ_{mCVaR}; a low mCVaR margin adversely affects the estimation task like a low mean margin adversely affects the classical MAB algorithm behaviors. A second factor is the risk level α, reflecting the probability of undesirable events one would like to be protected from. For α close to 0, the unwanted events are extremely rare, and thus their impact is very difficult to assess (e.g., only one sample out of 100 could be used for $\alpha = .01$). The estimation task is likewise made more hazardous when considering short time horizons: specifically, $\frac{T}{K}$ provides an indication of the average number of samples which can be expected to compute the desired estimates.
Accordingly, the artificial three problems illustrate different types and levels of difficulty:

- The first problem involves two arms with same mean and high mCVaR margin. The first arm is a Bernoulli variable of parameter .5 and the second arm returns a constant reward .5. The mCVaR margin Δ_{mCVaR} thus is high (.5) while the mean margin is 0.

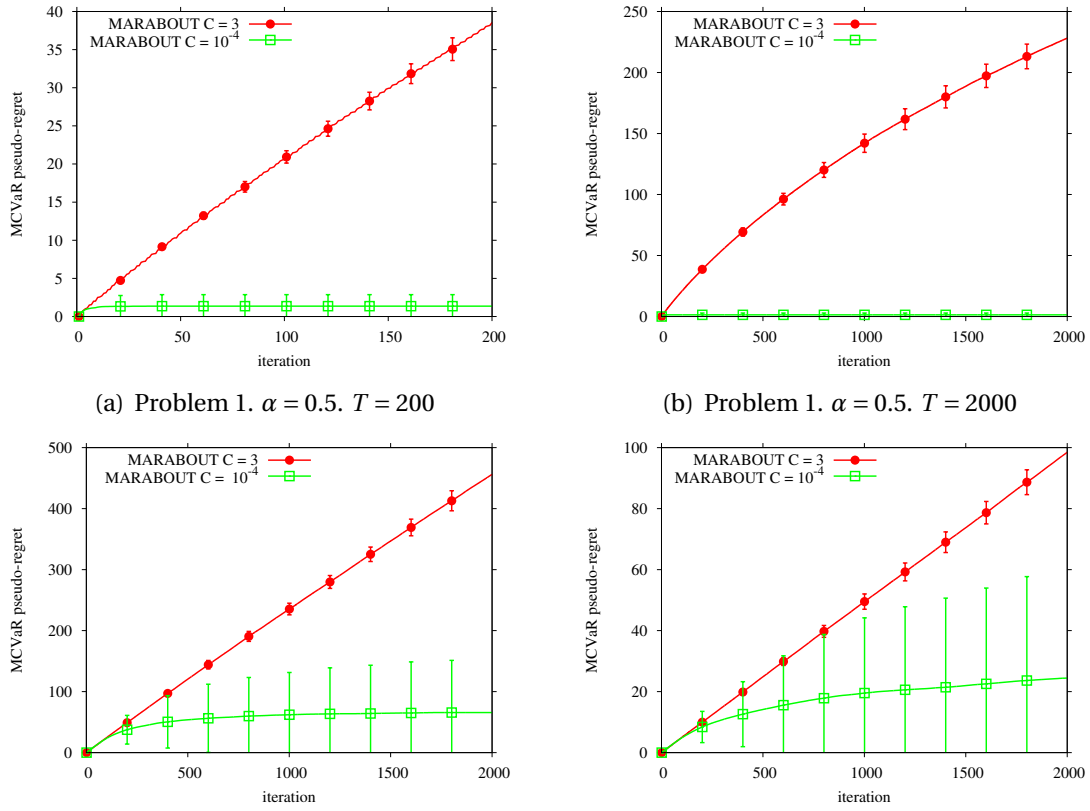
- The second problem involves two arms with low mean margin and high mCVaR margin. The first arm returns .5 with probability .01 and 1 otherwise; the second arm is a Bernoulli variable of parameter .99. The mean margin Δ thus is low (5.10^{-3}) while the mCVaR margin is high ($\Delta_{mCVaR} = .5$).

- The third problem involves two arms with very low mean margin and low mCVaR margin. The first arm returns .1 with probability .01 and 1 otherwise; the second arm is a Bernoulli variable of parameter .99. The mean margin Δ thus is very low (10^{-3}) and the mCVaR margin is low ($\Delta_{mCVaR} = .1$).

The goal of experiments is to study: i) the performance in increasingly challenging settings, in relation with the desired risk level α ; ii) the sensitivity with respect to the parameter values, and in particular, in relation with the upper bound on the regret. The first question is investigated by setting $\alpha = .5$ for the first problem, and $\alpha = .01$ for the second and third problems. The second question is investigated by setting the

exploration coefficient C to two distincts values $C = 3$ and $C = 10^{-4}$.

The CVaR regret, averaged over 100 independent runs, is displayed on Fig. 6.7.



(a) Problem 1. $\alpha = 0.5$. $T = 200$

(b) Problem 1. $\alpha = 0.5$. $T = 2000$

(c) Problem 2. $\alpha = 0.01$. $T = 2000$

(d) Problem 3. $\alpha = 0.01$. $T = 2000$

Figure 6.7: $mCVaR$ Pseudo-Regret for MARABOUT averaged out of 100 runs, on three 2-armed artificial MAB problems (see text) with $C \in \{10^{-4}, 3\}$ and $\beta = 0$. Top row: problem 1 for time horizon $T = 200$ (left) and $T = 2000$ (right). Bottom line: problem 2 (left) and problem 3 (right).

The main lesson learned from the experimental results, depicted on Fig. 6.7 is that MARABOUT can achieve a logarithmic regret on all three problems, albeit with a small value of C ($C = 10^{-4}$). This suggests that when dealing with such short time horizons, the exploration strength must be limited; and that the theoretical lower bound for $C > 2$ might be practically too large. In a more detailed way:

- The first problem corresponds to an easy setting, with a large mCVaR margin ($\Delta_{mCVaR} = 0.5$) and a high risk level $\alpha = .5$, enabling to use a fair amount of information. In this case, the $mCVaR$ pseudo-regret is logarithmic even for high values of C ($C = 3$) provided that the time horizon is sufficiently large (Fig. 6.7, top row, right). In this problem, a standard MAB algorithm would

equally select both arms since the mean margin is 0, leading to a high variance of the rewards. In such situations, MARABOUT contributes to the stability of the gathered rewards.

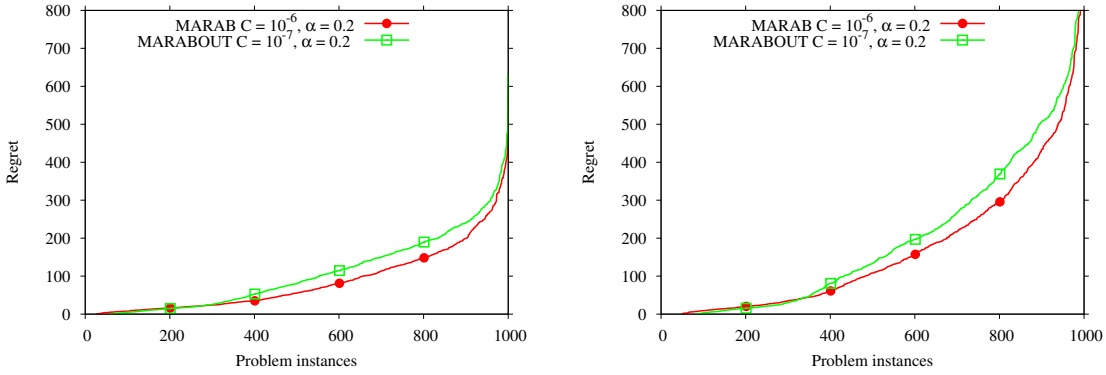
- For the second problem (Fig. 6.7(c)), the main challenge lies in the tiny value of $\alpha = 0.01$. Achieving a logarithmic regret rate in this challenging setting and over such short time horizons is a encouraging result for the applicability of the approach.
  Actually, such a setting is ideally suited to the use of risk-aware approaches. On the one hand, the low mean margin ($\Delta = 5.10^{-3}$) adversely affects classical MAB algorithms. On the other hand, since Δ_{mCVaR} is large, the loss encountered in worst cases is large, which is specifically what one wants to be protected from. In the meanwhile, the large mCVaR margin makes it easier for MARABOUT.
  Furthermore, the above does not depend on the value of α. The fact that logarithmic regrets are attained for low α values thus also suggests a wide range of successful application of the approach.
  Note that the standard deviation is increased compared to Pb 1; a tentative interpretation for this large standard deviation is that the estimate of $mCVaR_{\alpha,i}$ is very sensitive to the first samples; many trials might be required to revise the (too optimistic) estimate.

- On the third problem, the most challenging one out of the three, MARABOUT also successfully manages to reach a logarithmic regret. The same remarks as above hold, where the higher variance of the results is explained by the smaller Δ_{mCVaR}.
  It must however be said that a practitionner might want to use a standard MAB algorithm in such a setting, for the losses in the worst cases remain small ($\Delta_{mCVaR} = 10^{-3}$) and optimistic MAB algorithms might make a more efficient use of all the samples to compensate for the smaller margin Δ.

Comparison with MARAB

The purpose of this section is now to compare the behavior of MARAB and MARABOUT. Figure 6.8 presents the regret after $T = 2000 = 100K$ (left) and $T = 4000 = 200K$ (right) iterations. For a parameter C one order of magnitude lower than for MARAB, MARABOUT is able to reach comparable level of performances on all the collection of 1000 artificial problems. In the same manner, Figure 6.9 compares the average sorted rewards of MARAB vs MARABOUT for the problem with the smallest (left) and largest (right) variance on the optimal arm. Like previously, a larger sensitivity to the parameter tuning is seen in the low variance case. However, MARABOUT is able to reach in both cases the

performance of MARAB for $\alpha = 20\%$. Finally, Figure 6.10 is interested with the real-world energy problem and shows the ability of the approach to gather good rewards in the 37.5% worst cases when α is low (left). Fig. 6.10 also shows that, given a C value, the MARABOUT performance is not sensitive to α. Secondly, MARABOUT performance is optimal for very low values of C ($C = 10^{-7}$). In the meanwhile, MARABOUT is dominated by MARAB (with low sensitivity with respect to both α and C, Fig 6.6). This fact is explained by the pessimistic and conservative strategy of MARAB; the lack of exploration does not harm its performance in small time horizon.



(a) Time horizon 2,000

(b) Time horizon 4,000

Figure 6.8: Comparative distribution of empirical cumulative regret of MARAB and MARABOUT on 1,000 problem instances (independently sorted for each algorithm) for time horizons $T = 2,000$ and $T = 4,000$.

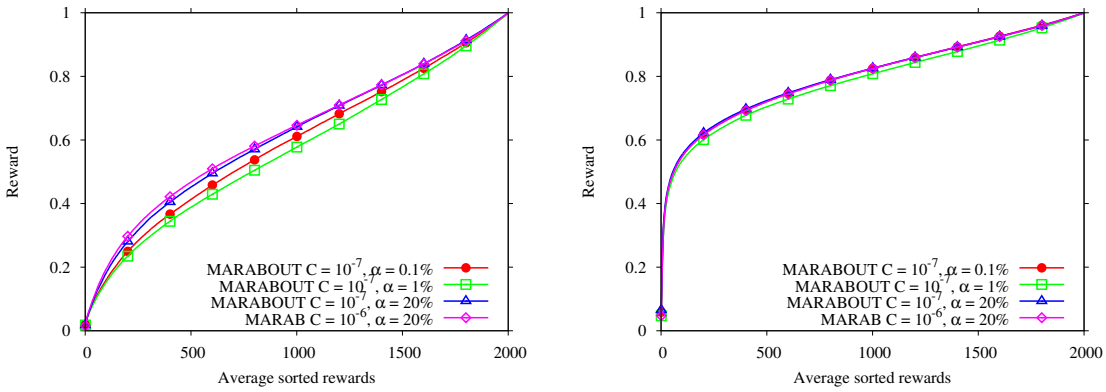


Figure 6.9: Comparative risk avoidance of MARAB and MARABOUT for time horizon $T = 2,000$ for two artificial problems with low (left column) and high (right column) variance of the optimal arm.

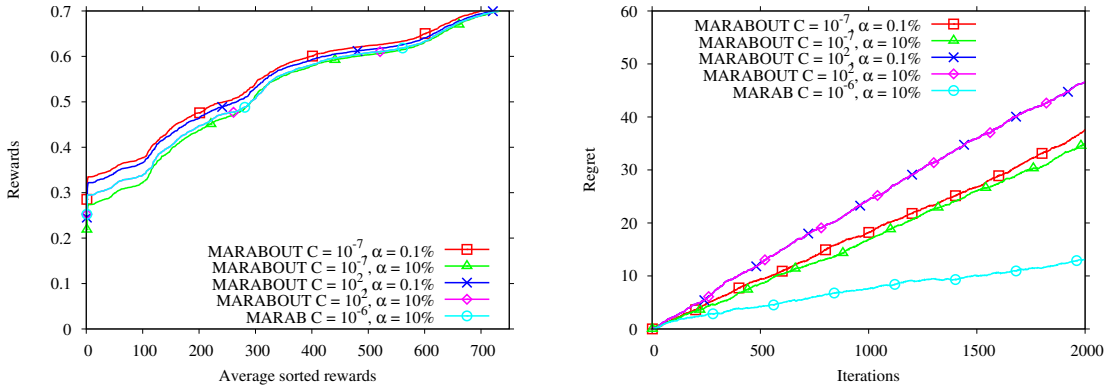


Figure 6.10: Comparative performance of MARAB and MARABOUT on a real-world energy management problem. Left: sorted instant rewards (truncated to the 37.5% worst cases for readability). Right: empirical cumulative regret with time horizon $T = 100K$, averaged out of 40 runs.

6.7 Discussion and perspectives

The first aim of our work is to extend the now standard *Exploration vs Exploitation* trade-off to a more complex and very applicatively relevant trade-off: *Exploration vs Exploitation vs Safety*.
The presented contributions are structured as follows:

- Under quite restrictive assumptions (about the arm distributions) and goals (where the risk level α goes to 0), it has been shown that MIN yields surprisingly good results, successfully competing with UCB.

- The MARAB algorithm was introduced to handle the general case of a risk level $\alpha > 0$; though with no provable guarantees, MARAB was found to outperform MVLCB and ExpExp in the general case on artificial[1] and real-world problems. MARAB is dominated by UCB at its best, i.e. for an optimally tuned C parameter; on the other hand, the sensitivity of UCB w.r.t. parameter C makes it ill-suited to risk-sensitive contexts. Quite the contrary, MARAB displays a very low sensitivity to parameter C, all the more appreciated as the burden of hyper-parameter tuning (Aut, 2015) is increasingly acknowledged in the Machine Learning field.

- An improved version of MARAB, MARABOUT enjoys a provable guarantee of logarithmic regret in term of the CVaR pseudo-regret, and experiments on artificial problems illustrating the different difficulties of risk-aware MAB confirm that

[1] Though with deteriorated results in the worst 30% cases comparatively to ExpExp; but it is true to say that ExpExp setting was optimally tuned, assuming that the time horizon is known.

the logarithmic regret is achieve in all cases (although with a much lower C value that the one considered in the analysis).

The research perspectives of this work are twofold. A medium term perspective, the MARAB and MARABOUT algorithms will be extended to the context of tree-structured search space to achieve safe sequential decision making, along the same lines as (Moldovan and Abbeel, 2012). A longer term perspective is to extend the MAB approaches to the general case where the quantity to be maximized is a function of the samples, following the pioneering work of (Neufeld et al., 2014).

Chapter 7

Subsampling for contextual linear bandits

Contents

This chapter presents our second contributions, focused on the *contextual linear bandit problem*. The contextual MAB setting extends the classical MAB setting by supposing that, at each time step, additional side information is provided to the learner and may be used to improve the arm choice. This situation holds true in various real world problems where partial but relevant data is available and exploitable.
After introducing the motivations for contextual MABs, this chapter presents the `BESA` algorithm (Baransi et al., 2014) and details our extension of the sub-sampling strategy

of BESA to the contextual linear case, forming the Contextual Linear Best sub-Sampled Arm (CL-BESA) algorithm.

7.1 Introduction

Like classical MABs, the contextual multi-armed bandit setting considers a finite set of arms with unknown associated reward distributions. The difference is that the learner is additionally provided with a **context** (or **state**) in each time step, knowing that this reward associated to the selected arm depends on the context.
The contextual MAB setting, introduced in chapter 3, is relevant in all application domains where some side information can be exploited about the current decision. For instance, in the medical domain where arms are classically associated with various treatments, data relative to the patient's medical background may be of great interest. Likewise, contextual bandits have also been used to customize online content delivery (Li et al., 2010).
The performances of a bandit algorithm are mainly assessed according to their cumulative regrets (chapter 3). Indeed, contextual MAB problems could likewise be handled through standard MAB algorithms, offering the practitionner the benefit of the regret guarantees associated with the state-of-the-art methods. Falling back on standard MAB algorithms is however likely to be suboptimal as the information provided by the context is dismissed. Different approaches have therefore been used for contextual and context-free MAB settings.
As already said, the (context-less and contextual) non-greedy MAB algorithms belong to two distinct categories:

1. The **optimistic** algorithms include: UCB (Auer et al., 2002), UCB-V (Audibert et al., 2009), KL-UCB (Cappé et al., 2013; Maillard et al., 2011), DMED (Honda and Takemura, 2010) and in the contextual case OFUL (Abbasi-Yadkori et al., 2011). These algorithms compute for each possible arm a confidence region containing an arm parameter (e.g. its value expectation) with high probability. Then, the chosen arm is the one considered to be the most promising in the sense that it maximizes the expected value augmented with an estimate of the confidence on this expectation. In doing so, the algorithm applies the *optimism in front of the unknown* principle.

2. The **Bayesian** algorithms essentially include the Thompson sampling in the stochastic MAB case (Thompson, 1933; Agrawal and Goyal, 2012b) and the linear contextual MAB (Agrawal and Goyal, 2012a) in the contextual case. This approach relies on maintaining and updating, for each arm and at each time step, a prior distribution over the best parameter (in term of mean reward

maximization). This prior is used at each time step to sample a parameter value and pick the corresponding best arm.

The work presented in this chapter proposes a third approach to the contextual bandits, rooted in the BESA algorithm first presented by (Baransi et al., 2014). The BESA algorithm proceeds by comparing any two arms using a subsampling strategy. Formally, let us consider two arms associated with respectively n_1 and n_2 samples, respectively denoted $\{X_{1,s}, s = 1 \dots n_1\}$ and $\{X_{2,s}, s = 1 \dots n_2\}$. Let us further assume with no loss of generality that $n_1 \geq n_2$. The intuition is that the choice should be based on comparing the two arms by using the same amount of information. Accordingly, one extracts a n_2-subsample from the 1st sample $\{X_{1,s}, s = 1 \dots n_1\}$, using uniform sampling without replacement, and compute its empirical average. Depending on whether this empirical average is higher or lower than the empirical average of the 2nd sample $\{X_{2,s}, s = 1 \dots n_2\}$, one selects the first or the second arm.
In the stochastic MAB case (Baransi et al., 2014), this sub-sampling strategy showed state-of-the-art performances, with several practical benefits over the other approaches. Compared to the UCB and Bayesian algorithms, it does not require any complex computation (e.g. empirical confidence interval, posterior update or sampling from a possibly complex distribution) and can be applied, *as is*, to a wide variety of arm distributions, without requiring any prior knowledge contrary to KL-UCB or Thompson sampling. Moreover, its great simplicity leads to a straigthforward implementation; furthermore, it does not require any hyper-parameter to be tuned, which is a huge advantage for practical applications, as already said. These properties motivated the extension of the sub-sampling to the contextual case.

7.2 Sub-sampling Strategy for Contextual Linear Bandit

This section introduces the Contextual Linear Best sub-Sampled Arm (CL-BESA) algorithm, in the $K = 2$ arms case. The extension to an arbitrary finite number of arms is straightforward by considering a tournament among the arms, as in (Baransi et al., 2014).

7.2.1 Notations

Let us use the same notations as in section 3.3, with the context space $\mathcal{X}$ and parameter space Θ included in $\mathbb{R}^d$. Some additional notations go as follows:
Let $\mathcal{S} = \big((X_{i_1}, Y_{i_1}), \dots, (X_{i_S}, Y_{i_S})\big)$ denote a context-reward sample set of size S. $\mathbf{X}(\mathcal{S}) = (X_{i_1}, \dots, X_{i_S})^T$ is the $S \times d$ context matrix and $\mathbf{Y}(\mathcal{S}) = (Y_{i_1}, \dots, Y_{i_S})^T$ is the S reward vector.

One wants to estimate θ such that $\mathbf{Y}(\mathscr{S}) = \mathbf{X}(\mathscr{S})\theta$.
The corresponding regularized least-square estimate $\widehat{\theta}$ is then defined as:

$$\widehat{\theta}_\lambda(\mathscr{S}) \stackrel{\text{def}}{=} \left(\mathbf{X}(\mathscr{S})^T\mathbf{X}(\mathscr{S}) + \lambda I_d\right)^{-1}\mathbf{X}(\mathscr{S})^T\mathbf{Y}(\mathscr{S}) \tag{7.1}$$

with I_d the $d \times d$ identity matrix and $\lambda > 0$ a regularization parameter.
It can be showed that $\widehat{\theta}_\lambda$ is the solution of the regularized least-squares problem defined by:

$$\widehat{\theta}_\lambda = min_\theta \|\mathbf{Y}(\mathscr{S}) - \mathbf{X}(\mathscr{S})\theta\|_2^2 + \lambda\|\theta\|_2^2$$

In particular, the regularization allows a unique closed-form estimation in the case of a rank-deficient matrix $\mathbf{X}(\mathscr{S})$.
Moreover, the following notations are introduced:

- $\mathscr{S}_{i,t} \stackrel{\text{def}}{=} \{(X_{t'}, Y_{t'}) : t' \leqslant t, I_{t'} = i)\}$ denotes the subset of observations corresponding to time steps when the arm i has been selected up to and including time t.
- $I \sim \texttt{Wr}(n, m)$ denotes a random set of n indices drawn uniformly over the set $\{1, \dots, m\}$. By convention, $I = \{1, \dots, m\}$ if $n \geqslant m$.
- By noting $\mathscr{S} = \{s_1, \dots, s_S\}$ a finite set of observations, one defines the subsampled set according to I by $\mathscr{S}(I) \stackrel{\text{def}}{=} \{s_i, i \in I\}$.
- $N_t(i) = \sum_{s=1}^{t} \mathbb{I}\{I_t = t\}$ is the number of pulls of arm i until time t.

Technical assumptions

Some further technical assumptions are required in order to derive theoretical regret bound.
First, both $\mathscr{X}$ and Θ are supposed to be convex, bounded and are known to the learner.
The noise η_t is supposed to be a i.i.d, centered and sub-Gaussian, i.e. there exists some known $R_\eta \in \mathbb{R}$, such that, for all $\lambda \in \mathbb{R}$,

$$\log\left[\mathbb{E}\exp(\lambda\eta_t)\right] \leqslant \frac{\lambda^2 R_\eta^2}{2} \tag{7.2}$$

The context mean is noted μ and one writes $X_t = \mu + \xi_t$, where ξ_t is a centered i.i.d noise, bounded almost surely by $\|\xi_t\|_2^2 \leqslant \frac{\sigma^2}{2}$ for some constant σ_X^2, and such that for all $\lambda \in \mathbb{R}^d$:

$$\log\left[\mathbb{E}\exp(\lambda^T\xi_t)\right] \leqslant \frac{\|\lambda\|_2^2\sigma_X^2}{2} \tag{7.3}$$

Finally, for convenience, one assumes that:

1. $\forall x \in \mathscr{X}, \forall \theta \in \Theta, |\langle x, \theta \rangle| \leqslant 1$

2. The radius of the parameter space Θ is bounded by some constant B:

$$\max_{\theta \in \Theta} \|\theta\|_2 \leqslant B \tag{7.4}$$

3. All distributions have density with respect to the Lebesgue measure.

7.2.2 Contextual Linear Best Sub-Sampled Arm

The Contextual Linear Best Sub-Sampled Arm (`CL-BESA`) is introduced (Algorithm 23) in the 2-arm case, where the two arms are respectively denoted a and b. As said, the extension to any finite number of arms proceeds by considering a tournament among the arms (Algorithm 24).

Algorithm 23 `CL-BESA` (a,b) for two arms

Require: Current time t, context X_t, parameter λ.

1: Sample $I^a_{t-1} \sim \texttt{Wr}(N_{t-1}(b); N_{t-1}(a))$ and $I^b_{t-1} \sim \texttt{Wr}(N_{t-1}(a); N_{t-1}(b))$.

2: Compute the estimates $\widehat{\theta}_{a,t-1} \overset{\text{def}}{=} \widehat{\theta}_\lambda(\mathscr{S}_{a,t-1}(I^a_{t-1}))$ and $\widehat{\theta}_{b,t-1} \overset{\text{def}}{=} \widehat{\theta}_\lambda(\mathscr{S}_{b,t-1}(I^b_{t-1}))$

3: Choose (break ties by choosing the least sampled arm)

$$I_t = \underset{a' \in \{a,b\}}{\operatorname{argmax}} \langle X_t, \widehat{\theta}_{a',t-1} \rangle \,. \tag{7.5}$$

Like `BESA`, `CL-BESA` focuses on comparing a and b based on an equal quantity of information. To this aim, the algorithm uniformly sub-samples from the most-visited arm a number of observations equals to the one of the least visited arm (Line 1). The arm parameters θ_a and θ_b are estimated thanks to Equation 7.1 (Line 2) and the arm with expected highest payoff is selected (Line 3).
Let us clarify the algorithm mechanism by rolling a simple example. Let us assume that two arms a and b have been respectively visited $N_{t-1}(a) = 10$ and $N_{t-1}(b) = 3$ times at time t. Then:

- Line 1 sub-samples, uniformly and without replacement, a subset I^a_{t-1} of size 3 from the set $\{1, \ldots, 10\}$.
- Line 2 computes the estimates $\widehat{\theta}_{a,t-1}$ and $\widehat{\theta}_{b,t-1}$ according to Eq. 7.1 and the indice sets I^a_{t-1} and the whole set of indices $\{1,2,3\}$ associated to arm b.
- Line 3 selects the most promising arm according to these estimates. Contrarily to the optimistic algorithms, there is no penalization of the estimate by a confidence bound.

Discussion

Like in the stochastic case, the approach can seem counter-intuitive and sub-optimal. Indeed, when interested in estimating a parameter, the intuition would suggest to take profit of any available information whereas sub-sampling imply discarding observations. This is especially the case where observations are largely imbalanced, for instance $N_{a,t-1} \gg N_{b,t-1}$. In the case of a single decision, it is clear that discarding information will not provide any benefits.

However, as previously stated, MAB algorithms are rooted on faithfully enforcing the Exploration vs. Exploitation trade-off, and MAB algorithms implement different ways of balancing the exploration and exploitation efforts. In this perspective, sub-sampling should be interpreted as yet another approach to enforce this balance. In particular, it must be noted that:

1. Consecutive comparisons between two arms a and b compensates for the discarded information as the subsampling of the most-visited arm sample yields (almost) independent subsets of observations, supporting a tight comparison between arms as detailled in (Baransi et al., 2014).

2. The linear reward structure assumption leads to estimate parameters θ_a and θ_b in each time step through a simple regularized least squares.

Algorithm 24 `CL-BESA` ($\mathscr{A}$)

Require: Current time t, context X_t, parameter λ, arm set $\mathscr{A}$

if $\mathscr{A} = \{a\}$ **then**
  $I_t = a$
else
  $I_t = \texttt{CL-BESA}\left(\texttt{CL-BESA}\left(\{1,\dots,\lceil\frac{K}{2}\rceil\}\right), \texttt{CL-BESA}\left(\{\lfloor\frac{K}{2}\rfloor,\dots,K\}\right)\right)$
end if

As said, `CL-BESA` can be extended to an arbitrary number K of arms after (Baransi et al., 2014), through a dichotomic process (Algorithm 24). The final choice of arm is given by organizing a tournament between arms. The arm set $\mathscr{A} = \{1,\dots,K\}$ is shuffled thanks to a random permutation $\sigma_t \in S_K$, thus preventing the apparition of learning biases by comparing same arms in each time iteration.

7.3 Contextual regret bound

7.3.1 Contextual regret

At each time step t, with context X_t, the best arm is defined as:

$$\theta_{t,\star} \stackrel{\text{def}}{=} \underset{i}{\text{argmax}} \langle X_t, \theta_i \rangle$$

Contrasting with the standard MAB context, the optimal arm is not fixed and it can change at each time iteration depending on the context X_t. A degenerate case is when the perturbation of the context information, noted σ_X, is small enough and the optimal arm becomes constant.

Given a time horizon $T \in \mathbb{N}^\star$, the **contextual regret** is thus defined as:

$$R_{X,T} = \sum_{t=1}^{T} \langle X_t, \theta_{t,\star} - \theta_{I_t} \rangle \tag{7.6}$$

The definition of *contextual* regret emphasizes the influence of the context X_t on the reward. This feature, specific to the contextual bandit case, implies a non-trivial contextual regret minimization.

7.3.2 Theoretical bound

Theorem 11 provides an upper-bound for the contextual regret of `CL-BESA`. Its proof takes inspiration from the theoretical study of both `BESA` (Baransi et al., 2014) and `OFUL` (Abbasi-Yadkori et al., 2011). However, contrarily to the case of `OFUL`, the least-squares matrices defining the vector estimates may change a lot between two consecutive time steps because of the sub-sampling, preventing the straightforward adaptation of the previous proof. Another proof design is thus proposed to address this difficulty.

Theorem 11. *Let R_η and σ_X be the sub-Gaussian parameters respectively defined by 7.2 and 7.3 and let B defined by Eq. 7.4. Run `CL-BESA` with non-decreasing regularization parameter (Eq. 7.1) $\lambda \geqslant 6\sigma_X^2 \log(T)$. Assume that*

$$\left|\langle \mu, \theta_a - \theta_b \rangle\right| \geqslant 8\sigma_X B + 2\sqrt{2}\frac{\|\theta_a\|_2 + \|\theta_b\|_2}{\sqrt{\lambda^{-1} + \|\mu\|_2^{-2}}}. \tag{7.7}$$

Then, the contextual regret of `CL-BESA` *after T rounds is upper bounded by*

$$\mathbb{E}[R_{X,T}] \leqslant \big(\max_{t\in[T]}\Delta_t\big)\frac{64}{\min_{t\in[T]}\Delta_t^2}\left[R_\eta\sqrt{2d\log\Big(\lambda^{1/2}T^2+\frac{T^3(\|\mu\|_2+\sigma_X)^2}{d\lambda^{1/2}}\Big)}+\lambda^{1/2}B\right]^2$$
$$+\big(\max_{t\in[T]}\Delta_t\big)\frac{24\sigma_X^2\log(T)-2\lambda}{\|\mu\|_2^2}+\sum_{t=1}^{T}\Delta_t\mathbb{1}\{\min_{t\in[T]}\Delta_t\leqslant\tau\}\Big)+O(1)\,.$$

where the expectation is with respect to the internal randomness of the algorithm and of the additive reward noise, and where

$$\tau \overset{\text{def}}{=} 2\sigma_X\left[R_\eta\sqrt{\frac{2d}{\lambda}\log\Big(\lambda^{1/2}T^2+\frac{T^3(\|\mu\|_2+\sigma_X)^2}{d\lambda^{1/2}}\Big)}+B\right]. \tag{7.8}$$

When the context perturbation is $\sigma_X=0$, then Δ_t reduces to $\Delta \overset{\text{def}}{=} |\langle\mu,\theta_a-\theta_b\rangle|$ and we obtain

$$\mathbb{E}[R_{X,T}] \leqslant \frac{128R_\eta^2 d}{\Delta}\log(2T^3\|\mu\|_2^2/d)+O(1).$$

Algorithm	Regret bound
`CL-BESA`	$O(d\log(T^3))$
`OFUL` (Abbasi-Yadkori et al., 2011)	$O(d\log^2(T))$
`Confidence Bound` (Dani et al., 2008)	$O(d^2\log^3(T))$
`Thompson sampling` (Agrawal and Goyal, 2012a)	$O(d^2\sqrt{T})$

Table 7.1: Theoretical regret bounds for contextual bandit algorithms

Remark 7.3.1. *For the sake of comparison, Table 7.1 lists the theoretical regret bounds associated with state of the art algorithms, showing the merits of* `CL-BESA`*: the associated regret scales linearly with the dimension d, and logarithmically with the time horizon T. This result establishes the applicability of the sub-sampling technique to the contextual multi-armed bandit problem.*

Remark 7.3.2. *The restriction regards the minimum gap $\min_{t\in[T]}\Delta_t$ (see Eq. 7.8). Note that a similar limitation is encountered in the distribution-dependent analysis of* `OFUL`*. Additionally, the authors explicitly assume a constant optimal arm in the distribution-dependent proof.*

Remark 7.3.3. *The assumption $\lambda \geqslant 6\sigma_X^2\log(T)$ is mainly formulated due to technical reasons. Also, experiments (see 7.4) suggest Equations 7.7 and 7.8 might be improved further. In practice, $\lambda=\lambda_t=\Omega(\sigma_X^2)\log(t)$ is recommended, even though a good robustness w.r.t. the choice of λ is observed in the experiments.*

Proof. **Step 1:** Let $\star_t$ be the optimal action at time t and $\neg\star_t$ the other action (Recall that there are only $K = 2$ arms $\mathscr{A} = \{a, b\}$).

By definition of the contextual regret at time T,

$$\begin{aligned} R_{X,T} &= \sum_{t=1}^{T} \langle X_t, \theta_{\star_t} - \theta_{I_t} \rangle \\ &= \sum_{t=1}^{T} \langle X_t, \theta_{\star_t} - \theta_{\neg\star_t} \rangle \mathbb{I}\{I_t = \neg\star_t\} \\ &= \sum_{t=1}^{T} \Delta_t \mathbb{I}\{\star_t = a, I_t = b\} + \sum_{t=1}^{T} \Delta_t \mathbb{I}\{\star_t = b, I_t = a\}, \end{aligned}$$

with $\Delta_t = \langle X_t, \theta_{\star_t} - \theta_{\neg\star_t} \rangle$ the instantaneous gap.

The event $\{I_t = \neg\star_t\}$ involves $\langle X_t, \widehat{\theta}_{\star_t,t-1} - \widehat{\theta}_{\neg\star_t,t-1} \rangle$, thus the instantaneous gap can be decomposed as

$$\begin{aligned} \Delta_t = &\langle X_t, \theta_{\star_t} - \widehat{\theta}_{\star_t,t-1} \rangle + \langle X_t, \widehat{\theta}_{\neg\star_t,t-1} - \theta_{\neg\star_t} \rangle \\ &+ \langle X_t, \widehat{\theta}_{\star_t,t-1} - \widehat{\theta}_{\neg\star_t,t-1} \rangle . \end{aligned} \tag{7.9}$$

Now, on the event $\{I_t = \neg\star_t\}$, either $\langle X_t, \widehat{\theta}_{\star_t,t-1} - \widehat{\theta}_{\neg\star_t,t-1} \rangle < 0$, or $\langle X_t, \widehat{\theta}_{\star_t,t-1} - \widehat{\theta}_{\neg\star_t,t-1} \rangle = 0$ and $N_{t-1}(\neg\star_t) < N_{t-1}(\star_t)$, or $\langle X_t, \widehat{\theta}_{\star_t,t-1} - \widehat{\theta}_{\neg\star_t,t-1} \rangle = 0$ and $N_{t-1}(\neg\star_t) = N_{t-1}(\star_t)$ and a random coin $\xi_t \sim B(0.5)$ is tossed that gets value 1 (without loss of generality). In any case, it holds that

$$\langle X_t, \widehat{\theta}_{\star_t,t-1} - \widehat{\theta}_{\neg\star_t,t-1} \rangle \mathbb{I}\{I_t = \neg\star_t\} \leqslant 0 .$$

The parameter $\widehat{\theta}_{\star_t,t-1} = \widehat{\theta}_\lambda(\mathscr{S}_{\star_t,t-1}(I_{t-1}^{\star_t}))$ involves the samples $\mathscr{S}_{\star_t,t-1}$ and the sub-sampling index set $I_{t-1}^{\star_t}$. For all deterministic x and constant $\delta \in (0,1)$, and for $\mathscr{S} = \mathscr{S}_{\star_t,t-1}$, it holds by the proof of Theorem 2 from Abbasi-Yadkori et al. (2011) that with probability higher than $1-\delta$ (w.r.t. $\mathscr{S}$),

$$\begin{aligned} &\left| \langle x, \widehat{\theta}_\lambda(\mathscr{S}) - \theta_{\star_t} \rangle \right| \leqslant \|x\|_{V_\lambda(\mathscr{S})^{-1}} B_{\lambda,\star_t}(\mathscr{S}) \text{ where} \\ &\qquad B_{\lambda,\star_t}(\mathscr{S}) = R_\eta \sqrt{2\log\Big(\frac{\det(V_\lambda(\mathscr{S}))}{\lambda^{d/2}\delta}\Big)} + \lambda \|\theta_{\star_t}\|_{V_\lambda(\mathscr{S})^{-1}} , \end{aligned}$$

where R_η comes from the sub-Gaussian assumption on the noise (7.2), and $V_\lambda(\mathscr{S}) = \mathbf{X}(\mathscr{S})^\top \mathbf{X}(\mathscr{S}) + \lambda I_d$. Since $I_{t-1}^{\star_t}$ is chosen independently on $\mathscr{S}_{\star_t,t-1}$, it is not difficult to see that the same bound holds for $\mathscr{S}_{\star_t,t-1}(I_{t-1}^{\star_t})$, with respect to all sources of randomness. Thus, combining this result together with the decomposition (7.9), and using the assumption that X_t is independent from $\bigcup_{a\in\mathscr{A}} \mathscr{S}_{a,t-1}$, one deduces that with

probability higher than $1-2\delta$

$$\Delta_t \mathbb{I}\{I_t = \neg\star_t\} \leqslant \Big[\sum_{a' \in \{\star_t, \neg\star_t\}} \|X_t\|_{V^{-1}_{a',t-1}} B_{a',t-1} \Big] \mathbb{I}\{I_t = \neg\star_t\}, \tag{7.10}$$

where the short-hand notations $V_{a,t-1} \overset{\text{def}}{=} V_{\lambda_{t-1}}(\mathscr{S}_{a,t-1}(I^a_{t-1}))$ as well as $B_{a,t-1} \overset{\text{def}}{=} B_{\lambda_{t-1},a}(\mathscr{S}_{a,t-1}(I^a_{t-1}))$ are introduced for convenience.

Step 2: Now, $\|X_t\|_{V^{-1}_{a,t-1}}$ appearing in Equation (7.10) is bounded. By definition of $V_{a,t-1}$, it holds that

$$\|X_t\|^2_{V^{-1}_{a,t-1}} = X_t^\top \Big(\lambda_{t-1} I_d + \sum_{i \in I^a_{t-1}} X_{s_i} X_{s_i}^\top \Big)^{-1} X_t .$$

This expression is decomposed by using the definition of $X_s = \mu + \xi_s$ for all s.
Thus, on the one hand:

$$\|X_t\|_{V^{-1}_{a,t-1}} \leqslant \|\mu\|_{V^{-1}_{a,t-1}} + \|\xi_t\|_2 \lambda_{t-1}^{-1/2}, \tag{7.11}$$

where one used the fact that minimum eigenvalue of $V^{-1}_{a,t-1}$ is lower-bounded by λ_{t-1}.
On the other hand, the following decomposition holds:

$$\begin{aligned} V_{a,t-1} =& \lambda_{t-1} I_d + |I^a_{t-1}| \mu\mu^\top + \sum_{i \in I^a_{t-1}} \xi_{s_i} \xi_{s_i}^\top \\ &+ \Big(\sum_{i \in I^a_{t-1}} \xi_{s_i} \Big) \mu^\top + \mu \Big(\sum_{i \in I^a_{t-1}} \xi_{s_i}^\top \Big) \\ =& V + E + E_1 + E_2 , \end{aligned}$$

where the following four matrices are introduced:

$$\begin{aligned} V =& \lambda_{t-1} I_d + |I^a_{t-1}| \mu\mu^\top , \\ E =& \sum_{i \in I^a_{t-1}} \xi_{s_i} \xi_{s_i}^\top , \\ E_1 =& \mu \Big(\sum_{i \in I^a_{t-1}} \xi_{s_i}^\top \Big) \text{ and} \\ E_2 =& \Big(\sum_{i \in I^a_{t-1}} \xi_{s_i} \Big) \mu^\top . \end{aligned}$$

Now, μ is an eigenvector of the matrix V with associated eigenvalue $\lambda_\mu \overset{\text{def}}{=} \lambda_{t-1} + |I^a_{t-1}| \|\mu\|_2^2$. Thus, it holds that $\mu^\top V^{-1} \mu = \mu^\top V^{-1} \frac{V\mu}{\lambda_\mu} = \frac{\|\mu\|_2^2}{\lambda_\mu}$. The minimum eigenvalue of E is non-negative. μ is also an eigenvector of the rank 1 matrix E_1 with eigenvalue $\lambda_{\mu,2} = \sum_{i \in I^a_{t-1}} \langle \xi_{s_i}, \mu \rangle$. Finally, the only non-zero eigenvalue of the rank 1 matrix E_2 is

$\sum_{i\in I^a_{t-1}}\langle \xi_{s_i},\mu\rangle$ (associated to the vector $\sum_{i\in I^a_{t-1}}\xi_{s_i}$).
Thus one deduces that the matrix norm of the vector μ cannot increase too much when V is perturbed by $E+E_1+E_2$: λ_μ is shifted by at most $\lambda_{\mu,2}+\min\{\lambda_{\mu,2},0\}$, which leads to the bound

$$\|\mu\|^2_{V^{-1}_{a,t-1}} \leqslant \frac{\|\mu\|_2^2}{\lambda_{t-1}+|I^a_{t-1}|\|\mu\|_2^2+2\min\{\sum_{i\in I^a_{t-1}}\langle\xi_{s_i},\mu\rangle,0\}},$$

under the condition that $\lambda_{t-1}+|I^a_{t-1}|\|\mu\|_2^2+2\min\{\sum_{i\in I^a_{t-1}}\langle\xi_{s_i},\mu\rangle,0\}>0$.
This condition happens with high probability, provided that the noise is small enough. Indeed, by the Chernoff method together with (7.3), it holds for all deterministic set I of size n, and for $\delta\in(0,1)$ that

$$\mathbb{P}\Big[\sum_{i\in I}\langle\xi_{s_i},\mu\rangle \leqslant -\|\mu\|_2\sigma_X\sqrt{2n\log(1/\delta)}\Big]\leqslant\delta\,.$$

Thus, since I^a_{t-1} is chosen independently on the samples, by a union bound over the possible values of the random size $|I^a_{t-1}|\leqslant t-1$ of the index set, it comes that on an event of probability higher than $1-\delta$,

$$\sum_{i\in I^a_{t-1}}\langle\xi_{s_i},\mu\rangle \geqslant -\|\mu\|_2\sigma_X\sqrt{2|I^a_{t-1}|\log((t-1)/\delta)}\,.$$

Thus, solving the condition $n\|\mu\|_2^2+\lambda_{t-1}-2\|\mu\|_2\sigma_X\sqrt{2n\log((t_1)/\delta)}>0$ in n, one observes that when $\lambda_{t-1}>2\sigma_X^2\log((t-1)/\delta)$, the condition is satisfied for all n.
Step 3: Plug-in this result and the bound on $\|\mu\|^2_{V^{-1}_{a,t-1}}$ in (7.11), and combining this together with (7.10), one deduces that at time t such that $I_t=\neg\star_t$, then with probability higher than $1-6\delta$,

$$\begin{aligned}\Delta_t \leqslant &\sqrt{\frac{\|\mu\|_2^2B^2_{\neg\star_t,t-1}}{\lambda_{t-1}+n_{t-1}\|\mu\|_2^2-2\|\mu\|_2\sigma_X\sqrt{2n_{t-1}\log(\frac{t-1}{\delta})}}}\\ &+\sqrt{\frac{\|\mu\|_2^2B^2_{\star_t,t-1}}{\lambda_{t-1}+n_{t-1}\|\mu\|_2^2-2\|\mu\|_2\sigma_X\sqrt{2n_{t-1}\log(\frac{t-1}{\delta})}}}\\ &+\|\xi_t\|_2\lambda_{t-1}^{-1/2}(B_{\neg\star_t,t-1}+B_{\star_t,t-1})\end{aligned}\tag{7.12}$$

where $n_{t-1}\overset{\text{def}}{=}|I^{\neg\star_t}_{t-1}|=|I^{\star_t}_{t-1}|=\min\{N_{t-1}(\neg\star_t),N_{t-1}(\star_t)\}$.
The next step is to simplify this expression by upper bounding both $B_{\neg\star_t,t-1}$ and $B_{\star_t,t-1}$. To this end, one notes that on the one hand $\lambda_{t-1}\|\theta_{\star_t}\|_{V^{-1}_{\neg\star_t,t-1}}\leqslant\lambda_{t-1}^{1/2}\|\theta_{\star_t}\|_2$, and on the other hand, using the fact that $\|X_t\|_2\leqslant C$ for all context vector X_t, where

$C \overset{\text{def}}{=} \|\mu\|_2 + \sigma_X$,

$$\begin{aligned}\det(V_{\neg\star_t,t-1}) &\leqslant \Big(\frac{\operatorname{trace}(V_{\neg\star_t,t-1})}{d}\Big)^d \\ &\leqslant \Big(\frac{\lambda_{t-1}d + n_{t-1}C^2}{d}\Big)^d \\ &\leqslant (\lambda_{t-1} + (t-1)C^2/d)^d\,.\end{aligned}$$

Thus, it holds for both $I' = \neg\star_t$ and $I' = \star_t$ that

$$B_{I',t-1} \leqslant R_\eta \sqrt{2d\log\Big(\frac{\lambda_{t-1}^{1/2} + \frac{(t-1)C^2}{d\lambda_{t-1}^{1/2}}}{\delta}\Big)} + \lambda_{t-1}^{1/2}\|\theta_{I'}\|_2\,. \tag{7.13}$$

For convenience, the first term on the left hand side in (7.13) is denoted $b_{t-1} = R_\eta\sqrt{2d\log\Big(\frac{\lambda_{t-1}^{1/2} + \frac{(t-1)C^2}{d\lambda_{t-1}^{1/2}}}{\delta}\Big)}$. Combining (7.13) together with (7.12), so far, and using the fact that $\|\theta\|_2 \leqslant B$ for all $\theta \in \Theta$, it is shown that for all t such that $I_t = \neg\star_t$, with probability higher than $1-6\delta$, then

$$\begin{aligned}\Delta_t \leqslant \Big[&\frac{2\|\mu\|_2}{\sqrt{\lambda_{t-1} + n_{t-1}\|\mu\|_2^2 - 2\|\mu\|_2\sigma_X\sqrt{2n_{t-1}\log(\frac{t-1}{\delta})}}} \\ &+ \|\xi_t\|_2\lambda_{t-1}^{-1/2}\Big]\Big[b_{t-1} + \lambda_{t-1}^{1/2}B\Big]\,,\end{aligned}$$

that is, after reorganizing the terms, and provided that the noise is not too large, i.e. $\Delta_t \geqslant \|\xi_t\|_2(b_{t-1}\lambda_{t-1}^{-1/2} + B)$, then

$$n_{t-1} - \frac{2\sigma_X}{\|\mu\|_2}\sqrt{2n_{t-1}\log(\frac{t-1}{\delta})} \leqslant \quad 4\Big(\frac{b_{t-1} + \lambda_{t-1}^{1/2}B}{\Delta_t - \|\xi_t\|_2(b_{t-1}\lambda_{t-1}^{-1/2} + B)}\Big)^2 - \frac{\lambda_{t-1}}{\|\mu\|_2^2}\,. \tag{7.14}$$

By introducing the threshold $\tau_t = \|\xi_t\|_2(b_{t-1}\lambda_{t-1}^{-1/2} + B)$, the regret is decomposed into:

$$R_{X,T} \leqslant \sum_{t=1}^{T}\Delta_t\mathbb{I}\{I_t = \neg\star_t, \Delta_t \geqslant \tau_t\} + \sum_{t=1}^{T}\Delta_t\mathbb{I}\{\Delta_t < \tau_t\}\,.$$

Step 4: In this step, two cases are considered. First, the case when $N_{t-1}(\neg\star_t) \geqslant N_{t-1}(\star_t)$. Second, when $N_{t-1}(\neg\star_t) < N_{t-1}(\star_t)$.

In the first situation, then $n_{t-1} = N_{t-1}(\star_t)$, and it comes from (7.14) that $N_{t-1}(\star_t)$ cannot be too large. Indeed, for positive constants A, B, the condition $n - A\sqrt{n} \leqslant B$ implies that $n \leqslant B + A^2/2\big(1+\sqrt{1+4B/A^2}\big)$, which in this case leads to $N_{t-1}(\star_t) \leqslant u_{\lambda,t}(\delta)$

where

$$u_{\lambda,t}(\delta) \overset{\text{def}}{=} 4\left(\frac{b_{t-1}+\lambda_{t-1}^{1/2}B}{\Delta_t - \|\xi_t\|_2(b_{t-1}\lambda_{t-1}^{-1/2}+B)}\right)^2 + \frac{4\sigma_X^2\log(\frac{t-1}{\delta})}{\|\mu\|_2^2} + \frac{4\sigma_X\sqrt{2\log(\frac{t-1}{\delta})}}{\|\mu\|_2}\left(\frac{b_{t-1}+\lambda_{t-1}^{1/2}B}{\Delta_t - \|\xi_t\|_2(b_{t-1}\lambda_{t-1}^{-1/2}+B)}\right) - \frac{\lambda_{t-1}}{\|\mu\|_2^2}.$$

Likewise, in the second case when $N_{t-1}(\neg\star_t) \leqslant N_{t-1}(\star_t)$, then one deduces that necessarily $N_{t-1}(\neg\star_t) \leqslant u_{\lambda,t}(\delta)$ with high probability, and thus a controlled regret.

From this point on, one can proceed similarly to the elementary proofs of the regret of the UCB algorithm (Auer et al., 2002), as in (Maillard et al., 2011; Bubeck, 2010) or of BESA (Baransi et al., 2014). More precisely, it holds, since $|\Delta_t| \leqslant 1$ by assumption, that

$$\begin{aligned}
R_{X,T} &\leqslant \sum_{t=1}^{T} \Delta_t \mathbb{I}\{I_t = \neg\star_t \cap \Delta_t \geqslant \tau_t \cap N_{t-1}(\star_t) > u_{\lambda,t}(\delta_t)\} \\
&\quad + \sum_{t=1}^{T} \mathbb{I}\{N_{t-1}(\star_t) \leqslant u_{\lambda_t}(\delta_t)\} + \sum_{t=1}^{T} \Delta_t \mathbb{I}\{\Delta_t < \tau_t\} \\
&\leqslant \sum_{t=1}^{T} \Delta_t \mathbb{I}\{I_t = \neg\star_t \cap \Delta_t \geqslant \tau_t \cap N_{t-1}(\neg\star_t) \leqslant u_{\lambda,t}(\delta_t)\} \\
&\quad + 6\sum_{t=1}^{T} \delta_t + \sum_{t=1}^{T} \mathbb{I}\{N_{t-1}(\star_t) \leqslant u_{\lambda_t}(\delta_t)\} + \sum_{t=1}^{T} \Delta_t \mathbb{I}\{\Delta_t < \tau_t\},
\end{aligned}$$

for any choice of $\delta_t \in (0,1)$ for $t \in \{1,\dots,T\}$. In particular, this holds for the choice $\delta_t = t^{-2}$.

The first sum is splitted into the sum over the time steps for which $\star_t = a$ and the sum over the time steps for which $\star_t = b$. Note that Δ_t is *not* independent from I_t. For the sum such that $\star_t = b$, it comes

$$\begin{aligned}
\sum_{t=1}^{T} \Delta_t \mathbb{I}\{I_t = a \cap N_{t-1}(a) \leqslant u_{\lambda,t}(\delta_t)\}\mathbb{I}\{\star_t = b\} &\leqslant (\max_{t\leqslant T}\Delta_t)\sum_{t=1}^{T}\mathbb{I}\{I_t = a \cap N_{t-1}(a) \leqslant \max_{t\leqslant T} u_{\lambda,t}(\delta_t)\}\mathbb{I}\{\star_t = b\} \\
&\leqslant (\max_{t\leqslant T}\Delta_t)(\max_{t\leqslant T} u_{\lambda,t}(\delta_t)).
\end{aligned}$$

Likewise, a similar control can be obtained for the sum corresponding to $\star_t = a$.

In order to control the maximum terms, one uses the fact that $\|\xi_t\|_2^2 \leqslant \sigma_X^2$. Thus, it holds that

$\max_{t\leqslant T} u_{\lambda,t}(\delta_t) \leqslant \overline{u}$ where

$$\overline{u} \stackrel{\text{def}}{=} 16\left[\frac{R_\eta\sqrt{2d\log\left(\lambda_T^{1/2}T^2+\frac{T^3C^2}{d\lambda_T^{1/2}}\right)}+\lambda_T^{1/2}B}{\min_{t\in[T]}\Delta_t}+\frac{\sigma_X\sqrt{6\log(T)}}{\|\mu\|_2}\right]^2-\frac{\lambda_T}{\|\mu\|_2^2},$$

provided that the context-noise is small enough that

$$\min_{t\in[T]}\Delta_t > \tau \stackrel{\text{def}}{=} 2\sigma_X\left[R_\eta\sqrt{\frac{2d}{\lambda_T}\log\left(\lambda_T^{1/2}T^2+\frac{T^3C^2}{d\lambda_T^{1/2}}\right)}+B\right].$$

To sum up this step, it has been shown so far that $R_{X,T}$ is bounded as

$$\begin{aligned}R_{X,T} \leqslant & 2\big(\max_{t\leqslant T}\Delta_t\big)\overline{u}+\pi^2+\sum_{t=1}^{T}\mathbb{I}\{N_{t-1}(\star_t)\leqslant u_{\lambda_t}(t^{-2})\}\\ &+\sum_{t=1}^{T}\Delta_t\mathbb{I}\{\min_{t\in[T]}\Delta_t\leqslant\tau\}.\end{aligned}$$

Step 5: In order to conclude this proof, the term that now needs to be controlled is

$$\sum_{t=1}^{T}\mathbb{I}\{N_{t-1}(\star_t)\leqslant u_{\lambda_t}(\delta_t)\}=\sum_{t\in T_a}\mathbb{I}\{N_{t-1}(a)\leqslant u_{\lambda_t}(\delta_t)\}+\sum_{t\in T_b}\mathbb{I}\{N_{t-1}(b)\leqslant u_{\lambda_t}(\delta_t)\},$$

where $T_a \stackrel{\text{def}}{=} \{t\in[T]:\star_t=a\}$ for $a\in\mathscr{A}$.
To this end, a procedure similar to that used in (Baransi et al., 2014) is employed. More precisely, following the exact same steps as Steps 3 and 4 of the proof of Theorem 1 in (Baransi et al., 2014), it comes that

$$\sum_{t\in T_b}\mathbb{P}\Big[N_{t-1}(b)\leqslant u_{\lambda_t}(\delta_t)\Big]\leqslant c+\sum_{t\in T_b,t\geqslant c}\sum_{j=1}^{\lfloor u_{\lambda_t}(\delta_t)\rfloor}\alpha_{b,a}(M_t,j)+O(1),$$

where c is a constant such that $t\geqslant c$ implies $t\geqslant u_{\lambda_t}(\delta_t)(u_{\lambda_t}(\delta_t)+1)$, $M_t\in\mathbb{N}$ is such that $M_t=O(\log(t))$ and where the function $\alpha_{b,a}(M,j)$ is defined by[1]

$$\alpha_{b,a}(M,j)=\mathbb{E}_{Z^b_{1:j}}\left[\mathbb{P}_{Z^a_{1:j},X}\Big(\langle X,\widehat{\theta}(Z^a_{1:j})-\widehat{\theta}(Z^b_{1:j})\rangle>0\Big)^M\right].$$

Here, one used explicitly the stochastic nature of the context X_t, to avoid having to deal with much more complex expressions. This comes at the price of restricting to

[1]Indeed, the tie event $\langle X,\widehat{\theta}(Z^a_{1:j})-\widehat{\theta}(Z^b_{1:j})\rangle=0$ has probability 0, since the distributions are diffuse.

cases when the noise it not too strong. Here, $Z^b_{1:j}$ denotes a set of j i.i.d. samples $Z^b_j = (X_j, Y_j)$ generated from the model considered in the introduction, when arm b is chosen (that is, such that $Y_j = \langle X_j, \theta_b \rangle + \eta_i$ and $X_j = \mu + \xi_j$), $Z^a_{1:j}$ denotes a similar set built using θ_a instead of θ_b, and $X = \mu + \xi$ is generated by Nature.

Step 6: The next step of the proof is to control the quantity $\alpha_{b,a}(M, j)$ for steps T_b (and likewise $\alpha_{a,b}(M, j)$ for steps T_a). In the sequel, the notation $b = \star$ is used to clarify that b is the optimal arm in time-steps T_b. One wants to show that this decays exponentially fast to 0 with either M or j, so that the contribution to the regret is controlled. To begin with, the dot product is decomposed according to the different random variables

$$\begin{aligned}\langle X, \widehat{\theta}(Z^a_{1:j}) - \widehat{\theta}(Z^\star_{1:j})\rangle =& \langle \mu, \theta_\star - \widehat{\theta}(Z^\star_{1:j})\rangle + \langle \mu, \widehat{\theta}(Z^a_{1:j}) - \theta_a\rangle + \langle \xi, \widehat{\theta}(Z^a_{1:j}) - \theta_a\rangle \\ &+ \langle \xi, \theta_\star - \widehat{\theta}(Z^\star_{1:j})\rangle + \langle \xi, \theta_a - \theta_\star\rangle - \Delta,\end{aligned}$$

where $\Delta = \langle \mu, \theta_\star - \theta_t \rangle$. Then, it holds for all $\varepsilon > 0$ that

$$\mathbb{P}_X\Big(\langle \xi, \theta_a - \theta_\star\rangle \geqslant \varepsilon\Big) \leqslant \exp\Big(-\frac{\varepsilon^2}{2\|\theta_a - \theta_\star\|_2^2 \sigma_X^2}\Big). \tag{7.15}$$

The term $\langle \xi, \theta_\star - \widehat{\theta}(Z^\star_{1:j})\rangle$ is controlled a bit differently, by

$$\mathbb{P}_X\Big(\langle \xi, \theta_\star - \widehat{\theta}(Z^\star_{1:j})\rangle \geqslant \varepsilon\Big) \leqslant \exp\Big(-\frac{\varepsilon^2 \lambda_{t_j}}{2\|\theta_\star - \widehat{\theta}(Z^\star_{1:j})\|^2_{V_{\star,t_j}} \sigma_X^2}\Big), \tag{7.16}$$

where $t_j \geqslant j$ corresponds to a time when arm a is sampled at least j times (note that this is a probability with respect to X, not $Z^\star_{1:j}$).

Likewise, $\langle \xi, \widehat{\theta}(Z^a_{1:j}) - \theta_a\rangle$ is controlled by

$$\mathbb{P}_{Z^a_{1:j}, X}\Big(\langle \xi, \widehat{\theta}(Z^a_{1:j}) - \theta_a\rangle \geqslant \varepsilon\Big) \leqslant \Big(\lambda_{t_j}^{1/2} + \frac{jC^2}{d\lambda_{t_j}^{1/2}}\Big) \exp\Big(-\frac{\lambda_{t_j}(\varepsilon - \sigma_X\|\theta^\star\|_2)^2}{2\sigma_X^2 R^2}\Big), \tag{7.17}$$

for all $\varepsilon > \sigma_X\|\theta^\star\|_2$.

Finally, it has already been shown that, with probability higher than $1 - \delta - \delta'$ with respect to $Z^a_{1:j}$, then

$$\langle \mu, \theta_a - \widehat{\theta}(Z^a_{1:j})\rangle \leqslant \frac{\|\mu\|_2 R_\eta \sqrt{2\log\Big(\frac{\lambda_{t_j}^{1/2} + \frac{jC^2}{d\lambda_{t_j}^{1/2}}}{\delta}\Big)} + \lambda_{t_j}^{1/2}\|\mu\|_2\|\theta_a\|_2}{\sqrt{\lambda_{t_j} + j\|\mu\|_2^2 - 2\|\mu\|_2\sigma_X\sqrt{2j\log(1/\delta')}}}. \tag{7.18}$$

Inverting this bound in δ, this gives for all $\varepsilon \geqslant \frac{\|\mu\|_2}{\sqrt{j\|\mu\|_2^2+\lambda_{t_j}}}$, and $\varepsilon' \geqslant \lambda_{t_j}^{1/2}\|\theta_a\|_2$, that

$$\mathbb{P}_{Z^a_{1:j}}\Big(\langle \mu, \theta_a - \widehat{\theta}(Z^a_{1:j})\rangle \geqslant \varepsilon\varepsilon'\Big) \leqslant \exp\Big(-\frac{\big(j\|\mu\|_2^2 + \lambda_{t_j} - \|\mu\|_2^2/\varepsilon^2\big)^2}{8\|\mu\|_2^2\sigma_X^2 j}\Big) + \Big(\lambda_{t_j}^{1/2} + \frac{jC^2}{d\lambda_{t_j}^{1/2}}\Big)\exp\Big(-\frac{\big(\varepsilon' - \lambda_{t_j}^{1/2}\|\theta_a\|_2\big)^2}{2R_\eta^2}\Big). \tag{7.19}$$

Thus, by combining equations (7.15), (7.16), (7.17) and (7.19) together it comes

$$\begin{aligned}
\mathbb{P}_{Z^a_{1:j},X}\Big(\langle X, \widehat{\theta}(Z^a_{1:j}) - \widehat{\theta}(Z^\star_{1:j})\rangle > 0\Big) \leqslant & \mathbb{P}_{Z^a_{1:j},X}\Big(\langle \mu, \widehat{\theta}(Z^a_{1:j}) - \theta_a\rangle + \langle \xi, \widehat{\theta}(Z^a_{1:j}) - \theta_a\rangle \\
& + \langle \xi, \theta_\star - \widehat{\theta}(Z^\star_{1:j})\rangle + \langle \xi, \theta_a - \theta_\star\rangle > \Delta - \langle \mu, \theta_\star - \widehat{\theta}(Z^\star_{1:j})\rangle\Big) \\
\leqslant & \inf_{\varepsilon_0,\dots,\varepsilon_4}\Big\{\exp\Big(-\frac{\big(j\|\mu\|_2^2 + \lambda_{t_j} - \|\mu\|_2^2/\varepsilon_0^2\big)^2}{8\|\mu\|_2^2\sigma_X^2 j}\Big) \\
& + \Big(\lambda_{t_j}^{1/2} + \frac{jC^2}{d\lambda_{t_j}^{1/2}}\Big)\exp\Big(-\frac{\big(\varepsilon_1 - \lambda_{t_j}^{1/2}\|\theta_a\|_2\big)^2}{2R_\eta^2}\Big) \\
& + \Big(\lambda_{t_j}^{1/2} + \frac{jC^2}{d\lambda_{t_j}^{1/2}}\Big)\exp\Big(-\frac{\lambda_{t_j}(\varepsilon_2 - \sigma_X\|\theta^\star\|_2)^2}{2\sigma_X^2 R_\eta^2}\Big) \\
& + \exp\Big(-\frac{\varepsilon_3^2\lambda_{t_j}}{2\|\theta_\star - \widehat{\theta}(Z^\star_{1:j})\|^2_{V_{\star,t_j}}\sigma_X^2}\Big) + \exp\Big(-\frac{\varepsilon_4^2}{2\|\theta_a - \theta_\star\|_2^2\sigma_X^2}\Big) : \\
& \varepsilon_0\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 \leqslant \Delta - \langle \mu, \theta_\star - \widehat{\theta}(Z^\star_{1:j})\rangle, \\
& \varepsilon_2 > \sigma_X\|\theta^\star\|_2, \varepsilon_0 \geqslant \frac{\|\mu\|_2}{\sqrt{j\|\mu\|_2^2+\lambda_{t_j}}}, \varepsilon_1 \geqslant \lambda_{t_j}^{1/2}\|\theta_a\|_2\Big\}.
\end{aligned}$$

By choosing $\varepsilon_0 = \frac{\sqrt{2}\|\mu\|_2}{\sqrt{j\|\mu\|_2^2+\lambda_{t_j}}}$, $\varepsilon_1 = 2\lambda_{t_j}^{1/2}\|\theta_a\|_2$, $\varepsilon_2 = 2\sigma_X\|\theta^\star\|_2$, $\varepsilon_3 = \sqrt{2}\sigma_X\|\theta^\star\|_2$, and $\varepsilon_4 = \sqrt{2}\sigma_X\|\theta^a - \theta_\star\|_2\kappa$, for some positive κ, the previous expression can be simplified. This way, and using the fact that $\lambda_T \geqslant \lambda_{t_j} \geqslant \lambda_j$, finally establishes the following bound

on the quantity $\alpha_{\star,a}(M_t, j)$ that has to be controlled:

$$\alpha_{\star,a}(M_t, j) \leqslant \mathbb{E}_{Z^\star_{1:j}}\Bigg[\Bigg(e^{-\kappa^2} + e^{-\frac{\left(j\|\mu\|_2^2+\lambda_j\right)^2}{32\|\mu\|_2^2\sigma_X^2 j}} + \left(\lambda_T^{1/2} + \frac{jC^2}{d\lambda_j^{1/2}}\right)\left[e^{-\frac{\lambda_j\|\theta_a\|_2^2}{2R^2}} + e^{-\frac{\lambda_j\|\theta^\star\|_2^2}{2R^2}}\right] + e^{-\frac{\lambda_j\|\theta_\star\|_2^2}{\|\theta_\star - \widehat{\theta}(Z^\star_{1:j})\|^2_{V_{\star,t_j}}}}\Bigg)^{M_t} \mathbb{I}\{\mathscr{E}\} + \mathbb{I}\{\mathscr{E}^c\}\Bigg]. \tag{7.20}$$

For convenience, the event $\mathscr{E}$ has been introduced:

$$\mathscr{E} \overset{\text{def}}{=} \left\{\langle \mu, \theta_\star - \widehat{\theta}(Z^\star_{1:j})\rangle \leqslant \Delta - \frac{2\sqrt{2}\|\mu\|_2\|\theta_a\|_2}{\sqrt{j\|\mu\|_2^2/\lambda_{t_j}+1}} - \sqrt{2}\sigma_X B(1+\sqrt{2}+2\kappa)\right\}.$$

At this point, one notes that since $\lambda_j \geqslant 6\sigma_X^2\log(j)$, all the exponential terms but $e^{-\kappa^2}$ decay polynomially fast to 0 with j, and that for $M_t = O(\log(t))$, the first half of the bound on $\alpha_{\star,a}(M_t, j)$ decays polynomially fast to 0 with t, at a rate that can always be adjusted to be $t^{-(1+\beta)}$ for some small $\beta > 0$. Thus, in order to control the regret term of Step 5, one deduces from this observation and (7.20) that it only remains to show that $\mathbb{P}_{Z^\star_{1:j}}(\mathscr{E}^c)$ is small enough under our assumptions on the noise. From equation, (7.19), it is not difficult to see that

$$\mathbb{P}_{Z^\star_{1:j}}(\mathscr{E}^c) \leqslant e^{-\frac{(j\|\mu\|_2^2+\lambda_j)^2}{32\|\mu\|_2^2\sigma_X^2 j}} + \left(\lambda_T^{1/2} + \frac{jC^2}{d\lambda_j^{1/2}}\right)e^{-\frac{\lambda_j\|\theta^\star\|_2^2}{2R_\eta^2}},$$

provided that the following condition holds

$$\frac{2\sqrt{2}(\|\theta_a\|_2 + \|\theta_\star\|_2)}{\sqrt{j/\lambda_T + \|\mu\|_2^{-2}}} \leqslant \Delta - \sqrt{2}\sigma_X B(1+\sqrt{2}+2\kappa).$$

Finally, in order for term $\mathbb{P}_{Z^\star_{1:j}}(\mathscr{E}^c)$ to decay fast enough with t (at least $t^{-(1+\beta)}$), it is enough to choose $\lambda_t \geqslant c\log(T)$ for all t for some constant c, which leads to

$$\sum_{t\in T_b, t\geqslant c_\Delta}\ \sum_{j=1}^{\lfloor u_{\lambda_t}(\delta_t)\rfloor} \alpha_{b,a}(M_t, j) = O(1).$$

Step 7: Now that $\alpha_{\star,a}(M_t, j)$ is controlled, $\alpha_{a,b}(M_t, j)$ can be treated similarly. Thus,

at the price of loosing a factor 2 (using $T_a \leqslant T$ and $T_b \leqslant T$), it has been shown that

$$\mathbb{E}[R_{X,T}] = \Big(\max_{t\in[T]} \Delta_t\Big)\overline{U} + \sum_{t=1}^{T} \Delta_t \mathbb{I}\{\min_{t\in[T]} \Delta_t \leqslant \tau\}\Big) + O(1)\,.$$

where

$$\overline{U} \stackrel{\text{def}}{=} \frac{64}{\min_{t\in[T]} \Delta_t^2}\left[R_\eta\sqrt{2d\log\left(\lambda_T^{1/2}T^2 + \frac{T^3C^2}{d\lambda_T^{1/2}}\right)} + \lambda_T^{1/2}B\right]^2 + \frac{24\sigma_X^2\log(T) - 2\lambda_T}{\|\mu\|_2^2}$$

and

$$\tau \stackrel{\text{def}}{=} 2\sigma_X\left[R_\eta\sqrt{\frac{2d}{\lambda_T}\log\left(\lambda_T^{1/2}T^2 + \frac{T^3C^2}{d\lambda_T^{1/2}}\right)} + B\right].$$

This concludes the proof after some cosmetic simplifications of $\overline{U}$ using $(a+b)^2 \leqslant 2(a^2+b^2)$ for positive a, b. □

7.4 Experimental study

This section presents numerical experiments illustrating the behavior of `CL-BESA` and supporting the discussion of its performances comparatively to the state of the art.

7.4.1 Experimental setting

For the sake of comparison, the presentation will focus on the 2-arm bandit problem, empirically comparing `CL-BESA` with the following baseline algorithms:

- OFUL, described in section 4.6.1, is a MAB algorithm which does not consider the contextual information (Abbasi-Yadkori et al., 2011); it assumes that the set of arms is known (with parameters θ_a and θ_b known), and learns a shared unknown parameter β with

  $$Y_t = \langle \beta^\star, \theta_{I_t} \rangle + \eta_t$$

  The computational trick, which consists in sparsely recomputing β_t is not used in the experiments due to the difficulty of tuning the underlying hyper-parameter C.

- `Thompson sampling` is a family of (pseudo-)Bayesian algorithms (Agrawal and Goyal, 2012a); like OFUL, it does not consider the contextual information (Abbasi-Yadkori et al., 2011); it assumes that the set of arms is known (with parameters θ_a and θ_b known), and maintains a prior distribution on the parameter $\beta^\star$.

- `LinUCB` is a linear contextual MAB algorithm, with same setting and same noise model as `CL-BESA`. Note however that (Li et al., 2010) considers $X_{i,t}$ context vector which can depend on the arm. This is a slightly more general framework than `CL-BESA`, although `CL-BESA` is able to cope with it.

As said, these algorithms are divided in two categories:

- `CL-BESA` and `LinUCB` learn an unknown parameter θ_i governing the reward gathered by the i-th arm.

- `OFUL` and `Thompson sampling` learn an unknown parameter β shared among all arms.

For the sake of a fair comparison with `CL-BESA`, the `OFUL` and `Thompson sampling` settings are adapted to match the `CL-BESA` setting. A straightforward scheme is to concatenate the unknown parameters θ_a and θ_b into one unknown vector $\overline{\theta} \stackrel{\text{def}}{=} (\theta_a \theta_b) \in \mathbb{R}^{dK}$ to be learned. The corresponding context vectors are $X_{t,a} = (X_t^T 0_d^T)^T \in \mathbb{R}^{dK}$ and $X_{t,b} = (0_d^T X_t^T)^T \in \mathbb{R}^{dK}$.
This adaptation enables to apply `OFUL` and `Thompson sampling`. Also, the resulting comparison will be fair as every algorithm has to learn a distinct model and cannot take advantage of any shared information. Indeed, `Thompson sampling` and `OFUL` in their original setting use *every* sample of *every* arm up to time t to reach a decision, leading to favor these algorithms especially in case of imbalanced trials among the arms.

7.4.2 Illustrative problem

Algorithms are evaluated on the following bi-dimensional MAB problems with $K = 2$ arms defined by $\mu = (0.5, 0.5)^T$, $\theta_a = (0.5, 0)^T$ and $\theta_b = (0, 0.5)^T + (0, 2\Delta)^T$. The gap is set to $\Delta = 10^{-1}$ and the time horizon to $T = 1000$. At each time t, a context X_t is uniformly drawn in the ball $\mathcal{B}(\mu, \sigma_X^2/2)$ and the immediate reward is given by $Y_t = \langle X_t, \theta_{I_t} \rangle + \eta_t$ with $\eta_t \sim \mathcal{N}(0, R_\eta^2)$.
This problem will be refered to as the **orthogonal problem** in the sequel.

7.4.3 Sensitivity analysis w.r.t. parameters

Table 7.2 lists the parameters involved in all baseline algorithms and in `CL-BESA`. Note that all four algorithms share one parameter governing the regularization of a least-squares problem: λ (implicitly set to 1 for `Thompson sampling` and `LinUCB`).

Table 7.2: Table of parameters

Algorithms	Parameters
`CL-BESA`	Regularization: $\lambda \in \mathbb{R}^\star$
`OFUL` (Abbasi-Yadkori et al., 2011)	Regularization: $\lambda \in \mathbb{R}^\star$ Confidence level: $\delta \in (0,1)$ Sub-Gaussian additive noise: $R_{OFUL} \in \mathbb{R}_+$ Parameter upper bound $S_{OFUL} = \mathbb{R}_+$
`LinUCB` (Li et al., 2010)	Confidence level: $\delta \in (0,1)$
`Thompson sampling` (Agrawal and Goyal, 2012a)	$\varepsilon \in (0,1)$ Confidence level: $\delta \in (0,1)$ Sub-Gaussian additive noise: $R_{TS} \in \mathbb{R}_+$

CL-BESA

As said, `CL-BESA` features a single parameter, λ. This is an important feature for both its implementation and its applications as parameter tuning often is a daunting problem with potentially huge influence on the performances.

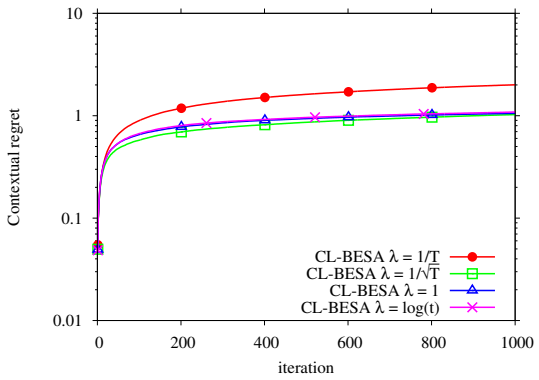


Figure 7.1: Contextual regret, in logarithmic scale as a function of the number of iterations on the orthogonal problem with $\Delta = 10^{-1}, \sigma_X/\sqrt{2} = \Delta, R_\eta = \Delta$ of `CL-BESA` for various choices of the regularization parameter λ. Results averaged over 1000 runs.

Fig. 7.1 shows the contextual regret of `CL-BESA` for various values of λ on the orthogonal problem with a noise level R_η and context perturbation $\frac{\sigma_X}{\sqrt{2}}$ equal to the gap Δ. The curves illustrate the robustness of the algorithm with respect to the tuning of λ, leading to a virtually parameter-free approach.

OFUL

OFUL involves three parameters: R_{OFUL}, S_{OFUL}, and δ, all of them influencing the size of the confidence ellipsoid. In the sequel, R_{OFUL} and S_{OFUL} will be set to their ideal values, respectively to $R_{OFUL} = R_\eta$ and $S_{OFUL} = \|\overline{\theta}\|_2$. $\delta = 10^{-4}$ as in the experimental study of (Abbasi-Yadkori et al., 2011).

This setting corresponds to a favorable case for OFUL as the optimal parameter values are not known in practice and may be difficult to estimate. Figure 7.4.3 shows the sensitivity of OFUL w.r.t. the parameter values and their impact on the contextual regret of OFUL. The same problem and the same noise and contextual perturbation levels are considered. Influence of λ, R_{OFUL} and S_{OFUL} are individually exhibited by fixing all the others parameter values.

Figure 7.2(a) illustrates the sensitivity of OFUL to parameter λ, as $\lambda = \frac{1}{T}$ leads to a high regret level. Also, Figures 7.2(b) and 7.2(c) exhibit the large deterioration of performances if R_{OFUL} and/or S_{OFUL} are one or two orders of magnitude too loosely estimated. This case happens not infrequently as the user mostly have no prior information about R_η or $\|\overline{\theta}\|_2$.

LinUCB

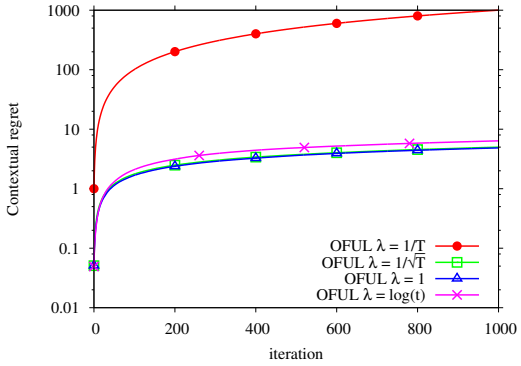
LinUCB requires a confidence parameter δ controlling the size of the confidence interval. As discussed in (Li et al., 2010), LinUCB is sensitive to the value of δ, and α as advised and defined in (4.32) generally provides an over-conservative bound.

Thompson Sampling

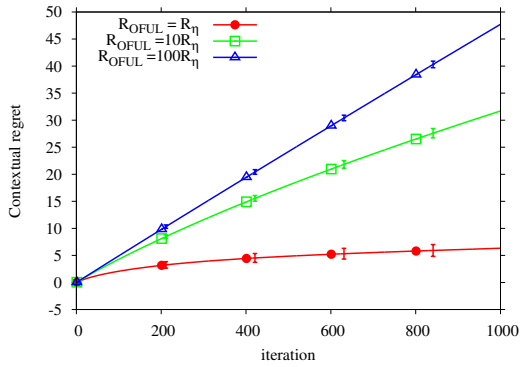
As for OFUL, Thompson sampling requires a confidence level $\delta \in (0,1)$ and bound R_{TS} on the sub-Gaussian noise level R_η and are respectively set to 10^{-4} and R_η. As advised in (Agrawal and Goyal, 2012a), ε will be set as $\varepsilon = \frac{1}{\log T}$ in the following experiments, ensuring a $\widetilde{O}(d^2\sqrt{T})$ regret.

7.4.4 Influence of noise and perturbations levels

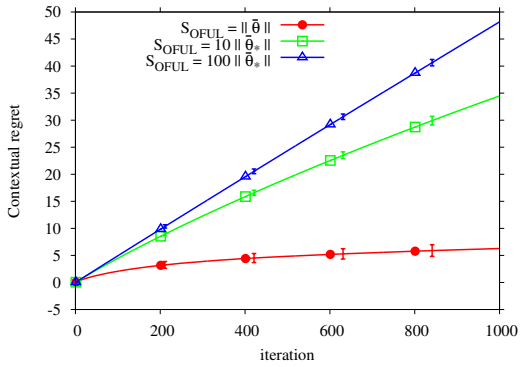
Figure 7.3 represents the contextual regret of CL-BESA, OFUL, LinUCB and Thompson sampling on the orthogonal problem for different levels of additive noise $R_\eta \in \{\Delta, 10\Delta\}$ and context perturbation $\frac{\sigma_X}{\sqrt{2}} \in \{\frac{\Delta}{10}, \Delta, 10\Delta, 100\Delta\}$. Algorithms are tuned as described in the previous section and, in particular, OFUL and Thompson sampling are provided with the exact values of $\|\overline{\theta}\|_2$ and R_η.



(a) Influence of λ. R_{OFUL} and S_{OFUL} are optimally tuned.



(b) Influence of R_{OFUL}. S_{OFUL} optimally tuned. $\lambda = 1$



(c) Influence of S_{OFUL}. R_{OFUL} optimally tuned. $\lambda = 1$

Figure 7.2: Parameter influence on the contextual regret of OFUL on the orthogonal problem with $\Delta = 0.1$, $(\frac{\sigma_X}{\sqrt{2}}, R_\eta) = (\Delta, \Delta)$ and $\lambda = 1$. Left: $R_{OFUL} \in \{R_\eta, 10R_\eta, 100R_\eta\}$, $S_{OFUL} = ||\bar{\theta}||_2$. Right: $S_{OFUL} \in \{||\bar{\theta}||_2, 10||\bar{\theta}||_2, 100||\bar{\theta}||_2\}$, $R_{OFUL} = R_\eta$. Results are averaged over 1000 runs.

Expected contextual regrets

The plots confirm the good performances of CL-BESA in all considered settings. More precisely, its expected contextual regret remains controlled, logarithmic and is quasi-optimal: its rank is 2nd or 1st on all settings, and furthermore when it is ranked second its regret is close to that of the first algorithm. More generally, the comparison of the curves shows that the algorithm ranks widely vary depending on the setting. For instance the third line ($\frac{\sigma_X}{\sqrt{2}} = 10\Delta$) shows that OFUL outperforms CL-BESA slightly and LinUCB significantly when $R_\eta = \Delta$ (left) but is clearly outperforms by LinUCB when $R_\eta = 10\Delta$ (right). The stability of CL-BESA and the general good level of its performances thus make it a good choice for the practitionner in the general case of unknown noise and/or perturbation level, ensuring a low if not optimal contextual

regret.
The bad performances of `Thompson sampling` are coherent with the $O(\sqrt{T})$ theoretical regret bound and are certainly due to over-conservative bounds and the known difficulty of parameter tuning.

Stability

Figure 7.3 also illustrates the stability of the four studied algorithms. A first observation is that, for each algorithm, the variance of the regret is increasing with the additive noise level R_η (from left to right) and with the contextual perturbation $\frac{\sigma_X}{\sqrt{2}}$ (from top to bottom), with an apparent higher sensitivy to R_η.
However, one can observe a larger instability for `CL-BESA` to a additive high noise R_η (noise ten times larger than the gap Δ) compared to `OFUL` and `LinUCB`. Nevertheless, the expected regret of `CL-BESA` remains competitive; this is explained (and confirmed by inspection) by the presence of catastrophic and excellent runs for the same setting. Finding a way of preventing the worst runs would lead to significant improvements of the sub-Sampling technique, and will define a perspective for further research.
In such extreme (and practically unlikely) situations, a practitioner may favor an alternative to `CL-BESA` depending of his or her objectives and available prior knowledge. For instance, in the case of risk-aversion, i.e. the user preferring reasonably high regret in worst cases rather than optimal expected regret, and reliable estimations of R_η and $\|\overline{\theta}\|_2$ are known, then `OFUL` or `LinUCB` should be preferred to `CL-BESA` depending of these estimates.
Otherwise, `CL-BESA` remains one of the best options. Note that in truly experimental conditions, `OFUL` would not benefit from the optimal parameters values S_{OFUL} and R_{OFUL}. In the considered benchmark, the exact knowledge of these (practically inaccessible) quantities favors `OFUL`. In particular the exact value of the additive noise level R_η, source of difficulty, is provided to the algorithm. As previously showed, imprecisions in the setting of the parameters would certainly lead to an increase of the expected contextual regret of `OFUL`. Likewise, `LinUCB` is a sub-optimal alternative for $R_\eta = \Delta$.

7.4.5 Influence of the dimension

Lastly, the sensitivity of the algorithms with respect to the space dimension is examined, varying the dimension d in $2 \dots 200$. For $d \geqslant 2$, one considers the d-dimensional MAB orthognal problem with $K = 2$ arms defined by $\mu = (\frac{1}{\sqrt{2d}}, \dots, \frac{1}{\sqrt{2d}})^\top$, $\theta_{a,d} = (\sqrt{2/d}\Delta, 0, \dots, \sqrt{2/d}\Delta, 0)^\top$ and $\theta_{b,d} = (0, -\sqrt{2/d}\Delta, \dots, 0, -\sqrt{2/d}\Delta)^\top$. The margin is set to Δ.

Algorithms are parametrized in the same fashion as in the former bi-dimensional problem case: $R_{OFUL} = R_{TS} = R_\eta = \Delta$, $\delta = 10^{-4}$ (`OFUL` and `Thompson sampling`) and $S_{OFUL,d} = \|\overline{\theta}_d\|_2$ with $\overline{\theta}_d = (\theta_{a,d}^T \theta_{b,d}^T)^T$.
Figure 7.4 shows the contextual regret of `OFUL`, `Thompson sampling`, `LinUCB` and `CL-BESA` for $T = 1000$, $\Delta = 0.1$ when d varies from 1 to 200 (results are averaged over 20 runs).
`LinUCB` suffers from numerical instability in large dimensions, causing a high regret and variance. For this reason, no results were obtained with `LinUCB` for $d \geq 40$. `Thompson sampling` and `CL-BESA` obtain a stable regret level and `OFUL` regret slightly increases with the dimension.
As the problems as been normalized so that Δ remains constant regardless of the dimension, this results remains coherent with the regret bound derived in Theorem 11, scaling linearly with d.
As already said, the baseline algorithms have been parameterized using the optimal parameters (in particular for R_η and $\|\overline{\theta}_d\|_2$), to get their performances at their best. This makes it even more impressive that `CL-BESA` only requires parameter λ to be tuned and still delivers best and stable results over all dimensions d ranging in $1 \ldots 200$.

7.5 Discussion and perspectives

In this chapter, the goal is to provide a virtual parameter-free approach to linear contextual bandits, extending the seminal work of (Baransi et al., 2014) to the contextual setting. The contributions are structured as follows:

- Firstly, the performance of `CL-BESA` confirms that the subsampling approach should be viewed as a third category of MAB algorithms, along with the optimistic and (pseudo-)Bayesian approaches.

- On the theoretical side, the analysis of the contextual regret of `CL-BESA` is conducted, showing a *logarithmic* scaling with the time horizon, demonstrating the applicability of sub-sampling to contextual linear bandits. This result is counter-intuitive as one may think that sub-sampling implies a detrimental loss of information. Moreover, the proof of this regret bound requires non-trivial adaptations of previous proofs due to the sub-sampling.

- On the practical side, `CL-BESA` offers an easy implementation, low computational cost comparatively to the optimistic (respectively Bayesian) approaches, which require the maintenance of confidence sets (resp. of the posterior distribution).

- Finally, the empirical validation of CL-BESA comparatively to the state of the art confirms: i) the logarithmic regret scaling; ii) the robustness of CL-BESA w.r.t. the tuning of its unique parameter λ; iii) the robustness w.r.t. the dimension of the contextual problem; iv) last and not least, the stability of the CL-BESA performances, ranking first or second in a wide range of artificial experimental settings.

The research perspectives of this work are manifold. On the practical side, one would like to improve the (currently naive) implementation of the sub-sampling scheme, which might be very useful for high-dimensional contexts, and when the gap in expected payoff between arms is small (and thus there is significant overlap between subsamples at different rounds). Making efficient re-use of the regularized least-squares solution across similar sub-samples in this situation would result in an appreciable computational speed up.
On the theoretical side, it is seen that the regret has possibly high variance in difficult scenarios (large noise); a test, detecting unpromising runs, would significantly improve the average results in such cases.
A longer term perspective of research is to address the general (non-linear) contextual bandit problem and extend the subsampling approach to the general reinforcement learning setting, where current decisions affect future outcomes.

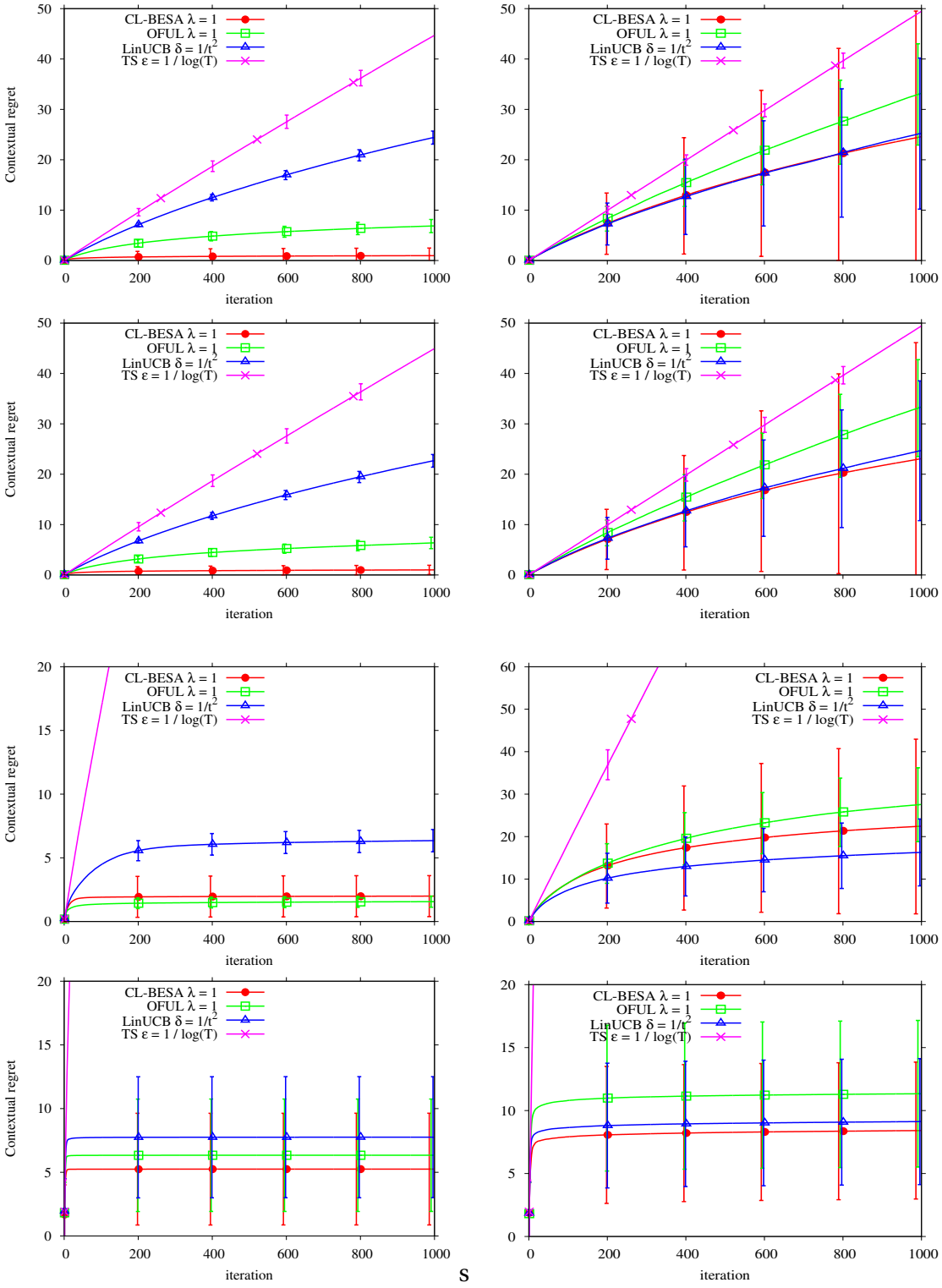


Figure 7.3: **Context perturbation and additive noise level:** Contextual regret as a function of the number of iterations on the orthogonal problem with $\Delta = 10^{-1}$ of `CL-BESA`, `OFUL` , `LinUCB` and `Thompson sampling`. $\delta_{OFUL} = \delta_{TS} = 10^{-4}$, $R_{OFUL} = R_{TS} = R_\eta$ and $S_{OFUL} = \|\theta\|_2$ (optimal values). Top to bottom: $\sigma_X/\sqrt{2} \in \{\frac{\Delta}{10}, \Delta, 10\Delta\, 100\Delta\}$. Left: $R_\eta = \Delta$; Right: $R_\eta = 10\Delta$.

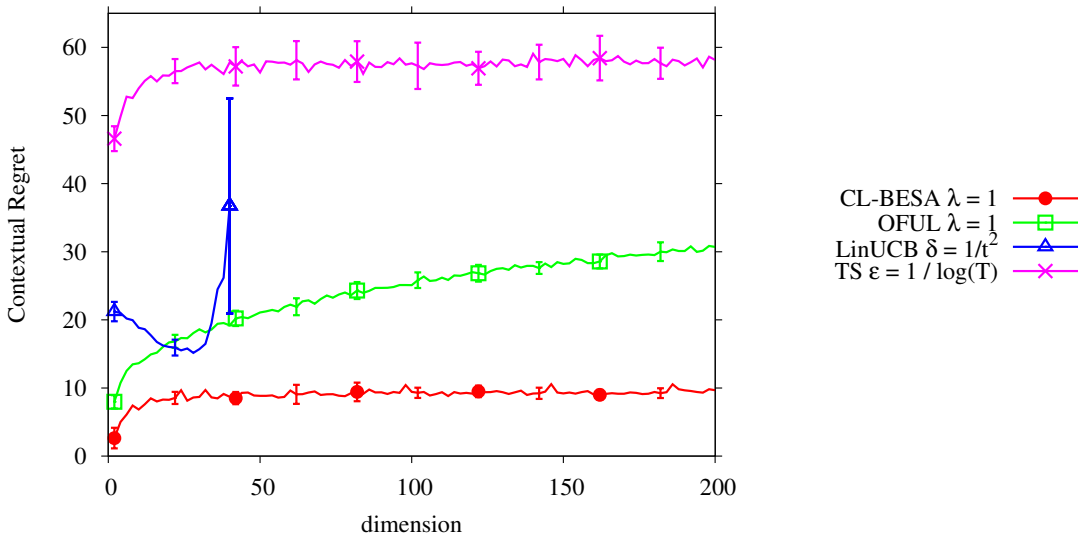


Figure 7.4: **Contextual regret** of `CL-BESA`, `OFUL`, `LinUCB` and `Thompson sampling` as a function of the dimension on an orthogonal problem with $\Delta = 10^{-1}, \sigma_X/\sqrt{2} = \Delta$, $R_\eta = \Delta$, $T = 1000$. $\delta_{OFUL} = \delta_{TS} = 10^{-4}$, $R_{OFUL} = R_{TS} = R_\eta$ and $S_{OFUL,d} = \|\overline{\theta}_d\|_2$ (optimal values).

Part III

Conclusions

Chapter 8

Conclusions and perspectives

This final chapter summarizes our contributions and presents future work directions.

8.1 Contributions

Our contributions include both theoretical and algorithmic extensions of the Multi-Armed Bandit setting:

1. Risk-awareness for the stochastic MAB (Chapter 6).

2. Sub-sampling for contextual linear bandits (Chapter 7).

8.1.1 Risk-Awareness for the stochastic Multi-Armed Bandits

In the Multi-Armed Bandit framework, the chosen definition of the risk is the possibility of getting low instantaneous rewards, potentially leading to hazards in real-world situations (e.g., for clinical testing). More generally, risk avoidance is an essential aspect in many applications; in such situations, it is suggested that one should focus on a three component trade-off: Exploration *vs* Exploitation *vs* Safety, instead of the classical two component trade-off.
While previous works focused on a mean-variance (Sani et al., 2012a) or a log-Laplace (Maillard, 2013) criterion, we study two new criteria:

- the essential infimum of the arms, leading to the algorithm `MIN`.

- the (modified) conditional value at risk at level $\alpha \in (0,1]$ ($CVaR_\alpha$), defined as the average value of the arms in the α% worst cases. This led to the derivation of two bandit algorithms termed *Multi-Armed Risk-Aware Bandits* (`MARAB`) and *Multi-Armed Risk-Aware Bandits Outhandled* (`MARABOUT`).

For MIN and MARABOUT, a theoretical analysis is proposed, based on concentration inequalities. These studies show a logarithmic regret of MIN under restrictive assumptions (probability mass lower-bounded by $A > 0$ in the neighborhood of the minimum, same optimal arms for min and expectation). For MARABOUT a logarithmic regret rate is shown for a (modified) CVaR-related regret.
Moreover, experiments conducted on both synthetic and real-world problems show the applicability of MARAB on short or moderate time horizons with low and stable empirical regrets in comparison with state-of-the-art methods. Also, experiments show the good performances of MARABOUT on challenging problems, even though the exploration constant C should be below the theoretical threshold. The empirical comparison between MARAB and MARABOUT shows a slightly lesser performance of MARABOUT on synthetic problems and a larger regret on the real-world energy problem. This performance loss is explained by the conservative (pessimistic) MARAB strategy, better suited to short-term horizons.

8.1.2 Sub-Sampling for Contextual Linear Bandits

The Contextual Bandit setting extends the stochastic MAB setting by adding side information available at each round and influencing the decision process. The Contextual Linear Bandits further add the assumption that the reward function is a linear function of the selected arm. In this framework, we introduce Contextual Linear Best Sub-Sampled Arm (CL-BESA), extending BESA (Baransi et al., 2014) and its sub-sampling approach.
We consider the challenging problem of estimating one parameter per arm (**disjoint model**) and derive an algorithm comparing two arms based on the same quantity of information. In the same fashion than BESA, we sub-sample from the most played arm as many observations as for the least sampled arm. This simple, though counter-intuitive, idea shows strikingly good properties in the stochastic case (simplicity, quasi-optimality, parameterlessness).
In the contextual linear case, the contextual regret is defined, encompassing the fact that the optimal arm might change at every round depending on the context. A non-trivial logarithmic bound for this regret is derived and empirical results are presented on synthetic problems, in comparison with state-of-the-art methods. It shows the excellent overall performances of the approach, under a wide variety of noise and perturbation conditions as well as in high-dimension settings. Indeed, the algorithm is the unique solution always ranked first or second in the worst case, without requiring the fine-tuning of hyper-parameters. This result demonstrates the wide applicability of the approach, providing the practitioner with best chances to obtain low contextual regret in an unknown environment (noise/perturbation levels). However, it also shows that CL-BESA seems more unstable (higher variance) than the competitors in the case

of high additive noise on the reward.

8.2 Future Work

The purpose of this section is to present some of the new research directions extending the work presented in this document.

8.2.1 Improvements of MARABOUT proof

The current proof is essentially similar to the one of UCB with distinct concentration inequalities and confidence bound. Refined techniques like peeling (see for instance (Bubeck, 2010)) or self-normalization (de la Peña et al., 2004; de la Peña et al., 2009) might allow tighter regret bound for MARABOUT.

8.2.2 Risk-Aware Reinforcement Learning

As described in Chapter 1, bandits can be exploited to design planning algorithms dealing with sequences of decisions in an unknown environment with varying states (Markov Decision Processes).
The most famous example of such algorithm is *Upper Confidence bound applied to Trees* (UCT, (Kocsis and Szepesvári, 2006b)) where UCB is used to navigate through the search tree. However, a study of its regret in worst case show a $\Omega(\exp(\exp(D)))$ with D being the tree depth (see (Coquelin and Munos, 2007)). Another approach is OLOP (Bubeck and Munos, 2010) which is demonstrated to be minimax optimal in term of simple regret.
Adapting one of these algorithms by replacing the UCB and mean-focused criterion by a criterion with a suited risk measure will lead to a *anytime* risk-aware reinforcement learning, able to provide a solution within a given computational budget.
These solutions will be compared to the one of (Moldovan and Abbeel, 2012) where safety is defined in term of *ergodicity*, i.e. the ability to reach any (safe) state from the current state with a suitable policy.

8.2.3 Extensions of CL-BESA

The CL-BESA algorithm and the sub-sampling strategy opens up new research directions:

- First, the higher noise sensitivity of CL-BESA in comparison to OFUL, Thompson sampling or LinUCB (Chapter 7), will be further investigated. Indeed, the high variance suggests room for improvement by mean of a characterization of the

bad runs of CL-BESA. At least, one can expect theoretical insight into when not to use CL-BESA.

- On a related way, it might be interesting to derive a risk-aware contextual linear bandit algorithm as the need for a safe exploration remains important in this setting.
- Finally, CL-BESA will be extended to the non-linear contextual setting.

Appendices

Appendix A

Résumé de la thèse

A.1 Introduction

Cette thèse s'intéresse au problème des bandits manchots stochastiques à bras multiples (*Multi-Armed Bandits* Thompson (1933); Lai and Robbins (1985)), une formalisation à la fois simple et riche des problèmes de prise de décision séquentielle en environnement inconnu et notamment du compromis exploration-exploitation.
Le problème se définit ainsi : un joueur entre dans une salle de casino équipée de K machines à sous ou bras dont les récompenses suivent des lois de distribution $\{\nu_i\}_{i\leqslant K}$ inconnues. Le but est alors de maximiser la somme des récompenses récoltées au cours d'un horizon temporel (nombre d'essais) T fini. Le compromis exploration-exploitation s'exprime sous la forme suivante :

- **Exploration** : le joueur entre sans connaissance *a priori* sur les distributions. Il doit donc essayer des machines pas ou peu testées auparavant pour identifier les meilleures.
- **Exploitation** : pour maximiser la somme des gains cumulés, il est indispensable de tirer le plus souvent possible les bras identifiés comme les meilleurs.

Il est important de noter que l'objectif ici n'est pas d'obtenir une évaluation précise des distributions associées à chaque bras mais de discriminer les meilleurs bras *le plus vite possible*.
Historiquement, l'étude de ce problème a débuté avec les essais cliniques (Thompson, 1933, 1935). T patients arrivent séquentiellement (un patient à chaque pas de temps), et nous disposons de K médicaments (bras) dont les effets sont inconnus. Le but étant de sauver le plus de patients possibles, le cadre des bandits manchots est particulièrement adapté à ce genre de situations.
Dans le cadre des bandits manchots, ce manuscrit de thèse apporte des contributions théoriques et algorithmiques selon deux axes principaux :

1. Le premier axe est la prise en considération du risque et la définition d'un compromis à trois composantes : Exploration-Exploitation-Sécurité. En effet, si l'objectif est de discriminer le plus rapidement possible les meilleurs bras, la question du critère de qualité associée à chaque bras reste centrale. Classiquement, les bras sont jugés selon la récompense moyenne qui leur est associée. Si ce critère se prête à beaucoup de domaines, il existe des situations où il semble inapproprié. Par exemple, dans le cadre d'une application médicale, on préférera à un médicament bon en moyenne mais potentiellement dangereux un médicament sûr en pire cas même au prix d'une légère dégradation des performances moyennes. Cet axe est développé dans le chapitre 6.

2. Le second axe s'intéresse aux bandits linéaires contextuels qui enrichissent le cadre stochastique de base en formulant deux hypothèses. Premièrement, la récompense est une fonction linéaire des bras. Ensuite, on suppose révélé à chaque instant t un contexte au joueur qui sert de support à sa prise de décision. Ce cadre permet d'inclure une information additionnelle aidant le joueur. Dans l'exemple des traitements médicaux, le contexte peut résumer les informations relatives au patient arrivant à l'instant t. Dans cette configuration, nous avons étendu un algorithme de bandit stochastique et basé sur le sous-échantillonage appelé Best Estimated Subsampled Arm (`BESA` (Baransi et al., 2014)). L'algorithme résultant, Contextual Linear Best Estimated Subsampled Arm (`CL-BESA`), a fait l'objet d'une étude théorique et empirique développées dans le chapitre 7

A.2 Regret

Nous notons :

- $K \in \mathbb{N}^\star$ le nombre de bras.
- ν_i la distribution du bras i
- $\mu_i \overset{\text{def}}{=} \mathbb{E}[\nu_i]$ l'espérance du bras i.
- $\mu^\star \overset{\text{def}}{=} \max_{i \in \{1,\dots,K\}} \mu_i$ l'espérance du (ou des) meilleur(s) bras.
- $i^\star$ un indice tel que : $\mu_{i^\star} = \mu^\star$.
- Δ_i la marge du bras i définie par $\Delta_i \overset{\text{def}}{=} \mu^\star - \mu_i$.
- $T \in \mathbb{N} \cup \{\infty\}$ l'horizon temporel.

- $I_t \in \{1, \dots, K\}$ le bras choisi à l'instant t.
- $X_{i,s}$ la récompense instantanée du bras i après s tirages.
- $N_{i,t} = \sum_{s=1}^{t} \mathbb{I}\{I_s = i\}$ le nombre de tirage du bras i jusqu'à l'instant t inclus.
- $Y_t = X_{I_t, N_{i,t}}$ la récompense instantanée reçue à l'instant t.

A.2.1 Définitions

Avec ces notations, l'objectif est donc de trouver une politique maximisant la somme $\sum_{t=1}^{T} Y_t$. De manière équivalente, on peut minimiser la perte engendrée par rapport à un oracle tirant toujours le bras optimal.
Formellement, cette perte est appelée le **regret cumulé** de la politique.

Définition A.2.1 (Regret cumulé)**.** *Le **regret cumulé** d'un agent au temps t est défini par :*

$$R_t \overset{\text{def}}{=} \underbrace{\max_{i \in \{1,\dots,K\}} \sum_{s=1}^{t} X_{i,s}}_{\text{Récompenses cumulées par un oracle}} - \underbrace{\sum_{s=1}^{t} Y_s}_{\text{Récompenses cumulées par l'agent}} . \tag{A.1}$$

Définition A.2.2 (Pseudo-regret cumulé)**.** *Le **pseudo-regret cumulé** d'un agent au temps t est défini par :*

$$\overline{R_t} \overset{\text{def}}{=} \sum_{s=1}^{t} \left(\mu^\star - \mu_{I_s}\right) = \sum_{i=1}^{K} \Delta_i N_{i,t} . \tag{A.2}$$

A.2.2 Borne inférieure

Cette section donne une borne inférieure théorique sur le regret d'un agent. Cette borne repose sur la divergence de Kullback-Leibler entre deux distributions.

Définition A.2.3 (Kullback and Leibler (1951))**.** *Soit $\mathscr{P}([0,1])$ l'ensemble des distributions de probabilités sur $[0,1]$. La divergence de Kullback-Leibler entre deux distributions P et Q dans $\mathscr{P}([0,1])$ est définie par :*

$$KL(P,Q) = \begin{cases} \int_{[0,1]} \frac{dP}{dQ} \log \frac{dP}{dQ} dQ & si\ P \ll Q \\ +\infty & sinon \end{cases} \tag{A.3}$$

Théorème A.2.1 (Burnetas and Katehakis (1996))**.** *Soit $\mathscr{P} \subset \mathscr{P}([0,1)$ et soit un joueur consistent avec $\mathscr{P}$, i.e. pour tout bras sous-optimal i et tout $\beta > 0$, $\mathbb{E}[N_{i,t}] = o(T^\beta)$. Alors,*

pour tout bandit stochastique avec distributions dans $\mathscr{P}$, on a :

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R_t]}{\log t} \geqslant \sum_{i:\Delta_i>0} \frac{\Delta_i}{\mathcal{K}_{\inf}(\nu_i, \mu^\star)} \tag{A.4}$$

avec $\mathcal{K}_{\inf}(\nu_i, \mu^\star) = \inf\left\{KL(\nu_a, \nu) : \nu \in \mathscr{P} \text{ et } \mathbb{E}[\nu] > \mu^\star\right\}$.

A.3 Prise en charge du risque

Cette section résume le premier jeu de contributions de cette thèse présenté dans le chapitre 6. Elle consiste en la prise en charge du risque en bandits manchots; le risque étant caractérisé par le tirage d'un bras aux récompenses très faibles.

A.3.1 Approche max-min

La première approche proposée est l'algorithme `MIN` tirant à chaque pas de temps un bras avec le minimum empirique maximal : $I_t \in \underset{i\in\{1\dots,K\}}{\operatorname{argmax}}\{m_{i,t}\}$ avec $m_{i,t} \overset{\text{def}}{=} \min\left\{Y_u \text{ t.q. } I_u = i, u = 1,\dots t\right\}$. Le but de cet algorithme est de trouver le bras avec l'*infimum essentiel* défini comme suit.

Définition A.3.1. *Soit ν une distribution de probabilité et $X \sim \nu$ une variable réelle. L'**infimum essentiel** a_ν de ν est défini par :*

$$a_\nu \overset{\text{def}}{=} \max_{a\in\mathbb{R}} \{\mathbb{P}(X < a) = 0\}$$

Grâce à une inégalité de concentration explicitée dans le chapitre 6 (lemme 6.6), nous pouvons, sous certaines hypothèses détaillées ci-dessous, établir une borne sur le pseudo-regret de `MIN`.

Proposition A.3.1. *Soient $\nu_1, \dots, \nu_K$ les K distributions des bras avec un support borné dans $[0,1]$. On dénote μ_i (respectivement a_i) les moyennes (respectivement les infima essentiels) des ν_i. On suppose :*

- *Pour tout $i \in \{1,\dots,K\}$ et tout $t \in \{1,\dots T\}$, il existe une constante $A > 0$ telle que pour tout $\varepsilon > 0$,*
$$\mathbb{P}\left(X_{i,t} \leqslant a_i + \varepsilon\right) \geqslant A\varepsilon. \tag{A.5}$$
- *Les bras ayant une moyenne maximale $\mu^\star = \max_{i\in\{1,\dots K\}} \{\mu_i\}$ ont aussi un infimum essentiel maximal $a^\star = \max_{i\in\{1,\dots K\}} \{a_i\}$.*

En notant $\Delta_{\mu,i} = \mu^\star$ *(respectivement* $\Delta_{a,i} = a^\star - a_i$*) la marge des moyennes (respectivement des infima essentiels), le pseudo-regret cumulé est borné, avec probabilité au moins* $1-\delta$ *par :*

$$\overline{R_t} \leqslant \frac{K-1}{A} \frac{\Delta_{\mu,\max}}{\Delta_{a,\min}} \log\left(\frac{tK}{\delta}\right) + (K-1)\,\Delta_{\mu,\max} \tag{A.6}$$

avec $\Delta_{a,\min} \overset{\text{def}}{=} \min_{i:\Delta_{a,i}>0} \Delta_{a,i}$ *et* $\Delta_{\mu,\max} \overset{\text{def}}{=} \max_{i:\Delta_{\mu,i}>0} \Delta_{\mu,i}$*. De plus, l'espérance du pseudo-regret cumulé est bornée, pour* t *assez grand (*$t \geqslant \frac{K-1}{A}\frac{\Delta_{a,\min}}{\Delta_{\mu,\max}}$*) par :*

$$\mathbb{E}[\overline{R_t}] \leqslant \frac{K-1}{A} \frac{\Delta_{\mu,\max}}{\Delta_{a,\min}} \left(\log\left(\frac{t^2 KA}{K-1}\frac{\Delta_{a,\min}}{\Delta_{\mu,\max}}\right) + 1\right) + (K-1)\,\Delta_{\mu,\max}. \tag{A.7}$$

Sous une condition supplémentaire concernant la queue gauche de la distribution, un résultat plus fort peut être obtenu :

Proposition A.3.2. *En reconsidérant les notations et hypothèses de la Proposition A.3.1, et en supposant de plus que pour tout* $i \in \{1, ldots, K\}$*,* $\Delta_{\mu,i} \leqslant \Delta_{a,i}$*, alors, avec une probabilité au moins* $1-\delta$ *:*

$$\overline{R_t} \leqslant \frac{K-1}{A} \log\left(\frac{tK}{\delta}\right) + (K-1)\Delta_{\mu,\max}$$

De plus, pour $t > \frac{K-1}{A}$*, l'espérance de* $\overline{R_t}$ *est bornée par :*

$$\mathbb{E}\left[\overline{R_t}\right] \leqslant \frac{K-1}{A} \left(\log\left(\frac{t^2 KA}{K-1}\right) + 1\right) + (K-1)\,\Delta_{\mu,\max} \tag{A.8}$$

Discussion

La comparaison de la borne obtenue avec celle de l'algorithme `UCB` (Auer et al., 2002) montre que `MIN` obtient une meilleure borne de regret lorsque :

1. les deux objectifs coïncident, i.e. lorsque les bras avec la moyenne maximale ont l'infimum essentiel maximal.
2. les marges $\Delta_{\mu,i}$ sont petites,
3. A est grand,
4. $\Delta_{a,i} \geqslant \Delta_{\mu,i}$.

Il est important de noter que la dernière condition revient à considérer que les meilleurs bras (au sens de la moyenne) ont un support plus étroit et donc un risque plus faible. En l'absence de cette hypothèse, les deux objectifs (maximisation de la moyenne et minimisation du risque au sens du min) sont en conflit.

A.3.2 Valeur à risque conditionnelle

Nous considérons dans cette partie une nouvelle mesure de risque correspondant informellement à la moyenne obtenue dans les α% pires cas, avec $\alpha \in (0,1]$ un paramètre. Cette notion intuitive fait l'objet d'une définition formelle issue de la littérature économique et définie ci-dessous.

Définitions

Définition A.3.2 (Valeur à risque)**.** *Soit X une variable aléatoire et $\alpha \in (0,1]$ une variable aléatoire réelle. La **valeur à risque (**Value at risk**)** ou **quantile** d'ordre α est définie par :*

$$VaR_\alpha \stackrel{\text{def}}{=} \inf_{\xi \in \mathbb{R}} \{P(X \leqslant \xi) \geqslant \alpha\} \tag{A.9}$$

Définition A.3.3 (Valeur à risque conditionnelle)**.** *Soit X une variable aléatoire réelle et $\alpha \in (0,1]$. La **valeur à risque conditionnelle (**Conditional Value at risk**)** d'ordre α est définie par :*

$$CVaR_\alpha \stackrel{\text{def}}{=} \inf_{\xi \in \mathbb{R}} \left\{ \xi + \frac{1}{1-\alpha} \mathbb{E}\left[(X-\xi)_+ \right] \right\} \tag{A.10}$$

avec $(x)_+$ la partie positive de x définie par :

$$(x)_+ = \begin{cases} x & x > 0 \\ 0 & x \leqslant 0 \end{cases}$$

Proposition A.3.3. *Soit X une variable aléatoire réelle. Si $P(X = VaR_\alpha(X)) = 0$ (en particulier si X est une variable continue), on a :*

$$CVaR_\alpha(X) = \mathbb{E}[X | X > VaR_\alpha(X)] \tag{A.11}$$

Remarque A.3.1. *Intuitivement, X représente une perte et $CVaR_\alpha(X)$ est l'espérance de X dans les α% pires cas. $CVaR_\alpha(X)$ est une quantité que l'on souhaite minimiser.*

Remarque A.3.2. *La dénomination de "valeur à risque conditionnelle" est justifiée par l'Équation A.11.*

Il est habituel dans la littérature bandit de considérer un critère à maximiser. Pour cela, nous nous intéressons, pour une variable X à la quantité $mCVaR(X) \stackrel{\text{def}}{=} -CVaR(-X)$ ($= \mathbb{E}[X | X < VaR(X)]$ si X est continue) que nous nommerons **valeur à risque conditionnelle modifiée**.

Estimation de la valeur à risque conditionnelle (modifiée)

Définition A.3.4. *Soit $\alpha \in (0,1]$ un niveau de confiance, et soit $x_1, \dots, x_n$ un échantillon de n réalisation i.i.d. d'une distribution ν. En supposant, sans perte de généralité, que $x_1 \leqslant x_n$, un estimateur de $mCVaR_\alpha(X)$ avec $X \sim \nu$ est défini par :*

$$\widehat{mCVaR}_\alpha(x_1, \dots, x_n) \stackrel{\text{def}}{=} \frac{1}{\lceil n\alpha \rceil} \sum_{i=1}^{\lceil n\alpha \rceil} x_i \tag{A.12}$$

avec $\lceil n \rceil$ la partie entière par excès de n.

Remarque A.3.3. *Pour $\alpha = 1$, l'estimateur est la moyenne empirique des x_i.*

Remarque A.3.4. *D'après (Chen, 2008), $\widehat{mCVaR}_\alpha(X)$ est un estimateur consistant de $mCVaR_\alpha(X)$.*

Algorithme `MARAB`

Le pseudo-code de l'algorithme `MARAB` (*Multi-Armed Risk Aware Bandit*, Galichet et al. (2013)) est décrit dans l'Algorithme 25. Nous posons $\widehat{mCVaR}_{\alpha,i}(X_{i,1}, \dots, X_{i,N_{i,t}}) \stackrel{\text{def}}{=} \widehat{mCVaR}_\alpha(X_{i,1}, \dots, X_{i,N_{i,t}})$.

Algorithm 25 `MARAB` pour K bras

Require: Horizon temporel T; niveau de risque α; paramètre d'exploration $C > 0$.

1: **for** $t = 1 \dots K$ **do**
2: $I_t = t$; récupérer Y_t; initialiser $\widehat{mCVaR}_{\alpha,t}(X_{t,1}) = X_{t,1} = Y_t$
3: **end for**
4: **for** $t = K+1 \dots T$ **do**
5: Tirer (choix arbitraire en cas d'égalité)

$$I_t \in \underset{i \in \{1,\dots,K\}}{\operatorname{argmax}} \left\{ \widehat{mCVaR}_{\alpha,i}(X_{i,1}, \dots, X_{i,N_{i,t}}) - \sqrt{\frac{C \log \lceil t\alpha \rceil}{\lceil \alpha N_{i,t} \rceil}} \right\}. \tag{A.13}$$

6: **end for**

Discussion

L'algorithme `UCB` est basé sur le principe d'optimisme face à l'inconnu en sélectionnant le bras avec uns borne de confiance supérieure sur la moyenne maximale. Ici, `MARAB` fait preuve d'un comportement prudent et pessimiste dû au terme d'exploration négatif. Ainsi, et en opposition à `UCB`, plus la valeur du paramètre C est grande et plus l'algorithme se montre conservateur. De part la définition de l'estimateur en équation A.12, l'algorithme `MARAB` se comporte en deux phases :

1. Dans une première **phase d'initialisation** ($N_{i,t} < \frac{1}{\alpha}$ et $\lceil N_{i,t} \rceil = 1$), on a :

$$\widehat{mCVaR}_{\alpha,i}(X_{i,1},\dots,X_{i,N_{i,t}}) = X_{(1)} = \min_{s\in\{1,\dots,N_{i,t}\}} \{X_{i,s}\}$$

   et la qualité du bras est évaluée à partir de la valeur minimale obtenue (et décroit donc avec le temps). La durée de cette phase est contrôlée par α et augmente lorsque α décroit vers 0.

   Dans cette phase, la maximisation de $\widehat{mCVaR}_{\alpha,i}(X_{i,1},\dots,X_{i,N_{i,t}})$ se réduit à un problème max-min et le comportement de `MARAB` se rapproche de celui de `MIN` (à la seule différence que `MIN` n'a pas de terme d'exploration négatif). Dans ces premières itérations, l'exploration est seulement dûe à la décroissance de $\widehat{mCVaR}_{\alpha,i}(X_{i,1},\dots,X_{i,N_{i,t}})$ avec $N_{i,t}$, qui peut induire la revisite de bras moins essayés.

   Cependant, la nature pessimiste de l'approche empêche la visite de bras ayant procuré de mauvaises récompenses dans les premiers essais.

2. Une seconde **phase de stabilisation**, où l'estimée $\widehat{mCVaR}_{\alpha,i}(X_{i,1},\dots,X_{i,N_{i,t}})$ est calculée avec une précision acrrue, l'erreur d'approximation convergeant vers 0 comme $\sqrt{N_{i,t}}$ (Chen, 2008). Seul les bras les plus joués entre dans cette phase ($N_{i,t} \geqslant \frac{1}{\alpha}$) et la valeur empirique de $\widehat{mCVaR}_{\alpha,i}(X_{i,1},\dots,X_{i,N_{i,t}})$ tend à se stabiliser. Notons cependant qu'il n'y a pas de garantie de visiter chaque bras un nombre infini de fois à cause du terme négatif d'exploration.

Validation expérimentale de MIN et MARAB

`MIN` et `MARAB` sont comparés à `UCB` et aux algorithmes `MVLCB` et `ExpExp` (Sani et al., 2012a), conçus pour présenter également une aversion au risque. Trois configurations sont envisagées :

- Premièrement un problème simple où `MIN` est placé dans des conditions favorables (Équation A.5 satisfaite, ordre identique sur les bras pour la moyenne et l'infimum essentiel avec $\Delta_{a,i} > \Delta_{\mu,i}$ pour tout bras i).
- 1000 problèmes aléatoirement générés satisfaisant uniquement Équation A.5.
- Un problème réel d'allocation énergétique simplifié.

On considère $K = 20$ bras pour toutes ces expériences et $T = 100K$ ou $T = 200K$. L'ensemble des figures présentant les résultats sont disponibles dans la Section 6.5.1. Les conclusions des expérimentations sont les suivantes :

- Sur un problème favorable, MIN est capable de tirer avantage de marges $\Delta_{a,i}$ supérieures à $\Delta_{\mu,i}$ et d'obtenir de meilleures performances que UCB dont les performances sont également dégradées en cas de fortes variances sur les bras sous-optimaux. MARAB est de façon intéressante capable d'avoir le même comportement que MIN pour une large plage de valeur α et avec une faible sensibilité à son paramètre d'exploration C.

- Sur les 1000 problèmes artificiels, après paramétrisation optimale de tous les algorithmes, UCB retourne les meilleurs résultats globaux. MARAB montre un regret légèrement dégradé avec une sensibilité plus faible au paramètre C. ExpExp domine MVLCB et est dominé par MARAB pour environ 70% des problèmes. De manière intéressante, ces observations sont valides sur les deux horizons temporels. De plus, en observant la distribution des récompenses instantanées récoltées par les algorithmes sur les problèmes artificiels avec une forte et faible variance, on observe une grande sensibilité au paramètre C d'UCB. MARAB et ExpExp sont eux capables d'éviter les bras risqués et MVLCB a de faibles performances. MARAB présente cependant les avantages de ne pas nécessiter la connaissance *a priori* de l'horizon T et d'être robuste vis-à-vis de sa paramétrisation (C et α). Enfin, pour $\alpha \leqslant 0.2$, les expériences montrent que MARAB obtient de meilleurs résultats en pire cas que ExpExp.

- Ces tendances globales sont reproduites sur le problème d'énergie réel avec un regret minimal pour UCB paramétré idéalement, MVLCB dominé par tous les algorithmes, MARAB (et MIN) capable d'éviter les bras risqués pour une large plage de valeurs de C.

A.3.3 Algorithme MARABOUT

Cette section présente une modification de l'algorithme MARAB basé sur un autre estimateur avec un regret contrôlé.

Définitions

Définition A.3.5. *Soit X une variable aléatoire réelle de distribution ν et soit $\alpha \in (0,1]$ un niveau de risque.*
Soit $x_1, \ldots, x_n$ un n échantillon i.i.d. de ν. En supposant, sans perte de généralité que $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$, l'estimateur de la méthode des moments de $mCVaR(X)$ est défini par :

$$\widehat{mCVaR}_\alpha(x_1, \ldots, x_n) \stackrel{\text{def}}{=} x_{(\lceil n\alpha \rceil)} + \frac{1}{n\alpha} \sum_{i=1}^{\lfloor n\alpha \rfloor} \left(x_{(i)} - x_{(\lceil n\alpha \rceil)} \right) \tag{A.14}$$

Le pseudo-code de l'algorithme MARABOUT est décrit dans l'Algorithme 26. D'une façon analogue à UCB, l'algorithme calcule pour chaque bras i une borne supérieure de confiance (pour $mCVaR_i$) et sélectionne le bras avec la borne supérieure maximale (Équation A.15). Les paramètres C et β contrôlent la force de l'exploration (plus ces paramètres sont grands et plus l'exploration est forte).
La différence clé entre MARAB et MARABOUT est le caractère optimiste de l'algorithme MARABOUT qui assure que chaque bras sera visité asymptotiquement un nombre illimité de fois.

Algorithm 26 MARABOUT pour K bras

Require: Horizon temporel T, niveau de risque α, coefficients d'exploration $C > 2$ et $\beta \in [0,1]$.

1: **for** $t = 1 \dots K$ **do**
2: $I_t = t$; récupérer Y_t; Initialiser $\widetilde{mCVaR}_{\alpha,i}(X_{i,1}) = Y_t$
3: **end for**
4: **for** $t = K+1 \dots T$ **do**
5: Tirer (choix arbitraire en cas d'égalité)

$$I_t \in \underset{i\in\{1,\dots,K\}}{\operatorname{argmax}} \left\{ \widetilde{mCVaR}_{\alpha,i}(X_{i,1},\dots,X_{i,N_{i,t}}) + \sqrt{\frac{11(C\log t + \beta \log 3)}{\alpha N_{i,t}}} \right\}. \tag{A.15}$$

6: **end for**

Proposition A.3.4 (Wang and Gao (2010))**.** *Soit X une variable aléatoire avec $Supp(X) \subset [a,b]$. Pour tout $\varepsilon > 0$*

$$\mathbb{P}\left(\widetilde{mCVaR}_\alpha(x_1,\dots,x_n) \leqslant mCVaR_\alpha(X) - \varepsilon\right) \leqslant 3\exp\left(-\frac{1}{11}\alpha\left(\frac{\varepsilon}{b-a}\right)^2 n\right) \tag{A.16}$$

$$\begin{aligned}\mathbb{P}\left(\widetilde{mCVaR}_\alpha(x_1,\dots,x_n) \geqslant mCVaR_\alpha(X) + \varepsilon\right) &\leqslant 3\exp\left(-\frac{1}{5}\alpha\left(\frac{\varepsilon}{b-a}\right)^2 n\right) \\ &\leqslant 3\exp\left(-\frac{1}{11}\alpha\left(\frac{\varepsilon}{b-a}\right)^2 n\right)\end{aligned} \tag{A.17}$$

La proposition A.3.4 permet d'établir sur le même modèle que UCB une borne pour le pseudo-regret de MARABOUT. Pour établir cette borne, il faut adapter la définition de pseudo-regret au cas du $mCVaR$ comme suit.

Définition A.3.6. *Soit un problème de bandits à K bras de distributions de récompenses ν_i et soit $\alpha \in (0,1]$ un niveau de risque.*
On note $mCVaR_i \overset{\text{def}}{=} mCVaR_\alpha(X_i)$ avec $X_i \sim \nu_i$ et $mCVaR^\star \overset{\text{def}}{=} \max_{i\in\{1,\dots K\}} mCVaR_i$. On définit alors $\Delta_{mCVaR,i} \overset{\text{def}}{=} mCVaR^\star - mCVaR_i$ la marge associée au bras i.

Le pseudo-regret mCVaR d'un algorithme au temps t est défini par :

$$\overline{R_{mCVaR,t}} \overset{\text{def}}{=} t \times mCVaR^\star - \sum_{s=1}^{t} mCVaR_{I_t}$$
$$= \sum_{i=1}^{K} \Delta_{mCVaR,i} N_{i,t}$$

Borne supérieure sur le regret

Proposition A.3.5. *Avec les mêmes notations que plus haut, supposons que les supports de ν_i soient dans $[0,1]$ pour tout i. Alors pour $C > 2$ et $\beta \in [0,1]$, le pseudo-regret de* `MARABOUT` *est borné comme suit :*

$$\mathbb{E}\left[\overline{R_{mCVaR,t}}\right] \leqslant \sum_{i:\Delta_{mCVaR,i}>0} \left\{ \frac{44(C\log(t) + \beta\log(3))}{\alpha\Delta_{mCVaR,i}} + \Delta_{mCVaR,i}\left(1 + \frac{2\times 3^{1-\beta}}{C-2}\right)\right\} \quad \text{(A.18)}$$

Remarque A.3.5. *Comme déjà dit, la force d'exploration de* `MARABOUT` *est contrôlée par C et β. Pour β grand, le facteur de $\Delta_{mCVaR,i}$ est petit, pouvant aboutir à une meilleure borne théorique. Cependant, une valeur $\beta = 0$ est un bon choix pratique, puisque les expériences montrent que l'aversion au risque est favorisée par des algorithmes au comportement conservateur.*

Validation expérimentale de `MARABOUT`

`MARABOUT` est évalué en deux temps :

1. Premièrement, son pseudo-regret mCVaR est évalué sur trois problèmes artificiels à deux bras de difficultés croissantes avec $T = 1000K$ et sur 100 exécutions indépendantes.

2. Ensuite, son comportement est comparé à celui de `MARAB` sur les 1000 problèmes générés aléatoirement et sur le problème énergétique.

L'ensemble des figures présentant les résultats de ces expériences est disponible en Section 6.6.3.

Tests synthétiques

Les trois problèmes artificiels sont caractérisés par :

- $\alpha = 0.5$, $\Delta = 0$ et $\Delta_{mCVaR} = 0.5$.
- $\alpha = 0.01$, $\Delta = 5.10^{-3}$ et $\Delta_{mCVaR} = 0.5$

- $\alpha = 0.01$, $\Delta = 10^{-3}$ et $\Delta_{mCVaR} = 0.1$.

Les résultats montrent que dans les trois configurations et dès un horizon temporel court, MARABOUT est capable d'obtenir un pseudo-regret mCVaR logarithmique, pour C petit ($C = 10^{-4}$, en dehors des valeurs admissibles pour la Proposition A.3.5) et $\beta = 0$. Ces résultats sont encourageants notamment pour les deux derniers problèmes où la petite valeur de α interdit une estimation précise de mCVaR avec peu d'échantillons mais où l'algorithme conserve un regret logarithmique même en cas de marge Δ_{mCVaR} petite.
Pour α petit, on observe une variance plus large, vraisemblablement dûe à la sensibilité de l'estimation de mCVaR lors des premières itérations. Cette variance augmente également lorsque la marge Δ_{mCVaR} diminue.

Comparaison avec MARAB

Sur les 1000 problèmes aléatoires, $\alpha = 0.2$ et $C = 10^{-7}$ (un ordre de grandeur plus petit), MARABOUT obtient des résultats légèrement dégradés mais comparables à ceux obtenus par MARAB pour $T = 100K = 2000$ et $T = 200K = 4000$. En examinant les récompenses instantanées obtenues, on observe comme précédemment une plus grande sensibilité au paramétrage dans le cas où la variance est faible, même si MARABOUT est capable d'atteindre des performances du niveau de celle de MARAB dans les deux cas pour $\alpha = 20\%$. Finalement, sur le problème d'énergie réel, on observe que MARABOUT est capable d'obtenir de bonnes récompenses en pire cas, mais souffre d'un regret significativement supérieur à MARAB. Ces meilleurs résultats de MARAB sont certainement conséquence du caractère pessimiste de l'algorithme.

A.4 Sous-échantillonage pour les bandit contextuels linéaires

Nous nous plaçons dans le cadre des bandits contextuels linéaires. Il s'agit d'un problème classique de bandit stochastique auquel est ajouté deux hypothèses supplémentaires. La première est, qu'à chaque instant t, un contexte est dévoilé au joueur et permet d'inclure plus d'information pour la prise de décision. Deuxièmement, la récompense instantanée est supposée être une fonction linéaire. On étend dans cette section l'algorithme BESA (Baransi et al., 2014), basé sur le sous-échantillonage, à ce cadre.

A.4.1 Algorithme CL-BESA

Nous introduisons maintenant formellement l'algorithme CL-BESA pour $K = 2$ bras ainsi que son cadre d'utilisation.

Notations et hypothèses

Nous notons d la dimension du problème, l'espace des contextes $\mathcal{X} \subset \mathbb{R}^d$ et un espace de paramètres $\Theta \subset \mathbb{R}^d$. Pour chaque bras i, nous faisons l'hypothèse qu'il existe un paramètre $\theta_i \in \Theta$ à apprendre et la récompense instantanée (associée au bras I_t) s'écrit alors :

$$Y_t = \langle X_t, \theta_{I_t} \rangle + \eta_t \tag{A.19}$$

avec η_t bruit additif centré. Ce modèle est appelé **modèle linéaire disjoint** et est partagé par `LinUCB` (Li et al., 2010), mais ni par `OFUL` (Abbasi-Yadkori et al., 2011) ni par `Thompson sampling` (Agrawal and Goyal, 2012a) qui eux considèrent un paramètre θ partagé par tous les bras.
Nous faisons de plus les hypothèses suivantes :

- Les contextes $X_t \in \mathcal{X}$ sont tirés indépendemment par la Nature.
- Les paramètres $\theta_i \in \Theta$ sont deux à deux indépendants.
- Θ et $\mathcal{X}$ sont supposés convexes, bornés et connus du joueur.
- Le bruit η_t est sous-gaussien, i.e., il existe une constante $R_\eta \in \mathbb{R}$ telle que, pour tout $\lambda \in \mathbb{R}$:
$$\log\left[\mathbb{E}\exp\left(\lambda\eta_t\right)\right] \leqslant \frac{\lambda^2 R_\eta^2}{2} \tag{A.20}$$
- La moyenne des contextes est notée μ et on écrit $X_t = \mu + \xi_t$, où les ξ_t sont centrés et i.i.d, bornés presque sûrement par $\|\xi_t\| \leqslant \frac{\sigma_X^2}{2}$ pour une constante σ_X^2, et tels que, pout tout $\lambda \in \mathbb{R}^d$:
$$\log\left[\mathbb{E}\exp\left(\lambda^T\xi_t\right)\right] \leqslant \frac{\|\lambda\|_2^2\sigma_X^2}{2} \tag{A.21}$$
- $\forall x \in \mathcal{X}, \forall \theta \in \Theta, |\langle x, \theta \rangle \leqslant 1$
- Le rayon de l'espace de paramètre Θ est borné par une constante B :
$$\max_{\theta\in\Theta} \|\theta\|_2 \leqslant B \tag{A.22}$$
- Toutes les distributions ont une densité par rapport à la mesure de Lebesgue.

Nous considérons maintenant un ensemble de S échantillons contexte-récompense $\mathcal{S} \overset{\text{def}}{=} \left\{\left(X_{i_1}, Y_{i_1}\right), \ldots, \left(X_{i_S}, Y_{i_S}\right)\right\}$, $\mathbf{X}(\mathcal{S}) \overset{\text{def}}{=} \left(X_{i_1}, \ldots, X_{i_S}\right)^T$ la matrice $S \times d$ de contextes et $Y(\mathcal{S}) \overset{\text{def}}{=} \left(Y_{i_1}, \ldots, Y_{i_S}\right)^T$ le vecteur de récompense de dimension S associés.

Nous voulons estimer le vecteur θ tel que $\mathbf{Y}(\mathscr{S}) = \mathbf{X}(\mathscr{S})\theta$. L'estimateur régularisé des moindres carrés $\widehat{\theta}$ est défini par :

$$\widehat{\theta_\lambda}(\mathscr{S}) \overset{\text{def}}{=} \left(\mathbf{X}(\mathscr{S})^T\mathbf{X}(\mathscr{S}) + \lambda I_d\right)^{-1}\mathbf{X}(\mathscr{S})^T\mathbf{Y}(\mathscr{S}) \tag{A.23}$$

avec I_d la matrice identité $d \times d$ et $\lambda > 0$ un paramètre de régularisation.
Finalement, nous introduisons les notations suivantes :

- $\mathscr{S}_{i,t} \overset{\text{def}}{=} \left\{(X_{t'}, Y_{t'}) : t' \leqslant t, I_{t'} = i\right\}$ est le sous-ensemble d'observations où le bras i est choisi.
- $I \sim Wr(n,m)$ dénote un ensemble aléatoire de n indices tirés uniformément sans remise sur l'ensemble $\{1,\dots,m\}$. Par convention, $I = \{1,\dots,m\}$ si $n \geqslant m$.
- En notant $\mathscr{S} = \{s_1,\dots,s_S\}$ un ensemble fini d'observations, on défini l'ensemble sous-échantillonne par rapport à I par $\mathscr{S}(I) \overset{\text{def}}{=} \{s_i, i \in I\}$.

Algorithme

L'algorithme `CL-BESA` (*Contextual Linear Best Sub-Sampled Arm*) est introduit pour le cas de $K = 2$ bras dénotés a et b (Algorithme 27).

Algorithm 27 `CL-BESA` (a,b) pour deux bras

Require: Itération courante t, contexte X_t, paramètre λ.

1: Échantillonner $I^a_{t-1} \sim \texttt{Wr}(N_{t-1}(b); N_{t-1}(a))$ et $I^b_{t-1} \sim \texttt{Wr}(N_{t-1}(a); N_{t-1}(b))$.
2: Calculer $\widehat{\theta}_{a,t-1} \overset{\text{def}}{=} \widehat{\theta}_\lambda(\mathscr{S}_{a,t-1}(I^a_{t-1}))$ et $\widehat{\theta}_{b,t-1} \overset{\text{def}}{=} \widehat{\theta}_\lambda(\mathscr{S}_{b,t-1}(I^b_{t-1}))$
3: Tirer (choix du bras le moins tirés en cas d'égalité)

$$I_t = \underset{a' \in \{a,b\}}{\operatorname{argmax}} \langle X_t, \widehat{\theta}_{a',t-1}\rangle\,. \tag{A.24}$$

Comme `BESA`, `CL-BESA` compare deux bras a et b sur la base de la même quantité d'information. Pour ce faire, il sous-échantillonne parmi le bras le plus tirés un nombre d'échantillons égal à celui du bras le moins tirés (Ligne 1) et calcule les estimateurs sur ces sous-échantillons (Ligne 2). Le bras avec la récompense instantanée espérée maximale est tiré selon l'Équation A.24.

A.4.2 Borne sur le regret contextuel

Regret contextuel

Contrairement au cas standard, le meilleur bras peut varier en fonction du contexte instantané X_t. Nous définissons alors une notion de regret contextuel de la façon

suivante :

Définition A.4.1. *Considérons un instant t et le contexte instantané X_t associé. Le meilleur bras est défini par :*

$$\theta_{t,\star} \overset{\text{def}}{=} \underset{i\in\{a,b\}}{\operatorname{argmax}} \langle X_t, \theta_i \rangle .$$

*Étant donné un horizon temporel $T \in \mathbb{N}^\star$, le **regret contextuel** est défini par :*

$$R_{X,T} = \sum_{t=1}^{T} \langle X_t, \theta_{t,\star} - \theta_{I_t} \rangle . \tag{A.25}$$

Borne

Théorème A.4.1. *Soit R_η et σ_X les paramètres de bruit sous-gaussiens repectivement définis par les Équations A.20 et A.21 et soit B défini par Équation A.22. Soit λ un paramètre croissant (Éq. A.23) tel que $\lambda \geqslant 6\sigma_X^2 \log(T)$. Supposons que*

$$\left|\langle \mu, \theta_a - \theta_b \rangle\right| \geqslant 8\sigma_X B + 2\sqrt{2}\frac{\|\theta_a\|_2 + \|\theta_b\|_2}{\sqrt{\lambda^{-1} + \|\mu\|_2^{-2}}} . \tag{A.26}$$

Alors, le regret contextuel de `CL-BESA` après T pas de temps est borné par :

$$\begin{aligned}\mathbb{E}[R_{X,T}] \leqslant & \left(\max_{t\in[T]} \Delta_t\right) \frac{64}{\min_{t\in[T]} \Delta_t^2} \left[R_\eta \sqrt{2d \log\left(\lambda^{1/2} T^2 + \frac{T^3 (\|\mu\|_2 + \sigma_X)^2}{d\lambda^{1/2}}\right)} + \lambda^{1/2} B \right]^2 \\ & + \left(\max_{t\in[T]} \Delta_t\right) \frac{24\sigma_X^2 \log(T) - 2\lambda}{\|\mu\|_2^2} + \sum_{t=1}^{T} \Delta_t \mathbb{I}\{\min_{t\in[T]} \Delta_t \leqslant \tau\} \Bigg) + O(1) .\end{aligned}$$

où

$$\tau \overset{\text{def}}{=} 2\sigma_X \left[R_\eta \sqrt{\frac{2d}{\lambda} \log\left(\lambda^{1/2} T^2 + \frac{T^3 (\|\mu\|_2 + \sigma_X)^2}{d\lambda^{1/2}}\right)} + B \right] . \tag{A.27}$$

Quand la perturbation sur le contexte $\sigma_X = 0$, alors Δ_t devient $\Delta \overset{\text{def}}{=} \left|\langle \mu, \theta_a - \theta_b \rangle\right|$ et nous obtenons

$$\mathbb{E}[R_{X,T}] \leqslant \frac{128 R_\eta^2 d}{\Delta} \log(2T^3 \|\mu\|_2^2 / d) + O(1).$$

A.4.3 Validation expérimentale

Cette section présente les résultats numériques de l'approche `CL-BESA` et la compare à d'autres approches de l'état de l'art.

A.4.4 Cadre expérimental

Nous comparons CL-BESA à trois algorithmes : Thompson sampling, OFUL et LinUCB. LinUCB et CL-BESA partagent le même modèle et LinUCB peut donc être appliqué tel que dans (Li et al., 2010). À l'inverse, OFUL et Thompson sampling font l'hypothèse d'un paramètre β partagé entre tous les bras et doivent donc être adaptés pour pouvoir être appliqué à notre modèle. Une façon simple est de concaténer les deux paramètres inconnus θ_a et θ_b en un seul paramètre à apprendre $\overline{\theta} \stackrel{\text{def}}{=} (\theta_a \theta_b) \in \mathbb{R}^{dK}$ et d'y associer les vecteurs de contexte $\theta_{t,a} \stackrel{\text{def}}{=} (X_t^T 0_d^T)^T \in \mathbb{R}^{dK}$ et $\theta_{t,b} \stackrel{\text{def}}{=} (0_d^T X_t^T)^T$. Cette adaptation permet une comparaison juste des algorithmes car OFUL et Thompson sampling ne pourront pas tirer parti d'information partagée entre les bras.

A.4.5 Résultats

L'objet de cette section est de résumer les résultats numériques obtenus. L'ensemble des figures peut être consulté dans le chapitre 7.

Problème orthogonal

Les algorithmes sont évalués sur le problème suivant avec $K = 2$ bras définis par $\mu = (0.5, 0.5)^T$, $\theta_a = (0.5, 0)^T$ et $\theta_b = (0, 0.5 + 2\Delta)^T$. La marge Δ est fixée à $\Delta = 10^{-1}$ et $T = 1000$. À chaque pas de temps, un context X_t est tiré uniformément dans la boule $\mathscr{B}(\mu, \frac{\sigma_X^2}{2})$ et la récompense est donnée par $Y_t = \langle X_t, \theta_{I_t} \rangle + \eta_t$ avec $\eta_t \sim \mathcal{N}(0, R_\eta^2)$

Sensibilité au paramètre

Un avantage majeur en terme d'applicabilité et d'implémentation de CL-BESA est le fait qu'il ne nécessite qu'un seul paramètre. De plus, les expériences montrent que le regret contextuel est peu sensible à la valeur donnée à ce paramètre. Cette robustesse est à mettre en comparaison avec, d'un part le grand nombre de paramètres et/ou la sensibilité des autres algorithmes à leur paramétrage. Par exemple, il est montré qu'OFUL est sensible aux paramètres R_{OFUL} et S_{OFUL} respectivement idéalement fixés aux valeurs (en pratique inconnues) $R_{OFUL} = R_\eta$ et $S_{OFUL} = \|\overline{\theta}\|_2$.

Sensibilité au bruit et à la perturbation du contexte

CL-BESA, OFUL, LinUCB et Thompson sampling sont testés sur le problème orthogonal avec différent niveaux de bruit additif $R_\eta \in \{\Delta, 10\Delta\}$ et perturbation de contexte $\frac{\sigma_X}{\sqrt{2}} \in \{\frac{\Delta}{10}, \Delta, 10\Delta, 100\Delta\}$. OFUL et Thompson sampling ont accès aux valeurs optimales de R et S.

Les figures confirment les bonnes performances de CL-BESA dans toutes les configurations. De manière plus précise, son regret reste contrôlé, logarithmique et quasi-optimal : il est classé deuxième ou premier pour toutes les configurations, et reste proche de la meilleure option lorsque classé second. Cette propriété est unique parmi les algorithmes étudiés et fait de CL-BESA un bon choix en pratique en cas de régime de bruit et/ou de perturbation sur le contexte inconnus.
Les figures permettent également une discussion sur la stabilité des divers algorithmes. On remarque pour commencer que les courbes de regret présentent, de manière logique, une variance supérieure lorsque R_η ou σ_X augmentent, avec une sensibilité plus forte à R_η. Cependant, on peut observer dans ce dernier cas une sensibilité plus forte pour CL-BESA que pour les autres algorithmes. Dans ces conditions extrêmes et peu probables (bruit 100 fois supérieur à la marge), un autre algorithme que CL-BESA comme OFUL peut être envisagé et fournir de meilleurs résultats. Il faut cependant souligner que OFUL bénéficie alors de l'avantage de la connaissance exacte de R_η.

Influence de la dimension

Nous étudions finalement la sensibilité des algorithmes vis-à-vis de la dimension d variant de 2 à 200. Pour $d \geqslant 2$, on considère le problème défini par $K = 2$, $\mu = (\frac{1}{\sqrt{2d}}, \dots, \frac{1}{\sqrt{2d}})^T, \theta_{a,d} = (\sqrt{2/d}\Delta, 0 \dots, \sqrt{2/d}\Delta, 0)^T$ et $\theta_{b,d} = (0, -\sqrt{2/d}\Delta, \dots, 0, -\sqrt{2/d}\Delta)^T$. La marge est alors Δ.
LinUCB souffre d'instabilités numériques et ne peut être étudiés pour $d \geqslant 40$. Thompson sampling et CL-BESA obtiennent un regret stable et celui d'OFUL croit légèrement avec la dimension, celui de CL-BESA étant significativement le plus faible des quatre algorithmes étudiés. Ce résultat s'explique en partie par la paramétrage facile de CL-BESA qui lui permet d'avoir un bon comportement pour toutes les dimensions.

A.5 Conclusion et perspectives

Les contributions de cette thèse sont :

1. L'algorithme MIN et son étude théorique et pratique.

2. Les algorithmes MARAB et MARABOUT ainsi que l'étude comparative de ces deux approches sur problèmes artificiels et réels.

3. La borne sur le regret de MARABOUT.

4. L'introduction du sous-échantillonnage pour les bandits contextuels et de l'algorithme CL-BESA.

5. L'étude théorique du regret contextuel de CL-BESA ainsi que l'étude comparative numérique de l'algorithme sur différents niveaux de bruits et différents niveaux de perturbation du contexte.

Les possibilités d'extensions du travail présenté sont multiples, on pourrait notamment penser à :

1. Un raffinement de la preuve sur le regret de MARABOUT avec des techniques d'auto-normalisation ou de pelage.

2. Une extension de la risque aversion au cas de décision séquentiel d'une manière arborescente, en s'inspirant par exemple de l'extension d'UCB à UCT(Kocsis and Szepesvári, 2006a) ou d'OLOP (Bubeck and Munos, 2010).

3. Un algorithme de bandit contextuel avec aversion au risque serait intéressant car le besoin d'une exploration prudente demeure importante dans ce cadre.

4. L'extension de CL-BESA au cas du bandit contextuel non linéaire.

Bibliography

Automl workshop, 2015.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2312–2320. Curran Associates, Inc., 2011.

Carlo Acerbi and Dirk Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487 – 1503, 2002. ISSN 0378-4266.

Deepak Agarwal, Bee-Chung Chen, and Pradheep Elango. Explore/exploit schemes for web content optimization. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, ICDM '09, pages 1–10, Washington, DC, USA, 2009. IEEE Computer Society. ISBN 978-0-7695-3895-2. doi: 10.1109/ICDM.2009.52.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. *CoRR*, abs/1209.3352, 2012a.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, pages 39.1–39.26, 2012b.

Philippe Artzner, Freddy Delbaen, jean-Marc Eber, and David Heath. Thinking coherently, 1997. ISSN 0952-8776.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2785–2836, 2010.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theor. Comput. Sci.*, 410 (19):1876–1902, April 2009. ISSN 0304-3975. doi: 10.1016/j.tcs.2009.01.016.

J.Y. Audibert, S. Bubeck, and R. Munos. Bandit view on noisy optimization. In *Optimization for Machine Learning*. MIT Press, 2010.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3:397–422, March 2003. ISSN 1532-4435.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-arm bandit problem. In *36th Annual Symposium on Foundations of Computer Science, Milwaukee, Wisconsin, 23-25 October 1995*, pages 322–331, 1995. doi: 10.1109/SFCS.1995.492488.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Mach. Learn.*, 47(2-3):235–256, May 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352.

Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mechanisms. In *ACM Conference on Electronic Commerce (EC'09)*. Association for Computing Machinery, Inc., July 2009. Full version can be found on arxiv.org (http://arxiv.org/abs/0812.2291).

Akram Baransi, Odalric-Ambrym Maillard, and Shie Mannor. Sub-sampling for multi-armed bandits. In *Machine Learning and Knowledge Discovery in Databases*, page 115–131. Springer, 2014.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. ISBN 0521833787.

David B. Brown. Large deviations bounds for estimating conditional value-at-risk. *Oper. Res. Lett.*, 35(6):722–730, 2007.

Cameron Browne, Edward J. Powley, Daniel Whitehouse, Simon M. Lucas, Peter I. Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *IEEE Trans. Comput. Intellig. and AI in Games*, 4(1):1–43, 2012.

Sébastien Bubeck. *Bandits Games and Clustering Foundations*. Theses, Université des Sciences et Technologie de Lille - Lille I, June 2010.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012. doi: 10.1561/2200000024.

Sébastien Bubeck and Rémi Munos. Open loop optimistic planning. In *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 477–489, 2010.

Apostolos N. Burnetas and Michael N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Adv. Appl. Math.*, 17(2):122–142, June 1996. ISSN 0196-8858. doi: 10.1006/aama.1996.0007.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation. *Annals of Statistics*, 41(3):1516–1541, 2013. Accepted, to appear in Annals of Statistics.

Nicolò Cesa-Bianchi and Paul Fischer. Finite-time regret bounds for the multiarmed bandit problem. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconsin, USA, July 24-27, 1998*, pages 100–108, 1998.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2249–2257. Curran Associates, Inc., 2011.

Song Xi Chen. Nonparametric estimation of expected shortfall. *Journal of Financial Econometrics*, 6(1):87–107, 2008.

Cheng-Wei Chou, Ping-Chiang Chou, Jean-Joseph Christophe, Adrien Couëtoux, Pierre de Freminville, Nicolas Galichet, Chang-Shing Lee, Jialin Liu, David Lupien Saint-Pierre, Michèle Sebag, Olivier Teytaud, Mei-Hui Wang, Li-Wen Wu, and Shi-Jim Yen. Strategic choices in optimization. *J. Inf. Sci. Eng.*, 30(3):727–747, 2014.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pages 208–214, 2011.

Pierre-Arnaud Coquelin and Rémi Munos. Bandit Algorithms for Tree Search. In *Uncertainty in Artificial Intelligence*, Vancouver, Canada, 2007.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pages 355–366, 2008.

V. H. de la Peña, T. L. , and Q. M. Shao. *Self-Normalized Processes: Limit Theory and Statistical Applications*. Springer Series in Probability and its Applications. Springer, New York, first edition, January 2009. ISBN 3540856358.

Victor H. de la Peña, Michael J. Klass, and Tze Leung. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Ann. Probab.*, 32(3):1902–1933, 07 2004. doi: 10.1214/009117904000000397.

Eyal Even-dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *In Fifteenth Annual Conference on Computational Learning Theory (COLT*, pages 255–270, 2002.

Joan Fruitet, Alexandra Carpentier, Rémi Munos, and Maureen Clerc. Bandit Algorithms boost Brain Computer Interfaces for motor-task selection of a brain-controlled button. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, pages 458–466, Lake Tahoe, Nevada, United States, 2012. Neural Information Processing Systems (NIPS) Foundation.

Matteo Gagliolo and Jürgen Schmidhuber. Algorithm selection as a bandit problem with unbounded losses. In *Learning and Intelligent Optimization, 4th International Conference, LION 4, Venice, Italy, January 18-22, 2010. Selected Papers*, pages 82–96, 2010. doi: 10.1007/978-3-642-13800-3_7.

Nicolas Galichet, Michèle Sebag, and Olivier Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning, ACML 2013, Canberra, ACT, Australia, November 13-15, 2013*, pages 245–260, 2013.

Aurélien Garivier. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *In Proceedings of COLT*, 2011.

Romaric Gaudel and Michèle Sebag. Feature selection as a one-player game. In Johannes Fürnkranz and Thorsten Joachims, editors, *ICML*, pages 359–366. Omnipress, 2010. ISBN 978-1-60558-907-7.

Sylvain Gelly, Yizao Wang, Rémi Munos, and Olivier Teytaud. Modification of UCT with Patterns in Monte-Carlo Go. Research Report RR-6062, 2006.

Sylvain Gelly, Levente Kocsis, Marc Schoenauer, Michèle Sebag, David Silver, Csaba Szepesvári, and Olivier Teytaud. The grand challenge of Computer Go: Monte-Carlo tree search and extensions. *Commun. ACM*, 55(3):106–113, March 2012. ISSN 0001-0782. doi: 10.1145/2093548.2093574.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 757–765, 2014.

J. Gittins, K. Glazebrook, and R. Weber. *Multi-arned bandit allocation indices*. Wiley and sons, 2011.

Sudipto Guha, Kamesh Munagala, and Peng Shi. Approximation algorithms for restless bandit problems. *J. ACM*, 58(1):3, 2010.

Ryan Hayward. Mohex wins hex tournament. *Int. Comp. Games Assoc. J*, pages 114–116, 2009.

Leigh R. Hochberg, Daniel Bacher, Beata Jarosiewicz, Nicolas Y. Masse, John D. Simeral, Joern Vogel, Sami Haddadin, Jie Liu, Sydney S. Cash, Patrick van der Smagt, and John P. Donoghue. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, 485:372–377, 2012. doi: 10.1038/nature11076.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *COLT'10*, pages 67–79, 2010.

Junya Honda and Akimichi Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85(3):361–391, 2011. ISSN 0885-6125. doi: 10.1007/s10994-011-5257-4.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory - 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings*, pages 199–213, 2012. doi: 10.1007/978-3-642-34106-9_18.

Levente Kocsis and Csaba Szepesvári. Universal parameter optimisation in games based on SPSA. *Machine Learning*, 63(3):249–286, 2006a. doi: 10.1007/s10994-006-6888-8.

Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *Machine Learning: ECML 2006, 17th European Conference on Machine Learning, Berlin, Germany, September 18-22, 2006, Proceedings*, pages 282–293, 2006b. doi: 10.1007/11871842_29.

Pushmeet Kohli, Mahyar Salek, and Greg Stoddard. A fast bandit algorithm for recommendation to users with heterogenous tastes. In *Proceedings of the Twenty-Seventh*

AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA., 2013.

S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22 (1):79–86, 1951.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6(1):4–22, March 1985. ISSN 0196-8858. doi: 10.1016/0196-8858(85)90002-8.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 661–670, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772758.

Odalric-Ambrym Maillard. Robust risk-averse stochastic multi-armed bandits. In Sanjay Jain, Rémi Munos, Frank Stephan, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, volume 8139 of *Lecture Notes in Computer Science*, pages 218–233. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-40934-9.

Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Conference On Learning Theory*, 2011.

Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:2004, 2004.

Harry Markowitz. Portfolio Selection. *The Journal of Finance*, 7(1):77–91, March 1952. ISSN 00221082. doi: 10.2307/2975974.

Teodor Mihai Moldovan and Pieter Abbeel. Safe exploration in markov decision processes. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

James Neufeld, András György, Csaba Szepesvári, and Dale Schuurmans. Adaptive monte carlo via bandit allocation. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1944–1952, 2014.

Stefano Nolfi and Dario Floreano. *Evolutionary Robotics: The Biology,Intelligence,and Technology*. MIT Press, Cambridge, MA, USA, 2000. ISBN 0262140705.

Sandeep Pandey and Christopher Olston. Handling advertisements of unknown quality in search advertising. In *Advances in Neural Information Processing Systems*

19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006, pages 1065–1072, 2006.

GeorgCh. Pflug. Some remarks on the value-at-risk and the conditional value-at-risk. In StanislavP. Uryasev, editor, *Probabilistic Constrained Optimization*, volume 49 of *Nonconvex Optimization and Its Applications*, pages 272–281. Springer US, 2000. ISBN 978-1-4419-4840-3. doi: 10.1007/978-1-4757-3150-7_15.

Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 784–791, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390255.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

R. Tyrrell Rockafellar. Coherent approaches to risk in optimization under uncertainty. In *In Tutorials in Operations Research INFORMS*, pages 38–61, 2007.

R. Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, pages 1443–1471, 2002.

Philippe Rolet, Michèle Sebag, and Olivier Teytaud. Boosting active learning to optimality: A tractable monte-carlo, billiard-based algorithm. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II*, pages 302–317, 2009. doi: 10.1007/978-3-642-04174-7_20.

Benjamin Van Roy and Zheng Wen. Generalization and exploration via randomized value functions. *CoRR*, abs/1402.0635, 2014.

Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Math. Oper. Res.*, 35(2):395–411, May 2010. ISSN 0364-765X. doi: 10.1287/moor.1100.0446.

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 3284–3292, 2012a.

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-Aversion in Multi-armed Bandits. Research report, November 2012b.

Gilles Stoltz, Sébastien Bubeck, and Rémi Munos. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, April 2011. doi: 10.1016/j.tcs.2010.12.059.

Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.

Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010. doi: 10.2200/S00268ED1V01Y201005AIM009.

W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Bulletin of the American Mathematics Society*, 25:285–294, 1933.

W.R. Thompson. On the theory of apportionment. *American Journal of Mathematics*, 57:450–456, 1935.

Joannès Vermorel and Mehryar Mohri. Multi-armed bandit algorithms and empirical evaluation. In *In European Conference on Machine Learning*, pages 437–448. Springer, 2005.

Thomas J. Walsh, Istvan Szita, Carlos Diuk, and Michael L. Littman. Exploring compact reinforcement-learning representations with linear regression. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, pages 591–598, 2009.

Ying Wang and Fuqing Gao. Deviation inequalities for an estimator of the conditional value-at-risk. *Oper. Res. Lett.*, 38(3):236–239, 2010. doi: 10.1016/j.orl.2009.11.008.

Christopher John Cornish Hellaby Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, UK, May 1989.

Jia Yuan Yu and Evdokia Nikolova. Sample complexity of risk-averse bandit-arm selection. In Francesca Rossi, editor, *IJCAI*. IJCAI/AAAI, 2013. ISBN 978-1-57735-633-2.