Sequential decision problems in non-stationary environments

Yoan Russac

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Ecole Normale Supérieure

Sequential Decision Problems in Non-Stationary Environments

Soutenue par
Yoan RUSSAC
Le 01/03/2022

École doctorale n°386
Sciences mathématiques de Paris centre

Spécialité
Informatique

Composition du jury :

Florence d'Alché-Buc Professeure, Télécom Paris	*Présidente*
Junya Honda Associate Professor, Kyoto University	*Rapporteur*
Sébastien Gadat Professeur, TSE	*Rapporteur*
Alessandro Lazaric Research Scientist, Meta AI	*Examinateur*
Michal Valko Research Scientist, Deepmind	*Examinateur*
Arnak Dalalyan Professeur, ENSAE	*Examinateur*
Olivier Cappé Directeur de recherche, CNRS	*Directeur de thèse*
Aurélien Garivier Professeur, ENS Lyon	*Directeur de thèse*

1794 ENS | PSL

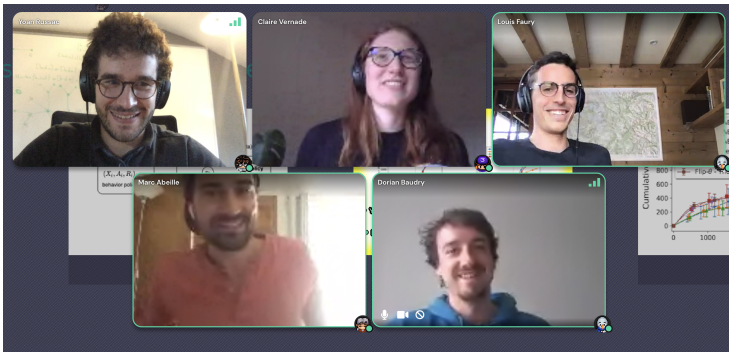
Remerciements

Il est temps pour moi d'écrire les remerciements pour cette thèse. Nombreuses sont les personnes qui ont été indispensables à mes trois années de recherche et à la rédaction de ce manuscrit. Je pense dans un premier temps à mes directeurs de thèse, Aurélien et Olivier. Aurélien, merci pour tes précieux conseils, et pour nos sessions au tableau noir, qui auront toujours été d'une grande utilité et qui m'ont marqué. Je regrette de ne pas avoir pu venir plus souvent à l'ENS à Lyon, malheureusement la situation sanitaire ne l'a pas permis. Olivier, merci pour ton encadrement pendant ces trois années. Je me sens très privilégié d'avoir eu tes nombreux conseils. Je suis ravi d'avoir eu un aussi bon exemple de chercheur et je garderais un très bon souvenir de nos nombreux échanges sur bien des sujets différents.

I would also like to express my sincere thanks to all the members of the jury. Je remercie Florence d'avoir accepté de présider mon jury de thèse. Arnak, Michal et Alessandro, c'est un honneur pour moi d'avoir pu compter sur votre présence pour ma soutenance et d'avoir eu vos retours sur mes travaux. J'ai eu la chance d'avoir chacun d'entre vous comme professeur durant mes années de Master, et vous avez tous joué un rôle dans ma décision de poursuivre en thèse. Sébastien, merci d'avoir pris le temps de relire mon long manuscrit et merci aussi pour ton rapport très détaillé malgré une contrainte temporelle. Junya thank you for accepting reviewing my manuscript. After being passionate about some of your work, it is an honour to receive some feedback from you on my own research.

Dans un second temps, j'aimerais remercier mes co-auteurs avec qui j'ai eu la grande joie de chercher des solutions à des problèmes complexes. Pour commencer, Louis pour qui les GLM n'ont plus de secret. Nos nombreux échanges et collaborations pendant le confinement et tout au long de la thèse auront été d'une aide précieuse. Merci à Marc pour ton accueil à Criteo et les sessions au tableau avec Louis. Je remercie aussi Dorian, que j'ai initialement encouragé à poursuivre en thèse sur les bandits et avec qui j'aurais eu un grand plaisir à approfondir les techniques de sous-échantillonnage. Je suis très heureux d'avoir embarqué un ami dans cette aventure qu'est la thèse! J'aimerais aussi remercier tout particulièrement Claire grâce à qui j'ai rencontré Olivier et avec laquelle j'ai rédigé mon premier article. Je n'aurais probablement pas choisi ce sujet de thèse sans ton précieux encadrement pendant mon stage du MVA, merci de m'avoir initié aux problématiques de bandits non-stationnaires. Merci aussi à Émilie de m'avoir accueilli à plusieurs reprises à Lille et pour nos échanges au cours de la thèse.

Je n'aurais malheureusement pas eu l'occasion de participer à de nombreuses conférences en présentiel mais j'ai pu découvrir les joies des conférences en ligne. A titre d'illustration, voici ci-dessous une photo d'un sous ensemble des personnes avec lesquelles j'ai eu la chance de collaborer pendant cette période en guise de remerciement!



Merci à l'ENS de m'avoir donné un cadre si exceptionnel pour préparer ma thèse. Merci au personnel administratif et en particulier à Lise Marie. Merci à Pierre Senellart et aux membres de l'équipe VALDA. Merci aussi aux membres de SCOOL pour leur accueil à Lille (Edouard, Julien, Omar, Sarah, Nathan, Xuedong, Mathieu). Merci aux autres doctorants avec lesquels j'ai eu plaisir à discuter, notamment avec Jean, Yann, Mathieu, Garance et Grégoire (notre quasi doctorant).

Je pense aussi à l'ensemble de mes amis qui m'ont soutenu pendant cette période un peu particulière. A Juliette, ma co-bureau, avec qui j'aurais partagé la plupart des péripéties de cette thèse. Je me souviens des soirées de deadline dans le grenier à l'ENS où l'on rédigeait chacun nos articles. Merci pour ta présence pendant ces trois années. A BSN et Gip qui auront suivi au plus près certains résultats de conférence après une semaine de randonnée et avec qui j'ai la chance de toujours avoir des discussions passionnantes. Je pense aussi aux "membres du ghet" (MBB, François, Antho, Sacha, Emilien) qui comprennent de loin le sujet de cette thèse et avec lesquels j'ai toujours plaisir à discuter. En particulier je tiens à remercier Pierrot qui m'aura hébergé pendant mon escale londonienne. Je pense aussi à mes ami(e)s de plus longue date (Dum, Anne-So, Clémouille, Dadoule, Delphine, Clarisse et Anne-So), pour qui le sujet de cette thèse est probablement plus obscur. A mes amis du club oeno, les floquettiers, avec qui j'ai toujours plaisir à refaire le monde (Baldovino, Servane, Aliette, Martin, Antoine, Olivia).

Pour finir, j'aimerais remercier ma famille et le soutien constant qu'elle m'a apporté tout au long de mes études. Mami, merci de m'avoir transmis ton goût pour les maths et de toujours t'être intéressée à mes sujets de recherche. Papi merci de m'avoir apporté le goût de l'apprentissage permanent. Ta source intarissable de savoir est une grande inspiration pour moi. Mes frères et sœurs, merci de m'avoir soutenu pendant ces années avec autant de bienveillance. Merci à mes parents de m'avoir encouragé à faire ce qui me plaisait et de m'avoir soutenu continuellement tout au long de ce chemin. Ce manuscrit n'aurait probablement pas vu le jour sans vous. Merci à Barbara, pour ses nombreux conseils durant ces trois années. Finalement, j'aimerais remercier mon épouse Pauline avec qui j'ai eu le bonheur de partager mon quotidien pendant ces 3 années et qui a toujours su m'écouter et me conseiller.

Abstract

The vanilla bandit model assumes that the rewards are independent and identically distributed. However, this assumption is restrictive: it prevents from modeling evolving behaviors that are common in real-world applications. In the medical domain, the efficiency of a treatment is likely to diminish over time. The opening rate of news articles fades for aging news. Fashion trends and consumers preferences evolve rapidly. Any recommender system ignoring the non-stationarity of the distributions of rewards is likely to make suboptimal choices. The objective of this thesis is the study of stochastic bandit algorithms in non-stationary environments. There are several ways to include non-stationarity into bandit models. We first study a variant of the best arm identification problem where the learner seeks to identify the set of arms that are better than a control arm in the presence of subpopulations. Those subpopulations can encode a temporal information (e.g. day of the week) and properly using them makes it possible to include non-stationarity in the pure exploration setting. We characterize the complexity of this learning task and propose optimal algorithms for solving it. We then propose theoretically grounded algorithms for minimizing the regret and discuss the exploration-exploitation trade-off the learner is facing in dynamically changing environments. Our findings concern three different settings: the well-known multi-armed bandit, the more general linear bandit but also generalized linear bandit. For each of those settings, we identify the technical challenges brought by non-stationarity.

Keywords: Sequential learning, bandit algorithms, non-stationary environments, regret minimization.

Resumé

La version classique du modèle de bandit suppose que les distributions de probabilité des récompenses sont indépendantes et identiquement distribuées. Pour autant, cette hypothèse est restrictive dans de nombreux cas, puisqu'elle ne permet pas de prendre en compte d'éventuels changements de comportements. Dans le domaine médical, l'efficacité d'un traitement peut diminuer au cours du temps. Pour un site internet d'information en temps réel, le taux de consultation d'une page diminue à raison de sa date d'ancienneté. Les tendances de mode et les préférences des consommateurs évoluent rapidement. Un algorithme de recommendation ignorant ces formes de non-stationarité est alors susceptible de faire des suggestions sous-optimales. Ainsi, l'objet de cette thèse est l'étude des algorithmes de bandits stochastiques dans des environnements non-stationnaires. La non-stationarité peut être incorporée de plusieurs manières dans les modèles de bandits. Dans un premier temps, nous étudions une variante du problème d'identification du meilleur bras. Cette variante correspond à un système d'apprentissage qui cherche à identifier l'ensemble des options qui sont meilleures qu'un bras de contrôle, et ce en présence de sous-populations. Entre autres, l'utilisation de sous-populations permet la modélisation de l'évolution temporelle des différents bras. Nous proposons ensuite des algorithmes avec des garanties théorique fortes pour la minimisation du regret et étudions le compromis exploration-exploitation pour de tels environnements. Nos recherches portent sur trois modèles différents: le bandit classique multi-bras, le bandit linéaire ou encore le bandit linéaire généralisé. Nous examinons les spécificités de chacun de ces trois modèles, ainsi que les défis techniques liés à la non-stationarité.

Mots clés : Apprentissage séquentiel, algorithmes de bandits, environnements non stationnaires, minimisation du regret.

Contents

Contributions and Thesis Outline

The organization of this thesis is the following:

Chapter 1 is an introduction to the settings that will be considered in this thesis and an overview of the different contributions. The stochastic multi-armed bandit model is first discussed with known upper and lower-bounds for quantifying the regret. Understanding this setting is useful for Chapter 3. We then describe the best arm identification task which is the starting point for Chapter 2. After that, we extend the multi-armed bandit framework and present the linear bandit model where contextual information can be used. Linear bandits are the point of departure for Chapter 4. Finally, in Chapters 5 and 6, we consider generalized linear bandits that consider richer reward models. In all of those settings, we present the most common forms of non-stationarity and how they impact the learning.

Chapter 2 presents our contribution to a *pure exploration* task where the objective is to identify all the arms that are better than a control arm in the presence of subpopulations (ABC-S). We discuss how the complexity of the ABC-S problem varies with the level of interaction with the subpopulations. We design strategies that are asymptotically optimal in the following sense: if τ_δ is the first time when the strategy is able to output the correct answer with probability at least $1-\delta$, then $\mathbb{E}[\tau_\delta]$ grows linearly with $\log(1/\delta)$ at the exact optimal rate. This rate is identified in three different settings: (1) when the experimenter does not observe the subpopulation information, (2) when the subpopulation of each sample is observed but not chosen, and (3) when the experimenter can select the subpopulation from which each response is sampled.

Chapter 3 first describes a non-parametric and asymptotically optimal algorithm LB-SDA based on subsampling for the vanilla bandit model. Starting from this chapter, the objective is to maximize the expected sum of rewards the agent collects while interacting with the bandits. In this chapter, we present a new technique for using a limited memory for the observations of each arm. We prove that storing $\Omega((\log T)^2)$ observations instead of T is enough to ensure asymptotic optimality in stationary environments. For non-stationary environments, we propose SW-LB-SDA, an adaptation of LB-SDA where subsampling is allowed only within a sliding window. We establish the minimax optimality of the approach and obtain guarantees in more general non-stationary settings where no existing algorithms were previously analyzed.

Chapter 4 gathers new results for stochastic linear bandits, a setting where the rewards follow a non-stationary linear regression model. In those environments, the unknown regression parameter is allowed to vary in time. We propose D-LinUCB a novel optimistic algorithm based on discounted linear regression where exponential weights are used to smoothly forget the past. We study the deviations of the sequentially weighted least-squares estimator under generic

assumption, and we provide theoretical guarantees for D-LinUCB in both slowly-varying and abruptly changing environments.

Chapter 5 considers a generalization of the stochastic linear bandit known as generalized linear bandit. In this setting, a layer of non-linearity is added on top of the linear regression model. In abruptly changing environments, we obtain novel confidence-based algorithm for the maximum-likelihood estimator with forgetting that better captures the non-linearity of the environment. We propose SC-D-GLUCB and obtain improved regret guarantees in abruptly changing environments.

Chapter 6 also considers generalized linear bandits but in different non-stationary environments. More precisely, we consider drifting environments where the level of non-stationarity is characterized by a general metric known as the variation budget. We uncover important mistakes in existing analyses for generalized linear bandits in this setting and we propose BVD-GLM-UCB the first algorithm with regret guarantees in those specific non-stationary environments.

Publications

The contributions from this thesis are the results of the work done under the supervision of my two PhD advisors Olivier Cappé and Aurélien Garivier, but also with other PhD students: Louis Faury and Dorian Baudry, and other researchers: Claire Vernade, Marc Abeille, Wouter M. Koolen, Clément Calauzènes, Christina Katsimerou.

List of peer-reviewed publications included in this thesis

- Russac, Y., Vernade, C., and Cappé, O. (2019). Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems*, 32:12040–12049
- Russac, Y., Faury, L., Cappé, O., and Garivier, A. (2021a). Self-concordant analysis of generalized linear bandits with forgetting. In *International Conference on Artificial Intelligence and Statistics*, pages 658–666. PMLR
- Faury, L., Russac, Y., Abeille, M., and Calauzènes, C. (2021b). A technical note on non-stationary parametric bandits: Existing mistakes and preliminary solutions. In *Algorithmic Learning Theory*
- Baudry, D., Russac, Y., and Cappé, O. (2021). On limited-memory subsampling strategies for bandits. In *International Conference on Machine Learning*, pages 727–737. PMLR
- Russac, Y., Katsimerou, C., Bohle, D., Cappé, O., Garivier, A., and Koolen, W. M. (2021b). A/B/n testing with control in the presence of subpopulations. *Advances in Neural Information Processing Systems*, 34

List of preprints

- Russac, Y., Cappé, O., and Garivier, A. (2020). Algorithms for non-stationary generalized linear bandits. *arXiv preprint arXiv:2003.10113*
- Faury, L., Russac, Y., Abeille, M., and Calauzènes, C. (2021a). Regret bounds for generalized linear bandits under parameter drift. *arXiv preprint arXiv:2103.05750*

List of Figures

List of Tables

1 | Introduction

Outline

1.1 Context of the Thesis

1.1.1 General Overview

In the past twenty years, we have seen the rise of Internet services such as online retailing, online advertising, video streaming, social networks. In this context, being able to recommend the most interesting products among a tremendous number of possibilities for a given customer is a crucial matter.

Multi-armed bandit algorithms have become a go-to paradigm for handling the "explore-exploit" dilemma for **sequential learning** tasks under uncertainty: after interacting with some users, the learning system needs to decide whether it commits to the seemingly best performing option (exploitation) or continues exploring for discovering potentially better candidates (exploration). In its vanilla formulation, the stochastic multi-armed bandit assumes that several unknown probability distributions are available. This is a **partial information** setting where the learner repeatedly picks an action among the different alternatives and only observes the reward associated with his actions. By interacting with the environment, the learner aims at maximizing the expected sum of rewards or at identifying the best performing arm. He needs to sequentially adapt his decision strategy in light of the information gained. Multi-armed bandits have been succesfully used for different applications, the most notable of which being Monte-Carlo Tree Search [Kocsis and Szepesvári, 2006a, Munos, 2014] and the success of AlphaGo [Silver et al., 2016]. In some cases, additional information is available to the learner such as the characteristics of the users and features describing the available items. In this scenario, contextual bandits offer an efficient solution for leveraging this additional data. Bandit algorithms have been successfully applied to various domains including recommender systems [Li et al., 2010, Bouneffouf et al., 2012], for displaying advertisements [Wang et al., 2017], for online learning to rank widgets [Radlinski et al., 2008], for auction designs [Nguyen, 2020, Achddou et al., 2021] and for A/B testing [Kaufmann et al., 2014].

Traditional machine learning methods start by collecting some data and use all available information to train a model. When facing unseen data, the algorithm will predict an outcome based on patterns learned from the training samples. The story is different for bandit algorithms: instead of accumulating data before making any prediction, the learning process is done in an online fashion. In some cases, accumulating data is costly and the learning system needs statistical guarantees as soon as possible. This is typically the case in [Durand et al., 2018] where an allocation strategy for treating mouse skin cancer is proposed. In this example, the learner cannot wait to have 1000 individuals before selecting a reasonable treatment.

However, whether the learning is sequential or not, an algorithm will be able to make accurate predictions on future data only if this data shares similarity with previously collected samples. When the efficiency of medical treatment diminishes with time, any algorithm assuming stationary data will make suboptimal decisions. Similarly, a recommender system that fails to recognize that users have ever-changing preferences is likely to propose irrelevant content.

Bandit algorithms are now used for an ever growing number of tasks where non-stationarity is an important aspect (e.g news recommendations, recommender systems). In order to design successful sequential learning algorithms for those applications, there is a need for a better understanding of bandits in non-stationary environments. This comes with new challenges that need to be solved: (1) the design of algorithms that learn sequentially with only partial feedback in uncertain and dynamically evolving environments. (2) a proper balance between

the exploration and exploitation in those more complicated environments. This brings us to the central question of this thesis:

> How can one design bandit algorithms for non-stationary environments with strong theoretical guarantees and with an emphasis on their practical applicability?

Throughout the thesis, we consider two different settings.

- A variant of the best arm identification setting (Chapter 2) where for a given level of risk δ, the learner tries to identify a set of arms satisfying specific constraints as fast as possible, while guaranteeing that the output set matches the true one with high probability.
- A *reward maximization setting* (Chapter 4-6) where the learner aims at maximizing his expected sum of rewards obtained through the interaction with the bandit.

For both settings, we explain how to deal with non-stationarity while preserving strong theoretical guarantees.

1.1.2 Motivating Examples

Let us introduce a few examples motivating and explaining the framework of our contributions.

E-commerce company. Pauline is an engineer in a large e-commerce company. She has millions of users coming to her website everyday and she is looking for the perfect webpage for a given category of product. She already has a webpage called *control* into production and she is testing different options (called *alternative A* and *alternative B*) to see if she should deploy some of them in addition to the existing one. She is willing to A/B test the different options, i.e. to randomly allocate some proportion of the traffic to the different versions of the webpage to compare their efficiency. Interestingly, she has discovered with the control version that depending on the time of the day, the opening rate of the product differs significantly. Along with the product team, they are wondering if among the different options some would perform better than the control version for some parts of the day. Assuming that the environment is stationary, any A/B test that randomly assigns users to the different versions will converge to the overall best performing version. From the marketing team, Pauline knows that users logging during the night have a specific behavior and are more likely to buy the product she is trying to sell but represent only 1% of the overall traffic. For those users, the control version has little success but they have empirical evidence that alternative A performs well. In this example, assuming a stationary environment, Pauline might miss the opportunity to push the alternative A into production and to satisfy the users logging during the night.

> In Chapter 2, we propose a model that includes non-stationarity and where users from different periods of the days can be treated differently. More precisely, the model we develop has the flexibility to value differently the users from different subpopulations. Based on this model, depending how the E-commerce company values the users logging during the night, Pauline will be able to discover that alternative A is worth deploying.

Coin-game. Let us assume that Clément is playing the following coin game: three coins are placed in front of him with different probabilities of giving a head when tossed. Clément has a total of 100 tosses. His score will be the total number of heads he will obtain over the 100 tosses. Clément is familiar with bandit algorithms and knows that this is typically a game where

a bandit algorithm would offer a solid policy for selecting the coin that should be flipped at each round. Clément's friend Arnaud finds the game boring and wants to further complicate it. He proposes the following alternative: during the game Arnaud will be able to exchange the three coins in front of Clément without notifying him. Clément ignores when the coins will be exchanged but knows that this will happen only once. Clément is thinking out loud: "How should I balance exploration and exploitation in this new game? I want to keep track of the best coin even after Arnaud messes things up."

> In Chapter 3, we propose to help Clément and design asymptotically optimal policies in any abruptly changing environment for the stochastic multi-armed bandit problem. Clément is right about one thing, there is a need to properly balance exploration and exploitation and we propose an efficient algorithm for doing this.

Mosquitoes repellent. This time let us step into the shoes of a producer of mosquitoes repellents. Depending on external factors such as the country (which impacts the species of mosquitoes we encounter), characteristics of users (tolerant or not to a specific molecule), we want to propose an effective repellent. For a given user with fixed characteristics if the environment is stationary, the best repellent will remain the same. Yet, mosquitoes can become more resistant to a repellent which progressively leads to an ineffective product. Adding a non-stationary component to the problem will allow the producer to adapt to more resistant mosquitoes and to keep offering effective treatments.

> In Chapters 4 to 6, we propose contextual bandit algorithms that leverage external information while taking into account potential non-stationarity in the learning task. We use forgetting strategies, that estimate the different parameters using the most recent observations only. By properly calibrating the amount of observations that are kept, we expect being able to offer efficient repellents even if the tolerance of the mosquitoes evolves over time.

1.2 Multi-Armed Bandit

In this section, we present the multi-armed bandit model and the two settings we consider in this thesis: the best arm identification task that will be the starting point for Chapter 2 and the regret minimization setting that will be the object of the following chapters. In each case, we specify what we call an optimal algorithm, present the best guarantees any optimal algorithm should be able to reach and give a brief overview of existing methods.

1.2.1 Presentation of the Model

A stochastic multi-armed bandit is characterized by K unknown probability distributions $(\nu_1, \dots, \nu_K)$ usually called "arms". The learner interacts sequentially with an environment (the K arms from the bandit) by selecting an action (an arm) at each round and receives the associated reward. At time t when selecting an action $A_t \in \{1, \dots, K\}$, a reward X_t drawn from the distribution ν_{A_t} is observed. The term *stochastic* refers to the assumption on how the environment generates rewards, i.e an independent reward sampled from the associated probability distribution. In a *full information* setting, after selecting action $A_t \in \{1, \dots, K\}$ the learner observes the K rewards he would have obtained from all the arms. On the contrary, the bandit feedback is a *partial information* setting where **only the reward associated to the**

selected action A_t is observed. Naturally, this *partial information* setting makes the learning more complicated. A bandit algorithm or policy consists in an allocation strategy for selecting the action A_t that is played at the different rounds.

This model was first studied in [Thompson, 1933] for clinical trials. At each time step, a patient is coming and the doctor can choose between K available treatments. Each of these treatments $a \in \{1, \dots, K\}$ can heal a patient with a probability p_a. When proposing the treatment A_t to a patient t, the doctor observes the health of this patient X_t. There is a clear tension for the doctor between focusing on seemingly well performing treatments and trying the other ones for making sure that they are not better. This exploration-exploitation trade-off is central in bandits. Quite far from medicine, the denomination "bandit" refers to a slot machine that is popular in casino.

The amount of information available to the learner at time t can be summarized with $\mathcal{F}_{t-1}$ the σ-algebra generated by the actions and the rewards collected up to round t, i.e. $\mathcal{F}_{t-1} = \sigma(A_1, X_1, \dots, A_{t-1}, X_{t-1})$. We assume that the action A_t selected by a bandit algorithm is $\mathcal{F}_{t-1}$-measurable, i.e the action can be selected based on past information only. Of course, the different probability distributions are unknown to the learner and an efficient policy will have to estimate the quantities of interest depending on the learning objective.

In the following, we present two different learning objectives: the *regret minimization* framework where the learner wants to maximize his expected sum of rewards, explaining the analogy with slot machines. We also consider *pure exploration* tasks where there is no cost for exploring the different options.

1.2.2 Regret Minimization

In the regret minimization setting, the learner seeks to maximize his expected sum of rewards. Let us denote $\mu_k := \mathbb{E}[\nu_k]$ the mean of the arm k. If the environment was perfectly known to the learner (i.e. the different probability distributions were known) only the arm $k^\star$ with the highest mean, denoted $\mu^\star := \text{argmax}_{k \in \{1,\dots,K\}} \mu_k$, would be pulled. The best expected cumulative reward one can expect when playing for T rounds is then $T \times \mu^\star$. The expected regret (usually only called regret for simplicity) quantifies how close the learner can get to this oracle when following a given policy for selecting the actions to play.

Definition 1.1. For a policy π, a bandit model $\nu = (\nu_1, \dots, \nu_K)$ and a T rounds interaction with the environment, the regret is defined as

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

We call sub-optimal an arm whose mean is strictly smaller than the highest mean $\mu^\star$. Interestingly, the regret can be related to the number of times the different sub-optimal arms have been pulled. Denoting $N_k(T) = \sum_{t=1}^{T} \mathbb{1}(A_t = k)$, the number of pulls of arm k up to time T, the regret can be rewritten as:

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

This equation features a key quantity $\Delta_k := \mu^\star - \mu_k$ called *gap* that represents how far on average a reward from arm k will be, compared to a reward from the optimal arm. Δ_k is strictly

positive for any suboptimal arm. The exploration-exploitation trade-off previously mentioned can already be understood from Equation (1.1). For obtaining a small regret, there is a need to identify the arms with small gaps and to pull them most of the time. The ideal scenario being a strategy that only selects the arm $k^\star$ and would not suffer any regret. Yet, the only way to gain information on arm k is to pull it, hence augmenting N_k and the regret. A trade-off has to be found between gaining more information about an arm to detect that its gap is smaller than previously thought, or focusing on the arm for which the estimated gap is the smallest.

Environment Class. A common assumption is that the K arms come from the same family of distributions that we call *environment class* and denote $\mathcal{E}$. For example, we call a Bernoulli bandit an instance for which $\nu_1, \dots, \nu_K$ are Bernoulli distributions. In Chapter 2 and Chapter 3, we consider the more general class of *exponential family bandit models* where any bandit instance is of the form $\nu = (\nu_{\theta_1}, \dots, \nu_{\theta_k})$, where $\theta_k \in \Theta$ for all k and the arms belong to the same one-parameter exponential family:

$$\mathcal{P} = \left\{ (\nu_\theta)_\theta : \frac{d\nu_\theta}{d\xi} = \exp(\theta x - b(\theta)) \right\} . \tag{1.2}$$

Here, ξ is a reference measure on $\mathbb{R}$ and b is the log-partition function that takes input in Θ and values in $\mathbb{R}$ and is assumed to be twice differentiable. For the exponential family bandit model, the mean of a distribution ν_θ is entirely characterized by θ through $\mathbb{E}[\nu_\theta] = \dot{b}(\theta)$. In this case, $\mathcal{E} = \{\nu = (\nu_{\theta_1}, \dots, \nu_{\theta_K}) \text{ s.t } \forall k \in \{1, \dots, K\}, \nu_{\theta_k} \in \mathcal{P}\}$.

The exponential family environment class contains most of usual distributions such as Bernoulli, Poisson, Gaussian with known variance. Our results for multi-armed bandits will be stated for this environment class. In the following, we formalize the natural idea that one wants to design strategies that work well for any $\nu \in \mathcal{E}$ and not only for specific instances.

On the Notion of Optimality. Going back to Equation (1.1), by upper-bounding Δ_k by the largest gap denoted $\Delta_{\max}$ and the expected number of pulls by T, for any policy we have $\mathcal{R}_\nu(T, \pi) \leq \Delta_{\max} T$. A linear regret means that a constant error is made at each step and that potentially no learning occurs throughout the interaction with the environment. For this reason, we aim for policies with *sub-linear* regret within an environment class $\mathcal{E}$ for which:

$$\forall \nu \in \mathcal{E}, \lim_{T \to \infty} \frac{\mathcal{R}_\nu(T, \pi)}{T} = 0 . \tag{1.3}$$

Equation (1.3) naturally brings the notion of *worst-case regret* of a policy π in an environment class $\mathcal{E}$ that is defined bellow.

Definition 1.2 (Worst case regret)**.** For a policy π, an environment class $\mathcal{E}$ and a T rounds interaction with the environment, the worst-case regret of π within $\mathcal{E}$ is defined as

$$\mathcal{R}_\mathcal{E}(T, \pi) := \sup_{\nu \in \mathcal{E}} \mathcal{R}_\nu(T, \pi) .$$

While we can always find a policy with null regret for a specific instance (by always selecting the best arm), obtaining a small regret over all possible instances of an environment class is a stronger and more meaningful metric for evaluating the performance of a policy. Let Π be a set of policies, it is sometimes interesting to understand how well the best policy within Π performs on the hardest instance of the environment class. This metric is the minimax regret and we define it below.

Definition 1.3 (Minimax regret)**.** For a set of policies Π, an environment class $\mathcal{E}$ and a T rounds interaction with the environment, the minimax regret of the set of policies Π within $\mathcal{E}$ is defined as

$$\mathcal{R}^\star_{\mathcal{E}}(T) := \inf_{\pi\in\Pi} \mathcal{R}_{\mathcal{E}}(T,\pi) = \inf_{\pi\in\Pi} \sup_{\nu\in\mathcal{E}} \mathcal{R}_\nu(T,\pi) \ .$$

A policy π is said to be minimax optimal when π satisfies $\mathcal{R}_{\mathcal{E}}(T,\pi) = \mathcal{R}^\star_{\mathcal{E}}(T)$.

While interesting, upper-bounds for the regrets are only meaningful when compared to lower-bounds. Lower-bounds essentially give us the best guarantee any algorithm can hope for in a particular setting. An algorithm can be said to be *optimal* when an upper-bound for the regret that matches a lower bound for the same setting can be obtained.

Minimax Lower-Bound. What types of guarantees can we expect for a finite time horizon for the hardest instance in the exponential family bandit model? Let us assume that $\nu_1, \dots, \nu_K$ are Gaussian distributions with unit variance for simplicity. The following theorem shows that for any policy π we can find a Gaussian instance such that the regret of the policy π for that instance is larger than $\mathcal{O}(\sqrt{KT})$.

Theorem 1.4 (Theorem 15.2 from [Lattimore and Szepesvári, 2020])**.** *Let $K > 1$ and $T \geq K-1$. Then, for any policy π there exists an instance of Gaussian bandits with unit variances and a mean vector $\boldsymbol{\mu} \in [0,1]^K$ denoted $\nu_{\boldsymbol{\mu}}$ such that*

$$\mathcal{R}_{\nu_{\boldsymbol{\mu}}}(T,\pi) \geq \frac{1}{27}\sqrt{(K-1)T} \ .$$

The main idea for obtaining this result is to design an instance such that the learner is not able to distinguish between the different arms. In the particular case where $K = 2$, the objective is to find the maximum gap such that we can fool the learner for the entire time horizon T and is illustrated on Figure 1.1. It can be shown that this maximum gap is of order $1/\sqrt{T}$. Hence, with a gap of this order, the learner will essentially choose randomly and will on average select the bad arm $T/2$ times. This implies a regret of order $T \times 1/\sqrt{T}$ and gives the announced dependence in T. For a given policy π and an instance $\boldsymbol{\mu}$, the proof of this result relies on building a well-designed instance $\boldsymbol{\mu}'$ such that the learner can not perform well on both $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$.

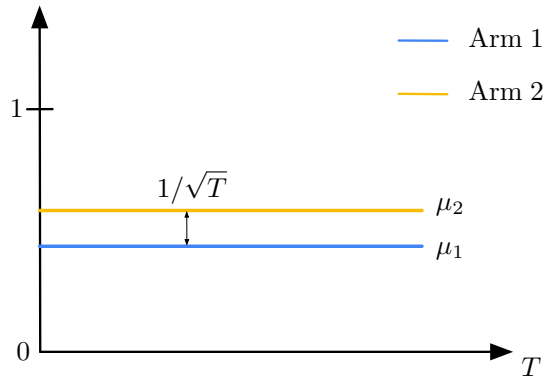


Figure 1.1: Indistinguishable means for a learner interacting for T steps with the bandit (μ_1, μ_2).

Given that the exponential family bandit model contains the Gaussian bandits with unit variance, an immediate consequence of Theorem 1.4 is that the minimax lower bound in the more general exponential family bandit model is at least of this order.

Problem-Dependent Lower-Bounds. The worst case regret measures a certain form of robustness of a policy but can be overly conservative. In particular, a problem-dependent bound that quantifies how well a policy will perform on a given instance is also an interesting indicator. For this reason, other forms of optimality have been considered. Let us first define the notion of *uniformly efficient* strategy [Lai and Robbins, 1985].

Definition 1.5 (Uniformly efficient policy)**.** Let $\mathcal{E}$ be the exponential family bandit model. A policy π is said to be uniformly efficient if, for all $\nu \in \mathcal{E}$ with a unique optimal arm, one has

$$\forall \alpha > 0, \ \lim_{T\to\infty} \frac{\mathcal{R}_\nu(T,\pi)}{T^\alpha} = 0 \ .$$

When the number of interactions with the environment increases, we expect a sound learner to make fewer mistakes and to select the best arm more frequently. With Definition 1.5 a uniformly efficient policy is effectively able to learn with a regret scaling sublinearly with the time horizon for every instance in the environment class.

Let $\theta_1, \theta_2 \in \Theta$ and consider ν_{θ_1}, ν_{θ_2} two distributions from $\mathcal{P}$ (defined in Equation (1.2)) with respective mean μ_1 and μ_2. The Kullback-Leibler divergence from ν_{θ_1} to ν_{θ_2} induces a divergence function d on $\dot{b}(\Theta)$. Knowing that $\dot{b}(\theta_1) = \mu_1$ and $\dot{b}(\theta_2) = \mu_2$, it is defined by:

$$d(\mu_1, \mu_2) = \mathrm{KL}(\nu_{\theta_1}, \nu_{\theta_2}) = b(\theta_2) - b(\theta_1) - \mu_1(\theta_2 - \theta_1) \ . \tag{1.4}$$

Following [Lai and Robbins, 1985], for the exponential family bandit model and for any instance with mean $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$, the regret can be lower-bounded with a problem-dependent quantity.

Theorem 1.6 (Lai and Robbins Lower Bound)**.** *Let $\mathcal{E}$ be the exponential family bandit model. For any instance ν with mean vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ with a unique optimal arm denoted $k^\star$, and any uniformly efficient policy,*

$$\liminf_{T\to\infty} \frac{\mathcal{R}_\nu(T,\pi)}{\log(T)} \geq \sum_{k\neq k^\star} \frac{\Delta_k}{d(\mu_k, \mu_{k^\star})} \ .$$

Theorem 1.6 is a direct consequence of a related result regarding the minimum expected number of pulls of any suboptimal arm for a uniformly efficient policy:

$$\forall k \neq k^\star, \ \liminf_{T\to\infty} \frac{\mathbb{E}_\nu[N_k(T)]}{\log T} \geq \frac{1}{d(\mu_k, \mu_{k^\star})} \ . \tag{1.5}$$

The proof of Equation 1.5 relies on changes of distribution and has interesting consequences for policies that have sublinear regret on every instance $\nu \in \mathcal{E}$. First, it proves that asymptotically every sub-optimal arm will be pulled infinitely often at a logarithmic rate. Second, this rate is exactly $1/d(\mu_k, \mu^\star)$. Note that for the Gaussian case with unit variance as $d(\mu_k, \mu^\star) = (\mu_k - \mu^\star)^2/2$, we recover the natural idea that the closer from the optimal arm, the more frequent a suboptimal arm will have to be pulled when a learner is trying to maximize his expected sum of rewards.

[Burnetas and Katehakis, 1996] have extended this result to more general environment classes that are out of the scope of this thesis. With this lower-bound at hand, we can define the notion of *asymptotically optimal policy*.

Definition 1.7 (Asymptotically optimal policy)**.** Let $\mathcal{E}$ be the exponential family bandit model. A policy π is said to be asymptotically optimal if if, for all $\nu \in \mathcal{E}$ with mean vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$:

$$\lim_{T\to\infty} \frac{\mathcal{R}_\nu(T,\pi)}{\log(T)} = \sum_{k\neq k^\star} \frac{\Delta_k}{d(\mu_k, \mu_{k^\star})} \, .$$

Some work has also been done to provide finite-time (instead of asymptotic) problem-dependent lower bounds e.g. [Garivier et al., 2019] but those will not be discussed here.

In Chapter 3, we build on the notion defined here and propose LB-SDA an algorithm that is asymptotically optimal for the exponential family bandit model. This remains true without the need to know the distribution of the arms (e.g. Bernoulli, Poisson or Gaussian with known variance). This makes the algorithm appealing because with the same implementation, we obtain strong theoretical guarantees for a broad class of distributions.

To balance exploration and exploitation and obtain asymptotically optimal strategies, two standard techniques have been studied in the literature: *Upper-Confidence Bound* or *Thompson Sampling*. We briefly introduce those concepts in the following sections and refer the interested reader to [Lattimore and Szepesvari, 2020] for other common approaches (resampling, epsilon-greedy, etc.).

1.2.2.1 Upper-Confidence Bounds

Throughout his interaction with the environment, the learner can build confidence intervals for estimating the K unknown means. At time t, for an arm k that has been played $N_k(t)$ times and a given level of risk δ, the learner can build a confidence region $\mathcal{I}_k(t,\delta) = [L_k(t,\delta), U_k(t,\delta)]$ such that:

$$\forall k \in \{1, \dots, K\}, \ \mathbb{P}(\mu_k \in \mathcal{I}_k(t,\delta)) \geq 1 - \delta \, .$$

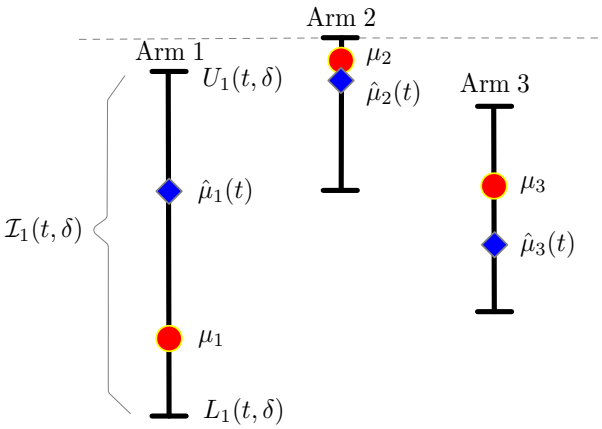


Figure 1.2: Upper confidence bounds for three arms at time t. When following the OFUL principle, the arm 2 (with the highest UCB) will be selected by the learner.

Optimism in the Face Of Uncertainty. The optimism in the face of uncertainty (OFUL) principle [Agrawal, 1995, Auer et al., 2002a] recommends to play "as if the environment was as nice as plausibly possible" [Lattimore and Szepesvári, 2020]. At time t, the mean of the arm k can be estimated using the upper bound of the confidence region $\mathcal{I}_k(t-1,\delta)$. In its simplest formulation this upper confidence bound (UCB) takes the form:

$$U_k(t,\delta) = \hat{\mu}_k(t) + c_k(t,\delta) \ ,$$

where $\hat{\mu}_k(t) = \frac{1}{N_k(t)} \sum_{i=1}^{N_k(t)} Y_{k,i}$ where $Y_{k,i}$ denotes the i-th sample obtained from distribution ν_k and $c_k(t,\delta)$ is the exploration bonus and depends on the distribution at hand. The OFUL principle recommends to play the arm with the largest UCB at each round. When the number of samples from arm k increases, the hope is that $\mathcal{I}_k(t,\delta)$ concentrates around the true mean.

Let us describe the simplest version of an algorithm based on the use of upper-confidence bounds. [Auer et al., 2002a] propose UCB-1 for stochastic bandits with distributions with support in $[0,1]$. In this case, at time t for an arm k that has been played $N_k(t-1)$ times, the UCB takes the following form:

$$U_k(t) = \hat{\mu}_k(N_k(t-1)) + \sqrt{\frac{2\log(t)}{N_k(t-1)}} \ . \tag{1.6}$$

Based in this quantity, the algorithm UCB1 works as follows

Input: $(\nu_1, \dots, \nu_K)$ with support in $[0,1]$
Initialization: $N_k(0) = 0$ for all $k \in \{1, \dots, K\}$
for $t \leq T$ **do**
  if $t \leq K$ **then**
    Play action $A_t = t$
  else
    Play action $A_t = \text{argmax}_{k\in\{1,\dots,K\}} \hat{\mu}_k(N_k(t-1)) + \sqrt{\frac{2\log(t)}{N_k(t-1)}}$
Receive reward $X_t \sim \nu_{A_t}$
Updating phase:
  $N_k(t) = N_k(t-1) + \mathbb{1}(A_t = k)$ for $k \in \{1, \dots, K\}$
  Update the empirical means and the Upper-confidence bounds.

Algorithm 1: UCB1 [Auer et al., 2002a]

Intuitively, the best arm is expected to have the largest UCB most of the time. Yet, the UCB of an arm that is not pulled anymore will increase up to the point where it reaches the highest UCB. At this moment, another observation will be collected for this arm and its UCB will be refined. This is in particular the case with UCB1 and the $\sqrt{\log(t)}$ term in the exploration bonus. This principle allows for a natural balance between exploration and exploitation and ensures that every suboptimal arm will be pulled infinitely often as $T \to \infty$, which is necessary according to the lower bound from Theorem 1.6 and Equation (1.5).

1.2.2.2 Thompson Sampling

An alternative to UCB based algorithm is a Bayesian algorithm called Thompson Sampling. This method was the first one to be proposed for studying bandit models by [Thompson, 1933]

when $K = 2$. Thompson Sampling, as the name suggests, is a method based on randomized arm pulls. In this Bayesian framework, the learner has an initial prior on the different distributions. Throughout the interaction with the environment, the learner computes the posterior distribution for each of the K arms based on the previous rewards and actions. At each time step, samples from the K posterior distributions are obtained and the learner plays the arm associated with the largest sample. Interestingly, in this setting the exploration is not imposed by construction but is the consequence of the randomness of the samples. An arm that has been barely pulled will have a poorly concentrated posterior and the rewards obtained by sampling from this posterior will vary a lot. On the contrary, the more an arm is pulled, the tighter the posterior concentrates around the true distribution and the learner starts exploiting.

Exploration by sampling will also be a central component of the subsampling algorithms we propose in Chapter 3.

1.2.2.3 Asymptotically Optimal Policies

In the exponential family bandit model where the K arms come from the same one-parameter exponential family, both UCB techniques and Thompson Sampling approaches are now known to be asymptotically optimal. [Cappé et al., 2013] designed kl-UCB a UCB based algorithm using KL divergences for obtaining the upper-confidence bound for the different arms. Using these refined bounds, the authors were able to prove the asymptotic optimality of this strategy. When selecting a proper prior, Thompson Sampling is also asymptotically optimal in the exponential family bandit model [Kaufmann et al., 2012, Korda et al., 2013]. To some extent, all of these approaches need to have some information on the rewards distributions at hand. Relaxing this while maintaining the asymptotic optimality is at the core of Chapter 3.

1.2.3 Pure Exploration Tasks

In some cases, the cost of exploration is less important and the regret is not the most appropriate metric. For example, in an A/B test experiment, an e-commerce company probably cares more about being able to identify as fast as possible the best version of a webpage at the cost of potentially losing a few clients. We refer to the tasks where there is no penalization for the exploration as *pure exploration tasks*. In those tasks, there is no exploration-exploitation trade-off and optimal policies for this setting are not necessarily strong candidates for the regret minimization setting and vice versa. The most studied pure exploration setting is the task of identifying the arm with the highest mean: it is entitled the *Best Arm Identification* task, and we describe it in more detail below.

Fixed Confidence vs. Fixed Budget Setting. Let $\hat{\mathcal{S}}_T \in \{1, \dots, K\}$ denote the arm recommended by a strategy after T interactions with the environment. Two different settings are of interest. In the *fixed budget* setting, the learner is given a number of rounds T and the objective is to design a sampling strategy that minimizes $\mathbb{P}(\hat{\mathcal{S}}_T \neq k^\star)$ where $k^\star$ is the arm with the highest mean $\mu^\star$. On the contrary, in the *fixed confidence* setting the number of rounds is not fixed in advance. Here, the learner seeks to design a policy that requires as few samples as possible for identifying $k^\star$ with a confidence of at least $1 - \delta$. Mathematically, for any risk level δ, the learner chooses a stopping rule (a stopping time) τ_δ adapted to the filtration $\mathcal{F} = (\mathcal{F}_t)_{t \geq 0}$ and aims at minimizing $\mathbb{E}[\tau_\delta]$ while satisfying the δ-correctness constraint:

$$\mathbb{P}(\tau_\delta < \infty, \hat{\mathcal{S}}_{\tau_\delta} \neq k^\star) \leq \delta \, . \tag{1.7}$$

The constraint ensures that if the algorithm is able to stop in a finite time, then the probability of recommending the wrong arm should be smaller than the risk δ. In this thesis, we only focus on the fixed confidence setting and we refer to [Bubeck et al., 2009, Audibert et al., 2010] for a better understanding of the fixed budget setting. From a high level, any policy tailored to the fixed confidence setting has three essential components: (i) a sampling rule: how should we select A_t based on $\mathcal{F}_{t-1}$? (ii) a stopping rule: when should we stop playing? and (iii) a recommendation rule: which arm should we output?

On the Complexity of Identifying the Best Arm. Similarly to the regret minimization setting, we want to understand the complexity of identifying the best arm for any bandit instance ν. This complexity will be the target for the expected number of pulls required by the different sampling methods we will develop. In the pure exploration setting, lower bounds are not only useful for characterizing optimal algorithms but they also guide the design of asymptotically optimal sampling strategies. As mentioned above, we focus on δ-correct strategies that guarantee that the correct arm is identified with probability higher than $1-\delta$. We define $\mathcal{L}$ the set of bandit instances with a unique optimal arm and where all the K distributions come from the same one-parameter exponential family. Recall that an instance $\nu \in \mathcal{L}$ is entirely defined by its mean vector $\boldsymbol{\mu}$. For a bandit instance $\nu_{\boldsymbol{\mu}} \in \mathcal{L}$ with mean vector $\boldsymbol{\mu}$, we denote $k^\star(\boldsymbol{\mu})$ the arm with the highest mean and $\Sigma_K = \{x \in [0,1]^K, \sum_i x_i = 1\}$ the K-dimensional simplex. For a given mean vector $\boldsymbol{\mu}$, we define the alternative model as the set of instances for which the optimal arm is different than for the instance $\boldsymbol{\mu}$. Formally:

$$\mathrm{Alt}(\boldsymbol{\mu}) := \left\{ \nu_{\boldsymbol{\lambda}} \in \mathcal{L} \;\middle|\; k^\star(\boldsymbol{\lambda}) \neq k^\star(\boldsymbol{\mu}) \right\} .$$

We now give a lower-bound for the complexity of identifying the best arm.

Theorem 1.8 (Theorem 1 in [Garivier and Kaufmann, 2016])**.** *Let $\delta \in (0,1)$. For any δ-correct strategy and any bandit model $\nu_{\boldsymbol{\mu}} \in \mathcal{L}$,*

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

where

$$T^\star(\boldsymbol{\mu})^{-1} := \sup_{w \in \Sigma_K} \inf_{\lambda \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{k=1}^{K} w_k d(\mu_k, \lambda_k) .$$

Once this complexity is obtained, the authors of [Garivier and Kaufmann, 2016] propose a sampling strategy called D-Tracking and a stopping criteria based on a Generalized Likelihood Ratio statistic for obtaining an asymptotically optimal algorithm in this setting. In the following, we explain how the stopping rule and the sampling strategy work as they will be our starting point for Chapter 2.

First note that the supremum for the optimal weights is indeed a maximum as established in [Garivier and Kaufmann, 2016]. Using this, we learn from the proof of Theorem 1.8 that the only way to obtain a policy matching the lower bound is to sample the different arms k with a proportion $w_k^\star(\boldsymbol{\mu})$, where:

$$w^\star(\boldsymbol{\mu}) = \mathrm{argmax}_{w \in \Sigma_K} \inf_{\lambda \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{k=1}^{K} w_k d(\mu_k, \lambda_k) .$$

This is the guiding principle that is used for building the sampling rule.

Sampling Rule. At time t, based on the previous interactions with the environment, the learner has access to $\hat{\boldsymbol{\mu}}_t$ an estimate of the means of the K arms. The most naive idea that we can have for estimating $w^\star(\boldsymbol{\mu})$ is to compute $w^\star(\hat{\boldsymbol{\mu}}_t)$ and to draw the next arm based on this value. This is the central idea of the D-Tracking strategy that is proposed in [Garivier and Kaufmann, 2016]. However, an additional forced exploration is necessary for ensuring that $\hat{\boldsymbol{\mu}}_t$ will converge to $\boldsymbol{\mu}$ when t tends to infinity. We introduce $U_t = \{k \in \{1, \dots, K\} \mid N_k(t) < \sqrt{t} - K/2\}$. Using the D-Tracking strategy, the arms are pulled as follows:

$$A_t \in \begin{cases} \text{argmin}_{k \in U_t} N_k(t) \text{ if } U_t \neq \emptyset \ , \\ \text{argmax}_{1 \leq k \leq K} \ t w_k^\star(\hat{\boldsymbol{\mu}}_t) - N_k(t) \text{ otherwise} \ . \end{cases}$$

Note that taking the argmax for $tw_k^\star(\hat{\boldsymbol{\mu}}_t) - N_k(t)$ is natural: it permits to identify the arm for which the gap between $N_k(t)/t$ and $w_k^\star(\hat{\boldsymbol{\mu}}_t)$ is the largest. By pulling it in the next round, we necessarily reduce this gap for the next steps and tend towards drawing the different arms with their optimal proportions $w^\star(\hat{\boldsymbol{\mu}})$ which will converge to $w^\star(\boldsymbol{\mu})$.

Stopping Rule. We now have the ingredients for explaining the stopping criterion. The objective with the stopping rule is to stop as soon as we have statistical evidence that one arm is better than all the others. For $k_1, k_2 \in \{1, \dots, K\}^2$, an efficient way for doing this is to consider the Generalized Likelihood Ratio statistic:

$$Z_{k_1,k_2}(t) \coloneqq \log \frac{\max_{\mu'_{k_1} \geq \mu'_{k_2}} p_{\mu'_{k_1}}(\mathcal{H}_{k_1}(t)) p_{\mu'_{k_2}}(\mathcal{H}_{k_2}(t))}{\max_{\mu'_{k_1} \leq \mu'_{k_2}} p_{\mu'_{k_1}}(\mathcal{H}_{k_1}(t)) p_{\mu'_{k_2}}(\mathcal{H}_{k_2}(t))} \ . \tag{1.8}$$

In the above expression, $\mathcal{H}_k(t) = \{X_s \mid A_s = k, s \leq t\}$ is the history of the rewards available at time t for arm k and $p_{\mu_k}(Z_1, \dots, Z_n)$ is the likelihood of n i.i.d observations from a distribution ν_{μ_k}. Recalling that we consider a one-parameter exponential family, $p_{\mu_k}(Z_1, \dots, Z_n)$ can be expressed as a function of μ_k following:

$$p_{\mu_k}(Z_1, \dots, Z_n) = \prod_{i=1}^{n} \exp\left(\dot{b}^{-1}(\mu_k) Z_i - b(\dot{b}^{-1}(\mu_k))\right) \ .$$

Using this expression a closed form expression for $Z_{k_1,k_2}(t)$ can be obtained. Intuitively, a large $Z_{k_1,k_2}(t)$ suggests that at time t there is statistical evidence that the arm k_1 has a largest mean than the arm k_2. It remains to quantify what "large" means for satisfying the δ-correct constraint. Introducing the threshold $\beta(t, \delta)$ (that needs to be tuned appropriately), the stopping rule is the following:

$$\tau_\delta \coloneqq \inf\left\{ t \geq 0 \ \middle| \ \exists k \in \{1, \dots, K\}, \forall k' \in \{1, \dots, K\}, \ Z_{k,k'}(t) > \beta(t, \delta) \right\} \ . \tag{1.9}$$

This stopping rule is usually referred to as Chernoff's stopping rule because it shares similarities with [Chernoff, 1959]. We now have all the ingredients for presenting the Track-And-Stop strategy introduced in [Garivier and Kaufmann, 2016] and reported in Algorithm 2.

With a proper tuning of the threshold $\beta(t, \delta)$, the δ-correct constraint is satisfied and Track-and-Stop reaches asymptotic optimality.

Proposition 1.9 (Proposition 12 in [Garivier and Kaufmann, 2016])**.** *Consider the exponential family bandit model. Let $\delta \in (0, 1)$ and $\alpha > 1$. There exists a constant $C = C(\alpha, K)$ such that whatever the sampling strategy, using Chernoff's stopping rule from Equation (1.9)*

Input: K arms, horizon T, $\beta(t, \delta)$ threshold
while $\max_{k \leq K} \min_{k' \leq K} Z_{k,k'}(t) \leq \beta(t, \delta)$ **do**
  if $U_t \neq \emptyset$ **then**
    Pick arm $A_{t+1} = \operatorname{argmin}_{k \in U_t} N_k(t)$
  else
    Pick $A_{t+1} = \operatorname{argmax}_{1 \leq k \leq K} tw_k^\star(\hat{\boldsymbol{\mu}}_t) - N_k(t)$
  Obtain sample X_{t+1} from $\nu_{A_{t+1}}$
Result: Recommend $\hat{\mathcal{S}}_{\tau_\delta}$

Algorithm 2: Track-and-Stop with D-Tracking sampling

with the threshold

$$\beta(t, \delta) = \log\left(\frac{Ct^\alpha}{\delta}\right)$$

ensures that for all $\nu \in \mathcal{L}$

$$\mathbb{P}_\nu(\tau_\delta < \infty, \hat{k}_{\tau_\delta} \neq k^\star) \leq \delta \, .$$

Proposition 1.9 proves that regardless of the used sampling strategy, the Chernoff's stopping rule is sufficient to guarantee the δ-correctness of a policy. It is also possible to establish the asymptotic optimality of the Track-and-Stop algorithm.

Theorem 1.10 (Theorem 14 in [Garivier and Kaufmann, 2016])**.** *Consider the exponential family bandit model. Let* $\alpha \in [1, e/2]$ *and* $r(t) = \mathcal{O}(t^\alpha)$*. Using Chernoff's stopping rule with* $\beta(t, \delta) = log(r(t)/\delta)$ *and the D-Tracking sampling strategy, the following holds*

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

In Chapter 2, we consider a variant of the best arm identification task where the objective is to find all the arms that are better than a baseline. We start by quantifying the complexity of this task and propose asymptotically optimal algorithms for solving it. We adapt the D-Tracking and the stopping rule presented in this section to this setting. The framework we propose allows us to model (among others) the practical scenario where a company is looking for the best variant of a webpage and want to identify the candidates that are better than a default version that is used in production. Chapter 2 is based on work done in collaboration with researchers from an online travel agency. A crucial aspect of this problem is the non-stationarity of the data as represented on Figure 1.3 with an actual A/B/n experiment run by this company. The metric of interest in this experiment is whether the visitor clicked at least once during the experiment to the next page after getting exposed to one of the variants. In this experiment, $K = 2$ copies compete against the control used in production. Due to global traffic, the data exhibits strong seasonality patterns within a day, as seen in Figure 1.3, in which every point corresponds to click-through rate per six hours (quarter of day) for 12 consecutive days. We explain how to adapt pure exploration tasks to this more general scenario in Chapter 2.

Understanding non-stationary environments where the distributions of the arms can evolve over time is of interest for practical applications. In the following section, we give an overview of

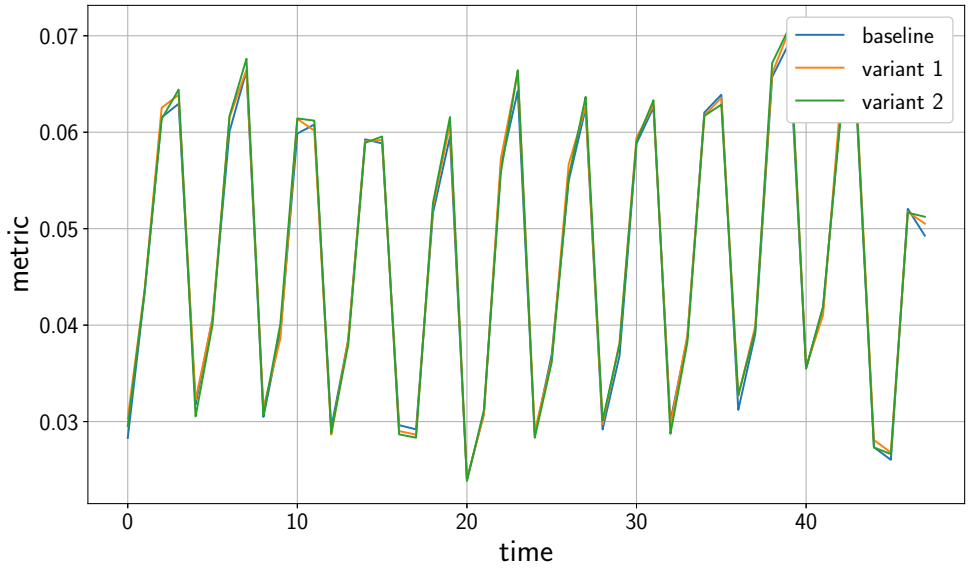


Figure 1.3: Click-through rate per 6 hours for 12 days for different variants of a webpage.

non-stationary stochastic bandits and we describe common assumptions made on the structure of the non-stationarity.

1.3 Non-Stationary Stochastic Bandits

Scenarios where non-stationarity is a crucial aspect of the learning have already been considered in the 1980s. In [Whittle, 1988], the author considers the case of the treatment of a mutating virus. A doctor has K treatments and their efficiency is evolving as soon as the virus mutates. In this example, even when a treatment is not selected by the doctor, its performance might change.

Historically, [Gittins, 1974] first proposed to solve the following problem that was raised as soon as the Second World War. A learner is offered to select among K different projects and only one project can be selected at a time. When selecting the project k, the state of this project is modified and the learner receives the reward from this project. The overall objective of the learner is to maximize the expected discounted reward.

[Whittle, 1988] proposes another martial example where non-stationarity naturally appears. Assume that a learner has one plane and aims at tracking n different enemy submarines. The objective is to properly allocate the plane to the different positions for monitoring the submarines. When a submarine is under observation, the learner gathers information about its position, its velocity, etc. For all the submarines that are not observed during that time, some information is lost and the uncertainty about their position increases. The future position of the submarine is hardly predictable. Those examples motivate the study of non-stationary environments where the state of an action can evolve depending on the action taken by the learner or not.

The model that we have presented so far assumes that the K distributions $(\nu_1, \dots, \nu_K)$ remain constant over time. Specifically, the notion of regret from Definition 1.1 can be adapted when the distributions are time-dependent. We now denote $\nu_{k,t}$ the distribution of arm k at time t. When maximizing the expected sum of rewards, an oracle that would know the distributions $(\nu_{k,t})$ of the arms at every time step will switch to the highest mean as soon as there is a change.

This requires defining a stronger competitor that selects the best arm at every time step, and the associated notion of regret.

Definition 1.11. For a policy π and a T rounds interaction in an environment with time dependent distributions, the dynamic regret is defined as

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

It is natural to wonder how to model non-stationarity. If the distributions can evolve in a completely arbitrary way, there is no hope for learning. For this reason, adding some structure on the non-stationarity seems reasonable. A variety of non-stationary bandits have been considered with different levels of assumption on the non-stationarity structure. At one extreme, non-stationarity follows probabilistic dynamics and we can try to predict the changes. In [Whittle, 1988] for example, the state of all the arms goes from $\mu_k(t)$ to $\mu_k(t+1)$ following a Markov rule specific to each arm. At the other extreme lies non-stochastic adversarial models where the rewards are arbitrary and are not even supposed to be drawn stochastically [Auer et al., 2002b]. Our interest in this thesis lies between these two extremes. We do not make a stochastic assumption on the non-stationarity itself, but the means of the rewards' distributions can not evolve completely arbitrarily. We focus on two models of non-stationarity that have received significant interest: *abruptly changing* environments that we describe in Section 1.3.1 and the *variation budget model* allowing for a broader class of non-stationary environments that is the subject of Section 1.3.2.

1.3.1 Abruptly Changing Environments

We first consider abruptly changing (or piece-wise stationary) bandits, where over a T rounds interaction, the environment can change at most Γ_T times. Formally, this requires

$$\sum_{t=1}^{T-1} \mathbb{1}\Big\{\exists k \in \{1,\dots,K\} \;\Big|\; \mu_{k,t} \neq \mu_{k,t+1}\Big\} \leq \Gamma_T .$$

The time instants $(t_1,...,t_{\Gamma_T})$ associated to these breakpoints define Γ_T+1 stationary phases where the reward distributions are fixed. This setting was probably considered because for each stationary phase there is hope for leveraging existing tools from the stationary bandit literature and to adapt them to piece-wise stationary environments. How hard is it to learn a policy in abruptly changing environments? The following theorem offers an interesting answer.

Theorem 1.12 (Theorem 31.2 in [Lattimore and Szepesvári, 2020])**.** *Let π be a policy, assume that $K=2$ and that $\mu_i(t)=\mu_i$ is constant for both arms. Assume that the arm 1 is optimal and define $\Delta=\mu_1-\mu_2>0$. If the regret of policy π on the instance (μ_1,μ_2) satisfies $\mathcal{R}_\mu(T,\pi)=o(T)$, then for T large enough, there exists a non-stationary bandit ν' with means $(\mu_1'(t),\mu_2'(t))$ for $t\leq T$ with at most 2 breakpoints, satisfying $\min_{t\leq T}|\mu_1'(t)-\mu_2'(t)|\geq\Delta$ such that*

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

Let us explain the implication of this result. If a policy π is such that it has a worst-case regret on stationary instances of order $\mathcal{R}(T,\pi)=\mathcal{O}(\sqrt{T})$ then Theorem 1.12 ensures that an instance with at most 2 breakpoints can be built with a regret of order $\Omega(\sqrt{T})$. On the

other hand, if a policy π has strong problem-dependent guarantees on a specific instance $\boldsymbol{\mu}$, a regret of order $\mathcal{R}_{\boldsymbol{\mu}}(T,\pi) = \mathcal{O}(\log T)$ is achievable (see asymptotically optimal policy from Definition 1.7). Theorem 1.12 implies that this policy would suffer a regret of order $\Omega(T/\log T)$ for a non-stationary instance. Essentially, this result smooths out the hope for obtaining policies with a regret of order $\log T$ in abruptly changing environments. The intuition by which this is true is quite natural. If a learner wants to hedge against an abrupt change, he has to explore suboptimal arms frequently to make sure that they have not changed. When an exploration of order $\log T$ was sufficient in stationary environments, in non-stationary environments, additional exploration is required: without it there is a risk of always making suboptimal decisions and to suffer linear regret.

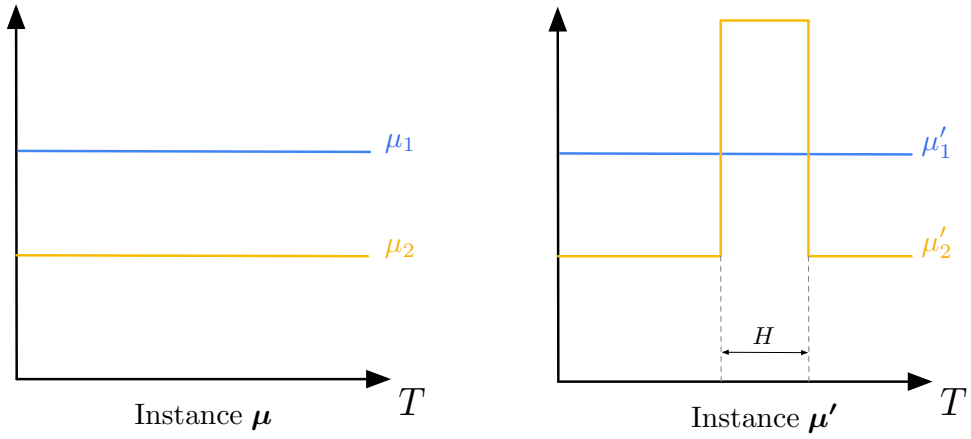


Figure 1.4: Trade-off between the regret on a stationary instance and a non-stationary instance. A policy π optimal on $\boldsymbol{\mu}$ will have a large regret on $\boldsymbol{\mu}'$.

Theorem 1.12 can be traced back to [Garivier and Moulines, 2008] where an instance $\boldsymbol{\mu}'$ is constructed such that it equals $\boldsymbol{\mu}$ for all time steps except on a period of length H, as illustrated on Figure 1.4. During those H steps, the suboptimal arm from $\boldsymbol{\mu}$ becomes optimal. By tuning the length of H inversely proportional to the expected number of pulls of the suboptimal arm from the instance $\boldsymbol{\mu}$, they obtain an equivalent of the aforementioned theorem. The trade-off is natural, the less frequent the suboptimal arm in $\boldsymbol{\mu}$ was pulled, the longer it will take for the learner to notice the change on the instance $\boldsymbol{\mu}'$. [Seznec et al., 2020, Proposition 4] extends Theorem 1.12 and shows that for any policy π there exists a piece-wise stationary environment with Γ_T breakpoints such that the regret satisfies:

$$\mathcal{R}(T,\pi) \geq \sqrt{KT\Gamma_T}\,. \tag{1.10}$$

Consequently, the best worst-case regret guarantee one can expect in abruptly changing environment with Γ_T breakpoints is of order $\mathcal{O}(\sqrt{KT\Gamma_T})$.

Two main mechanisms have been used to develop non-stationary bandit algorithms: strategies relying on change-point detectors and passively forgetting strategies that discard old data. Let us first briefly describe what change point detection is.

Change-Point Detection. The objective when using change-point detectors is to actively detect when a change in a distribution happens as fast and as accurately as possible. This implies being able to detect whether or not several changes might have occurred and to identify the times

of any such changes. Usually, the parameters in the change point detector have to be tuned for finding the trade-off between the *false alarm rate* (falsely detecting a change), the *misdetection rate* (missing a true change) and the *detection delay*. In Figure 1.5, we illustrate an example of time series data commonly used in change detection.

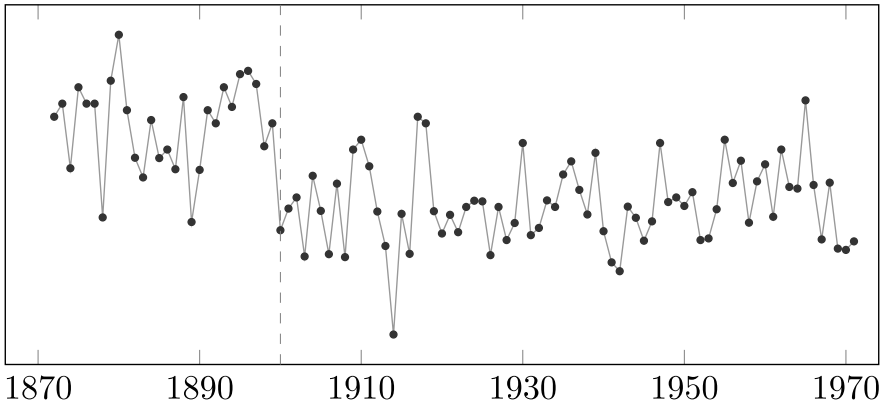


Figure 1.5: Yearly volume of the Nile river at Aswan. Dotted line denotes a detected change point. Extracted from `https://en.wikipedia.org/wiki/Change_detection`

Actively Adaptive Strategies. Algorithms based on change-point detectors contain three main components: (1) a bandit algorithm (any policy that was designed for stationary environments), (2) a change-point detector and (3) a fixed forced exploration rate. When a change is detected, the algorithm is restarted and the learner estimates the different quantities from scratch. The forced exploration is required. Indeed, without additional assumption on the structure of the non-stationarity, it is the only way to detect when a suboptimal arm becomes optimal after a breakpoint. Depending on the statistical test used for the change-point detector, several algorithms have been proposed. [Hartland et al., 2006, Srivastava et al., 2014] use a Page-Hinkley [Hinkley, 1971] algorithm for detecting the changes with the algorithms Adapt-EvE and SW-UCL respectively. [Liu et al., 2018] uses the cumulative sum (CUSUM) [Page, 1954] as a detector and obtains a regret of order $\mathcal{O}(\sqrt{KT\Gamma_T \log T})$ for CUSUM-UCB. GLR-klUCB proposed in [Besson et al., 2020] relies on sequential Generalized Likelihood Ratio Test for detecting the changes with a regret of order $\mathcal{O}(\sqrt{K\Gamma_T T \log T})$. [Cao et al., 2019] propose to compare running sample means over a sliding window for the change-point detection procedure. The algorithm M-UCB resulting from this procedure has the same regret as CUSUM-UCB. To some extent, all of those change-point detectors require scanning previously collected data and are computationally expensive. Furthermore, they are highly dependent on the nature of the non-stationarity and do not generalize well to smoothly changing environments. For these reasons, in this thesis we mostly focus on passively adaptive strategies that are computationally more attractive and more flexible. We describe them now.

Passively Adaptive Strategies. Instead of actively looking for a change in the rewards' distribution, another line of works has developed algorithms that passively discard old data considered as irrelevant for the estimation procedure. [Kocsis and Szepesvári, 2006b] suggest the use of discount factors for estimating the mean of the different arms with Discounted-UCB

(D-UCB). In this case, when using a discount factor γ the arm k at time t is estimated by:

$$\hat{\mu}_k(t) = \frac{\sum_{s=1}^{t-1} \gamma^{t-1-s} \mathbb{1}(A_s = k) X_s}{\sum_{s=1}^{t-1} \gamma^{t-1-s} \mathbb{1}(A_s = k)} \,. \tag{1.11}$$

[Garivier and Moulines, 2008, Garivier and Moulines, 2011] analyzes D-UCB with a regret of order $\mathcal{O}(K\sqrt{\Gamma_T T \log T})$ with a well-selected discount factor. They also propose SW-UCB based on a sliding window for approximating the means following:

$$\hat{\mu}_k(t) = \frac{\sum_{s=t-\tau}^{t-1} \mathbb{1}(A_s = k) X_s}{\sum_{s=t-\tau+1}^{t-1} \mathbb{1}(A_s = k)} \,. \tag{1.12}$$

When properly tuning the sliding window τ, they obtain a regret guarantee of order $\mathcal{O}(K\sqrt{\Gamma_T T \log T})$. The use of discount factors was also extended to Thompson Sampling in [Raj and Kalyani, 2017]. More recently, [Trovo et al., 2020] proposed SW-TS where Thomspon sampling is coupled with the use of a sliding window.

> In Chapter 3, we build upon the sliding window ideas from [Garivier and Moulines, 2008] and propose a subsampling method combined with a sliding window whose guarantees match the lower bound from Equation 1.10 when Γ_T is known. In addition to the traditional exploration-exploitation trade-off that a learner is facing in a sequential task, non-stationarity brings another trade-off with the tension between *remembering* old data for a better estimation of the means or *forgetting* data for keeping track of evolving means. We quantify at which rate the learner should forget past information for obtaining optimal strategies.

Knowledge of Γ_T. Except [Besson et al., 2020], all the methods discussed in the previous paragraphs require some information about the non-stationarity of the environment such as an upper-bound on the number of changes. Note that even when the number of breakpoints are known, none of those methods know when the breakpoints happen. One drawback of those approaches is that if the number of changes is not set correctly, the performance of the resulting algorithms might be significantly impacted. For this reason, some recent works aim at obtaining optimal regret guarantees without the knowledge of the number of changes [Chen et al., 2019, Auer et al., 2019]. Yet, those algorithms can not be used in practical scenarios (for the moment) or relies on strong assumption on the distribution of the arms and/or the form of the changes. For example [Mukherjee and Maillard, 2019, Komiyama et al., 2021] assume that all the arms change in a coordinated way. They show that the forced exploration can be canceled in this case and obtain guarantees of order $\mathcal{O}(\sqrt{T\Gamma_T})$ without knowing Γ_T. [Besson et al., 2020] requires Bernoulli distributions and additional assumption for the detectability of the gaps for obtaining similar guarantees.

1.3.2 Variation Budget

An alternative way to model non-stationarity is to introduce the variation budget and to restrict the total amount of variation of the means. A non-stationary instance can be characterized by the means of the different arms at the different time steps $\boldsymbol{\mu} = (\mu_{k,t})_{t \leq T, k \leq K}$. Formally, one can consider the instances satisfying:

$$\mathcal{B}_T(\boldsymbol{\mu}) \coloneqq \sum_{t=1}^{T-1} \max_{k \in \{1,\dots,K\}} |\mu_{k,t+1} - \mu_{k,t}| \leq B_T \,.$$

$\mathcal{B}_T(\boldsymbol{\mu})$ defines the true variation budget of an instance characterized by its mean $\boldsymbol{\mu}$ whereas B_T is an upper-bound for it. This setting allows for a rich class of non-stationary environments. In particular, it contains smoothly changing environments with a small drift for the means at every step [Krishnamurthy and Gopalan, 2021] with the constraint $\forall k \in \{1, \dots, K\}, \forall t \leq T, |\mu_{k,t+1} - \mu_{k,t}| \leq b$ but also abruptly changing environments as long as the number of breakpoints multiplied by the amplitude of the changes remain bounded by B_T.

The variation budget metric is more general. In some practical scenarios the environment is more likely to change smoothly, and obtaining guarantees for those environments is desirable. A lower bound established in [Besbes et al., 2014] quantifies the difficulty of learning in such environments similarly than in abruptly changing environments.

Theorem 1.13 (Theorem 1 in [Besbes et al., 2014])**.** *Assume that rewards have a Bernoulli distribution. Then, there is some absolute constant $C > 0$ such that for any policy π and for any $T \geq 1, K \geq 2$ and $B_T \in [1/K, T/K]$, the regret satisfies*

$$\mathcal{R}(T, \pi) \geq C \left(K B_T\right)^{1/3} T^{2/3} .$$

An important consequence of Theorem 1.13 is that when the variation budget scales linearly with respect to T the regret grows linearly and no learning is possible.

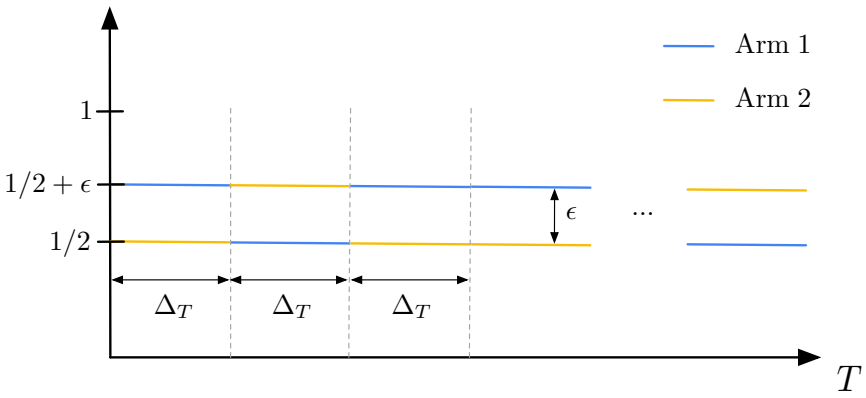


Figure 1.6: Example of instance used for obtaining the lower-bound for the variation budget setting when $K = 2$. For each block, the optimal arm is randomly selected in $\{1, 2\}$.

We now give a high level overview of the proof of this theorem. The entire time horizon T is separated in T/Δ_T blocks of length Δ_T. For each of this block an optimal arm with mean $\mu^\star = 1/2 + \epsilon$ is uniformly drawn from $\{1, \dots, K\}$ while the $K - 1$ remaining arms have a mean $\mu = 1/2$. The situation is illustrated on Figure 1.6 in the particular case $K = 2$. The identity of the best arm can only change at each block. For a fixed variation budget B_T, we have the constraint that $\mathcal{B}_T = \epsilon \times T/\Delta_T \leq B_T$ must be satisfied (the maximum gap for an arm between two consecutive blocks is ϵ). ϵ is tuned such that no policy can identify the best arm within a block. It can be shown that $\epsilon \approx 1/\sqrt{\Delta_T}$ is enough for ensuring this. With this choice, in the worst case, a constant regret of order ϵ can be suffered at every round within a block. This gives a regret of order $\epsilon \Delta_T \approx \sqrt{\Delta_T}$ for a block and of order $T/\sqrt{\Delta_T}$ for the entire horizon. The final step consists in tuning Δ_T for ensuring that $\epsilon \times T/\Delta_T \approx T/(\Delta_T)^{3/2} \leq B_T$ is satisfied. Taking $\Delta_T = B_T^{-2/3} T^{2/3}$ guarantees this and gives a regret of order $\mathcal{O}(T^{2/3} B_T^{1/3})$ as announced.

Overall, change-point detection based algorithms do not perform optimally in smoothly changing environments because they are designed to achieve strong guarantees on a stationary phase and to detect changes effectively. In some drifting environments, there is no stationary phase and detecting the change can be much harder. Relatively few works have tried to extend change-point detectors to smoothly changing environments. [Karnin and Anava, 2016] combine two statistical tests and obtain a suboptimal regret guarantee of order $\mathcal{O}(\log(T)T^{0.82}\mathcal{B}_T^{0.18}+\log(T)T^{0.771})$ without knowing $\mathcal{B}_T$. Most of existing methods that can be analyzed in this more general setting rely either on adversarial bandit algorithms and/or on passively forgetting strategies. [Besbes et al., 2014] propose Rexp3 an adaptation of the Exp3 algorithm [Auer et al., 2002b] with periodic restart and obtain regret guarantees of order $\mathcal{O}\left((K\log(K)B_T)^{1/3}T^{2/3}\right)$. [Wei et al., 2016] use periodic restarts combined with a UCB algorithm and propose Rerun-UCB-V. [Cheung et al., 2019] extend the analysis of SW-UCB to the variation budget setting and obtain order optimal bounds with respect to T and B_T.

In Chapter 4, we extend the discount factor principle from D-UCB to the variation budget. This is another advantage of passively forgetting strategies, they can be analyzed in more general non-stationary settings without any modification as long as some knowledge on the variation budget B_T is available.

1.3.3 Other Forms of Non-Stationarity

Here, we mention other non-stationary settings that are not discussed in this thesis. In some scenarios, it is natural to assume that the mean of an arm evolves only when the arm is played. Those are called rested bandits. For example in [Kleinberg and Immorlica, 2018, Pike-Burke and Grunewalder, 2019] when an arm is not played its mean will augment.

For a recommender system, a product is less interesting to a specific user once this user has bought it and this product should probably not be proposed in the close future. This is a case where the rewards of the different options depend on the algorithm choices. Sometimes, it is further assumed that the mean of an arm can only decrease when the arm is played. This setting is called rotting bandits and have been precisely studied in [Seznec et al., 2019, Seznec et al., 2020]. Quite interestingly, in [Seznec et al., 2019] the authors show that this non-stationary setting is not harder than stationary stochastic bandits. They design FEWA an algorithm with a problem-dependent regret guarantees of order $\mathcal{O}(\log(KT))$ and a worst case regret of order $\mathcal{O}(\sqrt{KT})$. They obtain these bounds without any knowledge of the decay function of the rotting bandit.

In the restless setting the means of the arms evolve independently from the algorithm choices. Abruptly changing environment and the variation budget discussed in the previous sections are special cases of restless bandits, however other variants have also been studied. For example in [Chen et al., 2020, Tracà et al., 2021] an additional seasonality in the reward distributions is assumed.

All the environments presented in this section are *unstructured* in the sense that by playing arm k nothing can be learned on the arm k' for $k' \neq k$. In the following sections, we consider richer *structured models* where it is sometimes possible to estimate precisely the reward from an action without even playing it.

1.4 Linear Contextual Bandits

Recalling our example for the mosquitoes repellent, a few elements have to be highlighted. First, depending on the mosquito species the effectiveness of a repellent might differ significantly. Second, depending on the characteristics of a user (age, tolerance to certain molecules, etc.) some repellents might be much more attractive than others. Algorithms presented in the previous sections can not use this sort of side information. Furthermore, if an efficient repellent has been found for a given species, this information might guide the learner to find a good repellent for a close species. Adding more structure to the multi-armed bandit model is tempting for generalizing from one set of characteristics to another. *Contextual bandit algorithms* have been considered for all of those reasons.

1.4.1 Stationary Linear Bandits

1.4.1.1 From Stochastic Contextual Bandits to Stochastic Linear Bandits

Unlike in the multi-armed bandit problem, at time t the learner receives a context c_t from a set of possible contexts $\mathcal{C}$. After seing this context, the learner picks an action u_t from an action set $\mathcal{U}$. In the stochastic contextual bandit setting, the rewards are assumed to satisfy:

$$X_t = r(c_t, u_t) + \eta_t \ .$$

Here $(\eta_t)_{t\geq 1}$ are assumed to be conditionally on the past σ-subgaussian (see Definition 1.14). r is a reward function that takes a context and an action as input. In online advertising, c_t would typically contain information about the user to be served whereas $\mathcal{U}$ would represent some features for the different ads that are available. It is usually convenient to assume that *contextualized actions* can be built from the different pairs of actions and contexts and to add some structure on the reward model. We assume that there are functions f and ϕ and an unknown vector $\theta^\star \in \mathbb{R}^d$ such that:

$$\forall c \in \mathcal{C}, \forall u \in \mathcal{U},\ r(c,u) = f(\langle \theta^\star, \phi(c,u)\rangle) \ .$$

Assuming such a structure and denoting $\mathcal{A}_t$ the different contextualized actions with $A_t = \phi(c_t, u_t)$ naturally brings to the study of a reward model of the form:

$$X_t = f(\langle \theta^\star, A_t\rangle) + \eta_t \ .$$

Depending on the assumption on the function f several models have been considered in literature. In Chapter 4 and in this section, we consider the case where there exists $\theta \in \mathbb{R}^d$ such that $f(x) = x$ a setting called linear contextual bandit and first studied in [Auer, 2002]. When $f(x) = \mu(x)$ with μ some link function, the setting is called generalized linear bandit model and was introduced by [Filippi et al., 2010]. This setting is studied in Chapter 5 and Chapter 6.

1.4.1.2 Reward Model

The reward model we consider is the following. At each time step t the learner receives $\mathcal{A}_t$ a set of K actions $(A_{t,1}, \dots, A_{t,K})$ lying in a d-dimensional space. We denote $a^\top$ the transpose of a vector a. The dot product between two vectors a and b is equal to $a^\top b$. By selecting the action A_t based on previous information, the learner receives a reward X_t satisfying

$$X_t = A_t^\top \theta^\star + \eta_t \ . \tag{1.13}$$

The parameter $\theta^\star \in \mathbb{R}^d$ is an unknown vector that has to be learned for finding the optimal action $A_t^\star$ at each step defined by:

$$A_t^\star = \text{argmax}_{a \in \mathcal{A}_t} a^\top \theta^\star \ .$$

Note that contrarily to the multi-armed bandit setting, even in stationary environments the best action can differ at every round because the learner receives **time-dependent set of actions**. Coming back to the online advertising example, allowing for time dependent action sets is necessary if we want to allow the candidate ads to differ from one user to another or to model the unavailability of ads for particular instants.

At round t the information collected so far can be summarized with a σ-field defined by $\mathcal{F}_t = \sigma(\mathcal{A}_1, A_1, \eta_1, \ldots, \mathcal{A}_t, A_t, \eta_t, \mathcal{A}_{t+1}, A_{t+1})$. $\mathcal{F}_t$ contains the information available at the end of round t in addition to the action selected at round $t+1$. Using this definition, η_t is $\mathcal{F}_t$-measurable and A_t is $\mathcal{F}_{t-1}$-measurable. We add A_t in $\mathcal{F}_{t-1}$ because the noise η_t might depend on the most recent action A_t. Additional assumptions on the noise are usually made, conditioned on the past the noise is centered i.e.

$$\mathbb{E}[\eta_t | \mathcal{F}_{t-1}] = 0 \quad \text{a.s.}$$

and the noise is conditionally σ-subgaussian.

Definition 1.14. A centered $\mathcal{F}_{t-1}$-measurable random variable η_t is said conditionally σ-subgaussian when

$$\forall \lambda \in \mathbb{R}, \ \mathbb{E}[\exp(\lambda \eta_t) | \mathcal{F}_{t-1}] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right) \quad \text{a.s.}$$

1.4.1.3 Definition of the Regret

In this setting, the performance of a policy π for an instance $\theta^\star$ is evaluated with a quantity usually referred to as pseudo-regret (that we will simply call regret) defined by:

$$R_{\theta^\star}(T, \pi) \coloneqq \sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} a^\top \theta^\star - \sum_{t=1}^{T} A_t^\top \theta^\star \ .$$

The regret is a random variable because of the term A_t which depends on the previous choices of the algorithm. Note that the comparison is done with $\sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} a^\top \theta^\star$ which is on average the best cumulative reward a learner can expect if he knows ahead of the game the true parameter $\theta^\star$ of the environment. We will also consider the expected regret that is defined as:

Definition 1.15 (Regret)**.** For a policy π, a linear bandit with an unknown vector $\theta^\star \in \mathbb{R}^d$ and a T rounds interaction with the environment, the expected regret is defined as

$$\mathcal{R}_{\theta^\star}(T, \pi) \coloneqq \mathbb{E}_{\theta^\star}[R_{\theta^\star}(T, \pi)] = \sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} a^\top \theta^\star - \mathbb{E}\left[\sum_{t=1}^{T} A_t^\top \theta^\star\right] \ .$$

Link to the multi-armed bandit. Note that the multi-armed bandit setting with K arms is recovered when $d = K$ and when the action set at every time step is equal to $\mathcal{A}_t = \{e_1, \ldots, e_d\}$ where e_i is the i-th canonical vector from $\mathbb{R}^d$ containing a one at the i-th coordinate and zero

elsewhere. This link has strong consequences for the regret guarantees as it implies that the worst case regret for linear bandits is at least of the order of the worst case regret from the multi-armed bandit.

1.4.1.4 Worst Case Regret

Similarly to the case of the multi-armed bandits, for any policy π, one can obtain a worst case regret lower bound by creating hard instances that aim at fooling the learner. Knowing that the linear setting is strictly more general than the non contextual case, we know that the worst case regret scales at least as $\sqrt{T}$ with respect to the time horizon. Yet, it remains unclear how it should scale with respect to the dimension d or the number of actions. The following theorem shows that a lower bound can be obtained even with an infinite action set.

Theorem 1.16 (Theorem 24.2 in [Lattimore and Szepesvári, 2020])**.** *Assume $d \leq 2T$ and assume that the action set is constant and equal to $\mathcal{A} := \{a \in \mathbb{R}^d \mid \|a\|_2 \leq 1\}$. For any policy π, there exists $\theta^\star \in \mathbb{R}^d$ such that*

$$\mathcal{R}_{\theta^\star}(T, \pi) \geq \frac{d\sqrt{T}}{16\sqrt{3}} .$$

We knew that the scaling in T should be at least of order $\sqrt{T}$. This bounds does not indicate that the scaling should be worse than this. Another interesting aspect is that the bound was obtained using an infinite action set. This suggests that the scaling in K should be rather mild when the number of actions is finite. By adding structure on reward model, the dependency in the number of arms is significantly reduced compared to the multi-armed bandit where the worst case regret was scaling as the square root of the number of arms (see Theorem 1.4). This comes at a cost which is the linear dependency in the dimension parameter d. Yet, for practical examples d is usually much smaller than K and reducing the dependency of K at the cost of an increase in the dependency of d remains desirable.

Linear bandits have been extensively studied in literature with for example [Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010, Abbasi-Yadkori et al., 2011] or [Chu et al., 2011]. As in the non-contextual case, the most popular methods for linear bandits rely either on upper-confidence bounds or on Thompson sampling. We only discuss how to obtain upper-confidence bounds in this setting and we refer to [Agrawal and Goyal, 2013b, Abeille and Lazaric, 2017] for an adaptation of Thompson Sampling to linear bandits.

1.4.1.5 Upper-confidence Bounds and Linear Bandits

A natural way for estimating $\theta^\star$ at time t using previous information is to consider the regularized least squares estimator defined as:

$$\hat{\theta}_t := \operatorname{argmin}_{\theta \in \mathbb{R}^d} \left(\sum_{s=1}^{t} (X_s - A_s^\top \theta)^2 + \lambda \|\theta\|_2^2 \right) . \tag{1.14}$$

This optimization program admits a closed form solution with:

$$\hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} A_s X_s \quad \text{where} \quad V_t = \sum_{s=1}^{t} A_s A_s^\top + \lambda I_d \, , \tag{1.15}$$

and I_d denotes the d-dimensional identity matrix.

An important characteristic in the multi-armed bandit model is that the UCBs are disjoint, i.e the upper-confidence bound from arm k carries no information about the other arms $k' \neq k$. The story is different in the linear setting where the different directions of $\mathbb{R}^d$ are interconnected through the dot product with $\theta^\star$. We now explain the form that takes the UCB in the linear setting. The objective is to find a confidence region $\mathcal{C}_t(\delta)$ depending only on $A_1, \dots, A_{t-1}$ and on $X_1, \dots, X_{t-1}$ that contains the true parameter $\theta^\star$ with high probability. One of the tightest UCB based on self-normalized tail inequality for vector-values martingales was proposed in [Abbasi-Yadkori et al., 2011] and has the following form.

Theorem 1.17 (Theorem 2 in [Abbasi-Yadkori et al., 2011])**.** *Assume that $\|\theta^\star\|_2 \leq S$, that for all $a \in \mathcal{A}_t, \|a\|_2 \leq L$ and that the noise η_t is conditionally σ-subgaussian. Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $t \geq 0$,*

$$\theta^\star \in \mathcal{C}_t(\delta) := \left\{ \theta \in \mathbb{R}^d \;\middle|\; \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}} \leq \sqrt{\lambda} S + \sigma \sqrt{d \log\left(\frac{1 + tL^2/\lambda}{\delta}\right)} \right\} .$$

Note that $\mathcal{C}_t(\delta)$ is an ellipsoid centered around $\hat{\theta}_{t-1}$ and with a radius $\beta_t(\delta)$ defined as $\beta_t(\delta) := \sqrt{\lambda} S + \sigma \sqrt{d \log\left(\frac{1+tL^2/\lambda}{\delta}\right)}$.

Optimism in the Face of Uncertainty for Linear Bandits (OFUL). Once a confidence region is obtained for the regression parameter $\theta^\star$, optimistic algorithms apply the optimism in the face of uncertainty principle by selecting the action:

$$A_t = \text{argmax}_{a \in \mathcal{A}_t} \max_{\theta \in \mathcal{C}_t(\delta)} a^\top \theta \; . \tag{1.16}$$

Among the statistically plausible values for $\theta^\star$ contained in $\mathcal{C}_t(\delta)$, an optimistic algorithm selects the most valuable action.

For a $d \times d$ positive semi-definite matrix M and $x \in \mathbb{R}^d$ we define the norm $\|x\|_M := \sqrt{x^\top M x}$ which is non-negative by definition. Note that once the confidence region is obtained, the learner can estimate the reward in every direction $a \in \mathcal{A}_t$ and the following can be deduced from Theorem 1.17

$$\forall a \in \mathcal{A}_t, \; a^\top \theta^\star \leq a^\top \hat{\theta}_{t-1} + \beta_t(\delta) \|a\|_{V_{t-1}^{-1}} \; . \tag{1.17}$$

Equation (1.17) gives an upper-bound on the mean reward the learner can expect by playing action $a \in \mathcal{A}_t$ at time t. With the specific form of the confidence region from Theorem 1.17 (an ellipsoid) Equation (1.16) can be rewritten and the maximum for θ in $\mathcal{C}_t(\delta)$ can be computed. The selection rule in this case is:

$$A_t = \text{argmax}_{a \in \mathcal{A}_{t+1}} a^\top \hat{\theta}_{t-1} + \beta_t(\delta) \|a\|_{V_{t-1}^{-1}} \; . \tag{1.18}$$

1.4.1.6 A Minimax Optimal Algorithm

We now have all the ingredients for building an algorithm relying on the confidence region $\mathcal{C}_t(\delta)$ that we have obtained. [Abbasi-Yadkori et al., 2011] propose OFUL an algorithm leveraging this confidence region that we report in Algorithm 3. This algorithm can also be seen under

the name LinUCB as it shares strong similarities with the algorithm LinUCB proposed in [Chu et al., 2011] except that the exploration term in OFUL contains $\beta_t(\delta)$ the radius of the confidence ellipsoid whereas in LinUCB this quantity is denoted α and is an input of the algorithm. [Abbasi-Yadkori et al., 2011] obtain strong guarantees for this algorithm.

Input: Failure probability δ, subgaussianity constant σ, dimension d, regularization λ, upper-bound for actions L, upper-bound for parameters S
Initialization: $V_1 = \lambda I_d$ and $\hat{\theta}_1 = 0_{\mathbb{R}^d}$
for $t \leq T$ **do**
  Receive $\mathcal{A}_t$
  Play action $A_t = \text{argmax}_{a \in \mathcal{A}_t} a^\top \hat{\theta}_{t-1} + \beta_t(\delta)\|a\|_{V_{t-1}^{-1}}$
Receive reward X_t
Updating phase: Compute $\hat{\theta}_t$ and update the confidence region $\mathcal{C}_{t+1}(\delta)$

Algorithm 3: OFUL [Abbasi-Yadkori et al., 2011]

Theorem 1.18 (Theorem 3 in [Abbasi-Yadkori et al., 2011]). *Assume that $\|\theta^\star\|_2 \leq S$, that $\forall a \in \mathcal{A}_t, \|a\|_2 \leq L$ but also that $\forall a \in \mathcal{A}_t, |a^\top \theta^\star| \leq 1$. Then, with probability at least $1-\delta$, the regret of OFUL satisfies*

$$\forall T,\ R_{\theta^\star}(T) \leq 4\sqrt{dT \log(\lambda + TL/d)}\beta_T(\delta)\ .$$

Using this theorem and by observing that the radius of the ellipsoid $\mathcal{C}_t(\delta)$ scales as the square root of the dimension, we easily obtain that under the same assumptions as Theorem 1.18 the regret of OFUL scales as:

$$\mathcal{R}_{\theta^\star}(T) = \widetilde{\mathcal{O}}(d\sqrt{T})\ , \tag{1.19}$$

where $\widetilde{\mathcal{O}}$ hides polylogarithmic terms depending on T. Up to logarithmic terms, this regret is of the same order as the worst case regret established in Theorem 1.16 which explains why OFUL is optimal in the class of instances defined by the constraints $\|\theta^\star\|_2 \leq S$ and $\forall a \in \mathcal{A}_t, \|a\|_2 \leq L$.

A typical application of bandit algorithms based on the linear model is online recommendation where actions are items to be, for instance, efficiently arranged on personalized web pages to maximize some conversion rate. However, it is unlikely that customers' preferences remain stable and the collected data becomes progressively obsolete as the interest for the items evolve. Hence, it is essential to design adaptive bandit agents rather than restarting the learning from scratch on a regular basis. This motivates the study of non-stationary linear bandits.

1.4.2 Non-Stationary Linear Bandits

Similarly to the multi-armed bandit, a non-stationary component can be added to the model. In the linear bandit model this takes the form of a time-dependent regression parameter $\theta_t^\star$. In this case, the reward obtained by selecting action A_t at time t satisfies:

$$X_t = \langle A_t, \theta_t^\star \rangle + \eta_t\ . \tag{1.20}$$

A learner that knows the exact evolution of the sequence $(\theta_t^\star)_{t \leq T}$ would obtain an expected cumulative reward of order $\sum_{t=1}^T \max_{a \in \mathcal{A}_t} a^\top \theta_t^\star$. For this reason, the dynamic regret in non-stationary environments is defined as $R(T, \pi) = \sum_{t=1}^T \max_{a \in \mathcal{A}_t} a^\top \theta_t^\star - A_t^\top \theta_t^\star$. Again, this is a

random quantity and we obtain the regret by taking the expectation over the choices of the algorithm to obtain:

Definition 1.19. For a policy π, a linear bandit with unknown vectors $(\theta_t^\star)_{t\leq T} \in \mathbb{R}^d$ for and a T rounds interaction with the environment, the dynamic expected regret is defined as

$$\mathcal{R}(T,\pi) \coloneqq \sum_{t=1}^{T} \max_{a\in\mathcal{A}_t} a^\top \theta_t^\star - \mathbb{E}\left[\sum_{t=1}^{T} A_t^\top \theta_t^\star\right] .$$

1.4.2.1 Abruptly Changing Environments

Compared to the multi-armed bandit model relatively few works have considered non-stationary linear bandits. Without surprise some structure on the non-stationarity is again necessary. By analogy with the multi-armed bandit model, some works consider abruptly changing environments where the regression vector $\theta_t^\star$ is only allowed to change Γ_T times. Mathematically, this means considering instances satisfying:

$$\sum_{t=1}^{T-1} \mathbb{1}\left(\theta_t^\star \neq \theta_{t+1}^\star\right) \leq \Gamma_T \, . \tag{1.21}$$

An interesting example of abruptly changing environment is music taste which can change really fast for specific period of the year. For example the taste of some users change quite drastically around Christmas or Halloween but come back to normal later. Any music recommendation system that fails to model those potential changes will make suboptimal recommendation during this period.

Detecting a change in the linear setting can be tedious and the geometry of the arms set $(\mathcal{A}_t)_t$ can make the detection even more complicated. Some works still follow this path and propose change-point detection strategies for linear bandits. [Hariri et al., 2015] use the CUSUM change-point detector and restart the learning when the learner is confident enough about a change. [Wu et al., 2018] assume the action set $\mathcal{A}$ remains static over time and build a pool of LinUCB learners called *slave models* as experts. A new model is added to the pool when no existing slave is able to give good prediction, that is, when a change is detected. [Di Benedetto et al., 2020] build on [Wu et al., 2018] and propose an algorithm for seasonal environments for which $\theta^\star$ is abruptly changing but where there is a periodicity of the non-stationary environments. [Ding et al., 2020] build a multiscale change-point detector and propose Multiscale-LinUCB that can be analyzed with time-dependent action sets. [Xu et al., 2020] use a change-point detector based on a sliding window and show good empirical performance on real-world data. However, the limitation of such approaches is that they can not adapt to some slowly-varying environments that we describe now.

1.4.2.2 Variation Budget

Quantifying the non-stationarity with the variation budget gives a more general framework for studying non-stationary linear bandits. Under this model, we consider only the instances $(\theta_t^\star)_{t\leq T}$ satisfying

$$\mathcal{B}_T \coloneqq \sum_{t=1}^{T-1} \|\theta_{t+1}^\star - \theta_t^\star\|_2 \leq B_T \, . \tag{1.22}$$

One of the advantage of this non-stationary structure is to encapsulate both abruptly changing environments and environments in which $\theta_t^\star$ drifts slowly. When dealing with user preferences, there is a no prior reason to believe that the preferences should change abruptly. Having this additional degree of freedom on the non-stationarity might be of interest for practical scenarios.

Once again, since linear bandits are more general than their multi-armed bandits counterpart, we know that the worst case regret in this setting should scale at least with $B_T^{1/3}T^{2/3}$. [Cheung et al., 2021] establishes a lower bound tailored to non-stationary linear bandits.

Theorem 1.20. *In the drifting linear bandit setting, for any $T \geq d$ and $B_T \in [d/\sqrt{T}, 8T/d^2]$, there exists a sequence of action sets $(\mathcal{A}_t)_{t\leq T}$ and parameters $(\theta_t^\star)_{t\leq T}$ and some absolute constant C such that the dynamic regret for any non-anticipatory policy π satisfies*

$$\mathcal{R}(T,\pi) \geq Cd^{2/3}B_T^{1/3}T^{2/3} \ .$$

This results suggests that the dependency in B_T and T remains the same than for non contextual bandits but that instead of scaling with $K^{1/3}$, in the linear setting only the dimension d matters with an order $d^{2/3}$.

Passively forgetting strategies were also developed for this setting. [Cheung et al., 2019, Cheung et al., 2021] extended the sliding window forgetting principle to linear bandits propose SW-UCB which is based on the following least squares estimator:

$$\hat{\theta}_t \coloneqq \operatorname{argmin}_{\theta\in\mathbb{R}^d} \left(\sum_{s=t-\tau+1}^{t} (X_s - A_s^\top\theta)^2 + \lambda\|\theta\|_2^2 \right) \ . \tag{1.23}$$

[Zhao et al., 2020] proposed an even simpler method with RestartUCB where the learning is restarted every H steps and H is tuned based on the knowledge of B_T. They show that similar guarantees than SW-UCB can be obtained. Both methods initially established regret bounds of order $\widetilde{\mathcal{O}}(d^{2/3}B_T^{1/3}T^{2/3})$ with a known B_T thus matching the lower bound from Theorem 1.20 up to logarithmic terms. [Touati and Vincent, 2020] find out that all passively forgetting method in linear setting [Cheung et al., 2019, Russac et al., 2019, Zhao et al., 2020] have a technical gap in their analysis and that the regret bounds yield degraded rates of order $\widetilde{\mathcal{O}}(B_T^{1/4}T^{3/4})$ without further assumption on the action sets.

Even when additional assumption on the action sets is made, we stress out that the minimax-optimality of the passively forgetting policies is conditioned on the knowledge of an upper-bound B_T on the true parameter drift $\mathcal{B}_T$. Naturally, the tighter this upper-bound, the better the performance. Yet, whether such a knowledge is available in real-life problems is questionable. In the linear setting, [Cheung et al., 2019] circumvent this drawback with the Bandit-Over-Bandit (BOB) strategy where $\mathcal{B}_T$ is learned by a master algorithm that tunes the window size of SW-UCB adaptively but yields degraded rates. Other methods have been designed for obtaining guarantees without knowing $\mathcal{B}_T$ under different modeling assumption.

[Luo et al., 2018] propose a change point detector for contextual bandits in drifting environments and assuming that the context-reward pairs are drawn from $\mathcal{D}_t$ at time t and denoting $\bar{\mathcal{B}}_T = \sum_{t=1}^{T-1}\|\mathcal{D}_{t+1} - \mathcal{D}_t\|_{\mathrm{TV}}$ the sum of total variations between consecutive distributions, they obtain a regret bound of order $\widetilde{\mathcal{O}}(\bar{\mathcal{B}}_T T^{2/3})$ without any information on the non-stationarity of the environment. [Chen et al., 2019] build upon this work and propose ADA-ILTCB$^+$ and improve the regret guarantees with a scaling of order $\mathcal{O}(\bar{\mathcal{B}}_T^{1/3}T^{2/3})$ but without emphasis on actual practical performance.

In Chapter 4, we address sequential learning problems in which the parameter of the linear bandit is evolving with time. For doing this, we consider the weighted least-squares estimator defined as

$$\hat{\theta}_t \coloneqq \operatorname{argmin}_{\theta\in\mathbb{R}^d} \left(\sum_{s=1}^{t} w_{s,t}(X_s - A_s^\top \theta)^2 + \lambda \|\theta\|_2^2 \right) ,$$

as an efficient method to progressively forget past interactions. We design a new confidence region for this estimator where we detail the particular role of the weights and propose the D-LinUCB algorithm. With additional assumption on the geometry of the action sets, we recover the $\widetilde{\mathcal{O}}(B_T^{1/3}T^{2/3})$ (omitting the dependency in d) optimal regret bound. Without this assumption, we show a regret bound of order $\widetilde{\mathcal{O}}(B_T^{1/4}T^{3/4})$. All previously existing algorithms based on forgetting principle suffer from the same technical flaw that we discuss. In particular, existing algorithms suffer a regret $\widetilde{\mathcal{O}}(B_T^{1/4}T^{3/4})$ without additional assumption on the action sets contrarily to what was announced in [Cheung et al., 2019, Russac et al., 2019, Zhao et al., 2020].

The linear bandit framework has proven to be an important paradigm for sequential decision making under uncertainty. It notably extends the multi-armed bandit framework to address the exploration-exploitation dilemma when the arm-set is large (potentially infinite) or changing over time. Several extensions show that under appropriate algorithmic changes, existing linear bandit concepts can be leveraged to handle non-stationarity in the reward model. Perhaps the main limitation of linear bandits resides in their inability to model specific (e.g binary, discrete) rewards. One axis of research to operate beyond linearity was initiated with the introduction of generalized linear bandit, the topic of the following section.

1.5 Generalized Linear Bandits

1.5.1 Setting

When selecting action A_t at time t, the main difference with generalized linear bandits is the addition of a function μ on top of the dot product $A_t^\top \theta^\star$ to allow for non linear reward models. This framework was introduced in [Filippi et al., 2010] and allows to handle reward which can be expressed as a generalized linear model. Those include for instance the Poisson model and logistic model. At time t the learner receives a set of actions $\mathcal{A}_t$. The agent then selects an action $A_t \in \mathcal{A}_t$ and receives the stochastic reward in the form of:

$$\mathbb{E}[X_t|\mathcal{F}_{t-1}] = \mu(A_t^\top \theta^\star) , \tag{1.24}$$

where $\mathcal{F}_t = \sigma(\mathcal{A}_1, A_1, X_1, \ldots, \mathcal{A}_t, A_t, X_t, \mathcal{A}_{t+1}, A_{t+1})$ is the σ-field containing all information before obtaining the reward at time $t+1$.

The reward model assumes that the conditional distribution of the reward x given some feature vector a belongs to the canonical exponential family as presented in Equation (1.2), i.e there is a reference measure ξ such that the density conditioned on a is of the form

$$\frac{d\nu_{\theta_a}}{d\xi} = \exp\left(x\theta_a - b(\theta_a)\right) .$$

Under this reward model, simple computations show that

$$\mathbb{E}[X|a] = \dot{b}(\theta_a) \quad \text{and} \quad \mathrm{Var}(X|a) = \ddot{b}(\theta_a) .$$

We deduce from this that $\mu \coloneqq \dot{b}$. μ is a real-valued function that is assumed to be twice differentiable and from the equality featuring the variance we get that μ is strictly increasing. μ is most often referred to as the inverse link function. Finally, is it also assumed in a generalized linear model that θ_a has a convenient form:

$$\exists \theta^\star \in \mathbb{R}^d \text{ such that } \theta_a = a^\top \theta^\star \ .$$

Note that the generalized linear bandit framework contains both the linear bandit framework when $\mu(z) = z$ and also the logistic bandits when $\mu(z) = 1/(1+\exp(-x))$. Similarly to the linear bandits, an oracle that knows $\theta^\star$ ahead of time can obtain an average cumulative reward of order $\sum_{t=1}^T \mu(a_{t,\star}^\top \theta^\star)$ with $a_{t,\star} = \text{argmax}_{a \in \mathcal{A}_t} \mu(a^\top \theta^\star)$. The regret is then the random quantity defined as $R_{\theta^\star}(T, \pi) \coloneqq \sum_{t=1}^T \mu(a_{t,\star}^\top \theta^\star) - \sum_{t=1}^T \mu(A_t^\top \theta^\star)$. Following the analogy with the linear model, the regret is defined as:

Definition 1.21 (Expected Regret)**.** For a policy π, a generalized linear bandit with a unknown vector $\theta^\star \in \mathbb{R}^d$ and inverse link function μ and a T rounds interaction with the environment, the expected regret is defined as

$$\mathcal{R}_{\theta^\star}(T, \pi) \coloneqq \sum_{t=1}^T \max_{a \in \mathcal{A}_t} \mu(a^\top \theta^\star) - \mathbb{E}\left[\sum_{t=1}^T \mu(A_t^\top \theta^\star)\right] \ .$$

We give the main assumptions under which this model can be analyzed. We assume that the bandit parameter $\theta^\star$ has a bounded norm following $\|\theta^\star\|_2 \leq S$. Further, we assume that the actions have bounded norms: $\|a\|_2 \leq L$ for all $a \in \mathcal{A}_t$. We denote $\Theta = \{\theta : \|\theta\|_2 \leq S\}$ the set of admissible parameters and $\mathcal{A} = \{a : \|a\|_2 \leq L\}$. We assume that the quantities L, S are known to the agent. For a given inverse link function μ, we will follow the notation from [Filippi et al., 2010] and denote:

$$k_\mu = \sup_{a \in \mathcal{A}, \theta \in \Theta} \dot{\mu}(a^\top \theta), \qquad c_\mu = \inf_{a \in \mathcal{A}, \theta \in \Theta} \dot{\mu}(a^\top \theta) \ .$$

Note that in the linear case, we obtain $k_\mu = c_\mu = 1$.

1.5.2 The Particular Role of c_μ

In the linear setting the least squares estimator had a convenient closed form expression. In this more general setting, the penalized maximum likelihood estimator is defined as:

$$\hat{\theta}_t = \text{argmax}_{\theta \in \mathbb{R}^d} \sum_{s=1}^t \left[X_s A_s^\top \theta - b(A_s^\top \theta)\right] - \frac{\lambda}{2}\|\theta\|_2^2 \ ,$$

and needs to be computed numerically.

Upon differentiating this expression that is strictly concave in θ, $\hat{\theta}_t$ is the unique solution satisfying

$$\sum_{s=1}^t X_s A_s - \dot{b}(A_s^\top \hat{\theta}_t) A_s - \lambda \hat{\theta}_t = 0 \ .$$

For this reason, $\hat{\theta}_t$ is the unique $\theta \in \mathbb{R}^d$ that satisfies

$$\sum_{s=1}^{t} \mu(A_s^\top \hat{\theta}_t) A_s + \lambda \hat{\theta}_t = \sum_{s=1}^{t} X_s A_s \,. \tag{1.25}$$

The analysis provided in [Filippi et al., 2010] aims at using existing tools in the linear bandits by linearizing the reward signal. For understanding how this is done, let us upper-bound $|\mu(A_{t+1}^\top \hat{\theta}_t) - \mu(A_{t+1}^\top \theta^\star)|$. We assume for now that $\hat{\theta}_t \in \Theta$ and define $g_t(\theta) \coloneqq \sum_{s=1}^{t} \mu(A_s^\top \theta) A_s + \lambda\theta$. The Fundamental Theorem of Calculus guarantees the following:

$$\begin{aligned}
g_t(\theta^\star) - g_t(\hat{\theta}_t) &= \sum_{s=1}^{t} (\mu(A_s^\top \theta^\star) - \mu(A_s^\top \hat{\theta}_t)) A_s + \lambda(\theta^\star - \hat{\theta}_t) \\
&= \sum_{s=1}^{t} \int_{u=0}^{1} \dot{\mu}(u A_s^\top \theta^\star + (1-u) A_s^\top \hat{\theta}_t)\, du A_s A_s^\top (\theta^\star - \hat{\theta}_t) + \lambda(\theta^\star - \hat{\theta}_t) \\
&= \underbrace{\left(\sum_{s=1}^{t} \int_{u=0}^{1} \dot{\mu}(u A_s^\top \theta^\star + (1-u) A_s^\top \hat{\theta}_t)\, du A_s A_s^\top + \lambda I_d \right)}_{G_t} (\theta^\star - \hat{\theta}_t) \\
&= G_t(\theta^\star - \hat{\theta}_t) \,.
\end{aligned} \tag{1.26}$$

In particular, under the assumption that $\hat{\theta}_t \in \Theta$, the following lower bound holds

$$G_t \geq c_\mu V_t \quad \text{where} \quad V_t = \sum_{s=1}^{t} A_s A_s^\top + c_\mu^{-1} \lambda I_d \,.$$

This lower bound ensures that G_t is invertible and we can now upper-bound $|\mu(A_{t+1}^\top \hat{\theta}_t) - \mu(A_{t+1}^\top \theta^\star)|$ in the following way:

$$\begin{aligned}
|\mu(A_{t+1}^\top \theta^\star) - \mu(A_{t+1}^\top \hat{\theta}_t)| &\leq k_\mu |A_{t+1}^\top (\theta^\star - \hat{\theta}_t)| \\
&= k_\mu |A_{t+1}^\top G_t^{-1} (g_t(\theta^\star) - g_t(\hat{\theta}_t))| \quad \text{(Equation (1.26))} \\
&\leq k_\mu \|A_{t+1}\|_{G_t^{-1}} \|g_t(\hat{\theta}_t) - g_t(\theta^\star)\|_{G_t^{-1}} \quad \text{(Cauchy–Schwarz inequality)} \\
&\leq \frac{k_\mu}{c_\mu} \|A_{t+1}\|_{V_t^{-1}} \Big\| \sum_{s=1}^{t} \Big(X_s - \mu(A_s^\top \theta^\star) \Big) A_s - \lambda \theta^\star \Big\|_{V_t^{-1}} \,.
\end{aligned} \tag{1.27}$$

In the last inequality, we have used Equation (1.26), the characterization of the maximum likelihood estimator from Equation (1.25) and the assumption on the reward model. The interesting aspect of the previous upper-bound is that except the term k_μ/c_μ the right hand side from Equation (1.27) does not feature any term associated with the non-linearity of the reward distribution. All the remaining terms can actually be bounded following existing analysis in the linear setting by remarking that $X_s - \mu(A_s^\top \theta^\star)$ is conditionally on the past a zero mean noise term.

Based on this upper-bound, the analysis from [Abbasi-Yadkori et al., 2011] can be applied and [Filippi et al., 2010] show that the GLM-UCB algorithm that they propose has regret guarantees of order $\widetilde{\mathcal{O}}(\frac{k_\mu}{c_\mu} d\sqrt{T})$. At first sight, this suggests that learning in this more general setting is not harder than learning in the linear bandit setting and the only major difference comes from the constant term k_μ/c_μ. Note however, that in the logistic case for example, k_μ/c_μ scales

exponentially with S the upper-bound of the true parameter $\theta^\star$. This has consequences on the practical applicability of the generalized linear bandit algorithms that will perform poorly in the logistic case when S is large.

We underline that this is not a specificity of [Filippi et al., 2010] analysis and that most of existing algorithms for generalized linear bandits have a similar scaling [Abeille and Lazaric, 2017, Kveton et al., 2020, Li et al., 2017]. In the paper from [Filippi et al., 2010], the authors already suggest an asymptotic argument suggesting that a scaling with $k_\mu/\sqrt{c_\mu}$ could be obtained. It is only recently that a first paper managed to obtain a regret bound of this order with a new concentration inequality tailored to generalized linear bandits. We elaborate on this in Chapter 5 and extend the concentration inequality of [Faury et al., 2020] to more general estimators.

Projection Step. In the previous reasoning we have assumed $\hat{\theta}_t \in \Theta$. There is a priori no reason why this should hold and [Filippi et al., 2010] circumvent this issue by considering another estimator $\widetilde{\theta}_t$ defined as

$$\widetilde{\theta}_t \coloneqq \text{argmin}_{\theta\in\Theta}\|g_t(\theta) - g_t(\hat{\theta}_t)\|_{V_t^{-1}} \,. \tag{1.28}$$

By construction $\widetilde{\theta}_t \in \Theta$. We can replace $\hat{\theta}_t$ by $\widetilde{\theta}_t$ and use the following.

$$\begin{aligned}
|\mu(A_{t+1}^\top\theta^\star) - \mu(A_{t+1}^\top\widetilde{\theta}_t)| &\leq k_\mu |A_{t+1}^\top(\theta^\star - \widetilde{\theta}_t)| \\
&= k_\mu |A_{t+1}^\top G_t^{-1}(g_t(\theta^\star) - g_t(\widetilde{\theta}_t))| \\
&\leq k_\mu \|A_{t+1}\|_{G_t^{-1}} \|g_t(\theta^\star) - g_t(\widetilde{\theta}_t)\|_{G_t^{-1}} \,.
\end{aligned}$$

We fall back on our feet using:

$$\begin{aligned}
\|g_t(\widetilde{\theta}_t) - g_t(\theta^\star)\|_{G_t^{-1}} &\leq \|g_t(\widetilde{\theta}_t) - g_t(\hat{\theta}_t)\|_{G_t^{-1}} + \|g_t(\hat{\theta}_t) - g_t(\theta^\star)\|_{G_t^{-1}} \\
&\leq 2\|g_t(\hat{\theta}_t) - g_t(\theta^\star)\|_{G_t^{-1}} \,.
\end{aligned}$$

Given the remarkable importance and widespread use of generalized models in practice, ensuring their resilience to non-stationarity stands as a crucial milestone in the parametric bandit literature. We present the first attempts for building non-stationary generalized linear bandit models in the following section.

1.5.3 Extension to Non-Stationary Environments

In this section, we consider the generalized linear bandit models in non-stationary environments. The main difference compared to the stationary counterpart is that the environment starts by picking a sequence of parameters $(\theta_t^\star)_{t\leq T}$. The reward model is modified as follows:

$$\mathbb{E}[X_t|\mathcal{F}_{t-1}] = \mu(A_t^\top\theta_t^\star) \,. \tag{1.29}$$

Assumption 6.1 also need to be adapted by imposing that the entire sequence $(\theta_t^\star)_{t\leq T}$ lies in the admissible set Θ. The key quantities k_μ and c_μ are unchanged. Similarly than in the previous chapter, we focus on abruptly changing environments and on smoothly changing environments.

1.5.3.1 Abruptly Changing Environments and Generalized Linear Bandits

[Russac et al., 2020] is the first paper to consider generalized linear bandits under abruptly changing environments where the bandit instances considered satisfy the same constraint as in the linear setting i.e $\sum_{t=1}^{T-1} \mathbb{1}\left(\theta_t^\star \neq \theta_{t+1}^\star\right) \leq \Gamma_T$. In this initial work, the analysis from [Filippi et al., 2010] is adapted to deal with non-stationarity. The authors combine GLB-UCB with passively forgetting mechanisms and propose SW-GLUCB based on a sliding window together with D-GLUCB using a discount factor.

> In Chapter 5, we build upon [Russac et al., 2020] to obtain better regret guarantees (with respect to c_μ) for generalized linear bandits in abruptly changing environments. This is done by leveraging the concentration inequality from [Faury et al., 2020] under the assumption that the inverse link function is *self-concordant*, a notion that will be explained in Chapter 5. We propose SC-D-GLUCB an algorithm with a regret guarantee of order $\widetilde{\mathcal{O}}(c_\mu^{-1/2} d\sqrt{\Gamma_T T})$ in any abruptly changing environments with at most Γ_T breakpoints.

1.5.3.2 Variation Budget and Generalized Linear Bandits

The study of generalized linear bandits where non-stationarity is characterized through the variation budget (Equation (1.22)) has received more interest. In this case, the variation budget is defined as in Equation (1.22) and only the instances satisfying $\mathcal{B}_T \coloneqq \sum_{t=1}^{T-1} \|\theta_{t+1}^\star - \theta_t^\star\|_2 \leq B_T$ are kept. All existing works rely on the same conceptual approach, i.e addressing the reward's drift by progressively forgetting past data [Cheung et al., 2021, Zhao et al., 2020]. This is achieved by maintaining estimation based on a truncated history of the data, judging that old observations no longer carry valuable signal about the current ground truth $\theta_t^\star$. Formally, the learning is canonically performed through the *quasi-maximum likelihood* principle, albeit equipped with a forgetting mechanism. Let b be a primitive of μ, $\lambda > 0$, $\{w_{s,t}\}$ the sequence of forgetting weights, and define:

$$\hat{\theta}_t \coloneqq \text{argmax}_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t} w_{s,t} \left[X_s A_s^\top \theta - b(A_s^\top \theta) \right] - \frac{\lambda c_\mu}{2} \|\theta\|_2^2 \,. \tag{1.30}$$

This formulation covers the sliding-window approach of [Cheung et al., 2019] with $w_{s,t} = \mathbb{1}(t - s \leq \tau)$ (τ being the length of the sliding window) and the exponential-weights with $\omega_{s,t} = \gamma^{t-s}$ and $\gamma \in (0, 1)$. The exploration is conducted according to the optimism-in-face-of-uncertainty principle: confidence regions for the ground-truth parameters are build around $\hat{\theta}_t$ and leveraged to ensure that the learner plays optimistic arms.

[Cheung et al., 2021, Zhao et al., 2020] extended their linear bandit analysis to generalized linear bandits with an inflated exploration bonus of order k_μ / c_μ to account for non-linearity. They claim regret bound of order $\mathcal{R}_T = \mathcal{O}(k_\mu c_\mu^{-1} d B_T^{1/3} T^{2/3})$ when given access to an upper-bound B_T on the true variation budget $\mathcal{B}_T$. Nevertheless both approaches disregard the fundamental non-linear aspect of generalized linear bandits. Following [Filippi et al., 2010], they rely on a linearization of the reward function around $\hat{\theta}_t$. Naturally, the linear approximation must accurately describe the *effective* behavior of the reward signal (characterized by the ground-truth $\theta_t^\star$). This translates in the structural constraint $\hat{\theta}_t \in \Theta$, which is implicitly assumed to hold in [Cheung et al., 2021, Zhao et al., 2020]. Unfortunately, there exists no proof guaranteeing that $\hat{\theta}_t \in \Theta$ could hold. Actually, existing deviation bounds [Abbasi-Yadkori et al., 2011, Theorem 1] rather suggest that in some directions, *even in the stationary case*, $\hat{\theta}_t$ can grow to be $\sqrt{d \log(t)}$ far from Θ. The situation is worse under non-stationarity since $\hat{\theta}_t$ can be B_t far from Θ.

Reward Model	Assumption	Regret Upper Bound
Linear (Chapter 4)	Orthogonal action sets	$\widetilde{\mathcal{O}}\left(B_T^{1/3}T^{2/3}\right)$
	✗	$\widetilde{\mathcal{O}}\left(B_T^{1/4}T^{3/4}\right)$
Generalized Linear (Chapter 6)	Orthogonal action sets	$\widetilde{\mathcal{O}}\left(B_T^{1/3}T^{2/3}\right)$
	✗	$\widetilde{\mathcal{O}}\left(B_T^{1/5}T^{4/5}\right)$

Table 1.1: Non-stationary contextuals bandits under drifting environments: regret upper-bounds presented in this thesis.

In Chapter 6, we propose the first correct analysis of generalized linear bandits under parameter drift. We detail the mistakes that were done by previous attempts. Under a geometric assumption on the action sets our algorithm BVD-GLM-UCB enjoys a regret of order $\widetilde{\mathcal{O}}(B_T^{1/3}T^{2/3})$. In the general case, we show that it suffers at most a $\widetilde{\mathcal{O}}(B_T^{1/5}T^{4/5})$ regret. At the core of our contribution is a generalization of the projection step introduced in [Filippi et al., 2010] and detailed in Section 1.5.2 adapted to the non-stationary nature of the problem.

In Table 1.1, we report the different regret guarantees that we obtain with the algorithm developed for the linear bandits and for the generalized linear bandits in drifting environments that are presented in this thesis.

2 | The ABC-S Learning Task

Motivated by A/B/n testing applications, we study in this chapter an alternative to the best arm identification task. We consider a finite set of distributions where one of them is treated as a *control*. We assume that the population is stratified into homogeneous subpopulations. At every time step, a subpopulation is sampled and an arm is chosen: the resulting observation is an independent draw from the arm conditioned on the subpopulation. The quality of each arm is assessed through a weighted combination of its subpopulation means. We propose a strategy for sequentially choosing one arm per time step so as to discover as fast as possible which arms, if any, have higher weighted expectation than the control. This strategy is shown to be asymptotically optimal in the following sense: if τ_δ is the first time when the strategy ensures that it is able to output the correct answer with probability at least $1-\delta$, then $\mathbb{E}[\tau_\delta]$ grows linearly with $\log(1/\delta)$ at the exact optimal rate. In this chapter, we identify this rate in three different settings: (1) when the experimenter does not observe the subpopulation information, (2) when the subpopulation of each sample is observed but not chosen, and (3) when the experimenter can select the subpopulation from which each response is sampled. We illustrate the efficiency of the proposed strategy with numerical simulations on synthetic data. The results from this chapter are based on [Russac et al., 2021b].

Outline

2.1 Introduction

A/B/n testing is a website optimization procedure where multiple versions of the content (called "arms" below) are compared, often in order to find the one with the highest conversion rate as represented on Figure 2.1. However, many e-commerce companies use A/B/n testing not only to deploy the best product implementation, but primarily to draw post-experiment inferences [Johari et al., 2015]. In addition to the experiment results several factors are taken into account before making a decision (e.g. the cost of scaling-up a solution). In this setting, each of the arms better than the default product (which we will refer to as the "control" arm) is a contender for being deployed and the interest is not only in the best arm.

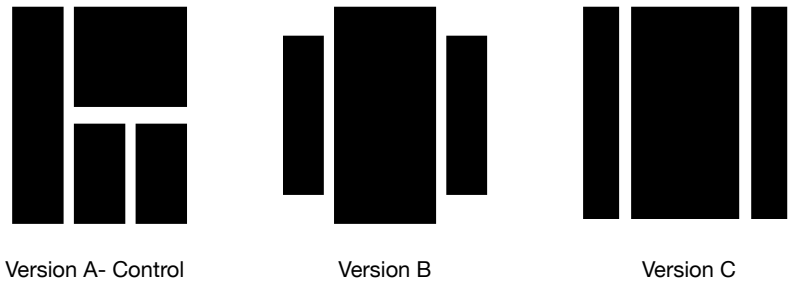


Figure 2.1: Three different versions of a webpage. The version A corresponds to the control arm that is currently used in production.

Given the control and $K \geq 1$ alternative implementations (variants), the simplest idea is to distribute the traffic uniformly among the arms; the arms that appear to be significantly better than the control at the end of the experiment are considered for deployment. While well-established, this process can be inefficient in terms of resources. Some alternatives are soon obviously worse (or better) than the control and would require fewer samples than the alternatives closer to the control. A second related shortcoming of the basic A/B/n testing approach is that setting the duration of the experiment –when done in advance– necessitates a very conservative approach by choosing a run-length that is sufficiently long to differentiate even the smallest possible changes.

To address these limitations, we consider in this chapter sequential testing policies that can both adjust the allocation of the samples and be stopped adaptively, in light of the data gathered during the experiment. This corresponds to a pure exploration task, see Section 1.2.3 for a brief introduction to this framework. A pure exploration strategy will typically choose every minute (say), an allocation of traffic that favors arms for which the uncertainty is the highest. The experiment is stopped as soon as the significance is considered sufficient for every arm. Approaches have been developed in [Even-Dar et al., 2006, Kalyanakrishnan et al., 2012, Garivier and Kaufmann, 2016] for the identification of the single arm with the highest mean. In particular, [Garivier and Kaufmann, 2016] propose a strategy that is asymptotically optimal in the *fixed confidence setting*, meaning that, given a risk parameter δ, it finds the best arm with probability at least $1 - \delta$, using an expected number of samples that is hardly improvable when δ is small. Later, [Yang et al., 2017] incorporated the special role of the control arm in BAI and proposed an algorithm that declares as winning arm the one with the highest mean only if it is significantly better than the control.

In this chapter, we propose a solution to the problem of identifying all the arms that are better than the control, in a framework that generalizes the fixed confidence setting.

In order to provide useful tools for practical A/B/n testing, we address two additional issues. First, traditional stochastic bandit models are based on the assumption that the arm samples are i.i.d., whereas real world data streams usually show trends or some form of inhomogeneity. A particular case of interest for website optimization are the seasonal patterns caused by time-of-day or day-of-week variations. We henceforth include in our model observed covariates (e.g. the time of the day, but possibly also the country of origin, or controlled covariates like the order in which partners appear on the page, etc.) that stratify the observations into homogeneous subpopulations. Using subpopulations can allow us to deal with non-stationary environment where the behavior of the users differs depending on the period of the day. We study different scenarios, depending on how much interaction is possible with these subpopulations. We provide a sample complexity analysis and an efficient algorithm in each case. In particular, we will show that using the subpopulation information efficiently can provide significant speedups of the decision-making. In the following, we will refer to the task of identifying the set of Arms that are Better than the Control in the presence of Subpopulations as the ABC-S problem.

Second, the practice of A/B/n testing often differs from a pure sequential experiment in that the experimenter cannot always fix a risk δ at the beginning and passively wait for the stopping time of the experiment without any time limitation. To address this issue, [Johari et al., 2015] proposed to define some notion of sequential "p-values" that can be monitored as the experiment progresses and used to terminate it. This notion was further used in the BAI setting in [Yang et al., 2017]. In this chapter, we elaborate on this idea by sequentially updating a suggested solution to the ABC-S problem *together with* a risk assessment for this suggestion. We show that, for any stopping time, the probability that the suggested solution is incorrect is indeed lower than the risk assessment. When the stopping time is selected as in usual fixed-confidence pure exploration, we recover the exact same guarantees but this view of the problem also provides useful results, for instance, if the experiment needs to be terminated prematurely.

Structure of the Chapter. The chapter is organized as follows. In Section 2.3, we present the mathematical model and study the information-theoretic complexity of the problem, extending the lower bound of [Garivier and Kaufmann, 2016] to the ABC-S setting. We show how the complexity of the problem depends on the degree of interaction that one has with the subpopulations, introducing different modes of interaction to be defined in Figure 2.4 below. We also consider in detail the Gaussian case which gives rise to more interpretable results. Section 2.4 describes how to implement the proposed strategy, which involves the numerical resolution of non-trivial optimization problems. Finally, we provide the results of numerical experiments on synthetic data sets in Section 2.5. All the missing elements for the results presented in this chapter are reported in Appendix.

2.2 Related Work

Pure exploration strategies have been studied in various settings: the identification of the best arm [Even-Dar et al., 2006, Garivier and Kaufmann, 2016], the identification of the top m arms [Chen et al., 2017, Kalyanakrishnan et al., 2012, Gabillon et al., 2012] identifying the arms that are better than a threshold [Locatelli et al., 2016, Cheshire et al., 2020], or identifying all ϵ-good arms [Mason et al., 2020]. As far as we know, the work presented here is the first

to consider the problem of identifying all the arms better than a control. It is also the first to consider subpopulations in pure exploration tasks. While motivated by the example of online companies, we believe that the proposed algorithms are relevant to other domains where randomized controlled trials are used for learning. An example could be clinical trials: one may wish to identify all the alternative treatments that work better that some reference medical treatment. This would permit to choose among them taking into account different characteristics (some could be cheaper, using another molecule for avoiding allergy, etc.).

Close to the notion of the control is the notion of threshold. [Locatelli et al., 2016] propose an algorithm for identifying all arms above a given threshold. Their algorithm samples according to the significance of a statistical test, and shares some similarities with the approach presented here in the Gaussian case; however, the perspective is rather different because the authors consider the *fixed-budget setting*. Here, the index of the control arm is known but its probability distribution is not.

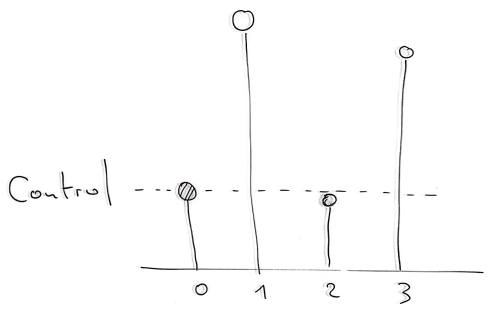


Figure 2.2: Mean of the control arm ($k = 0$) and of three other arms.

In Figure 2.2, we illustrate a setting with a control arm ($k = 0$) and three other arms. The four different probability distributions are unknown to the learner. We want to convince the reader that the problem of identifying the arms that are better than the control is different than (a) the BAI problem or (b) the problem of identifying arms that are better than a threshold. We focus on the fixed confidence setting. Quite intuitively, the larger the gaps between two means the easier it is to detect which one is the larger. For (a), the arm 1 is the one with the highest mean and is the target. Any reasonable algorithm will quickly discover that the arms 0 and 3 perform significantly worse than the arm 1. We expect an efficient policy to focus mostly on arm 1 and 3 for which the gap is smaller. For (b), the threshold problem assumes that the mean of the arm 0 is known. Again, the arms 1 and 3 will be quickly eliminated because they are far above the threshold. In this scenario, we expect the sampling strategy to sample mostly arm 2 because the mean of arm 0 is known. In our setting, the performance of the control arm is unknown and the arms better than the baseline are the arms 1 and 3. We expect the algorithm to pull the arms 0 and 2 most of the time in this case because those are the harder to distinguish and there is a need to estimate the control arm as well.

In this chapter, we also add an additional layer of complexity for measuring the performance of an arm. The quality of the different arms is assessed with a weighted combination of its subpopulations means. Minimizing the estimation error of a convex combination of means through adaptive sampling was considered in [Carpentier and Munos, 2011] with the introduction of a *stratified estimator* that will naturally appear in our analysis.

2.3 Complexity of the ABC-S Problem

2.3.1 Mathematical Framework

A problem instance consists of the following ingredients. Known to the learner are the number of arms $K \geq 1$ in addition to the designated control arm 0, the number of subpopulations J (a standard bandit being $J = 1$), and the vector $\boldsymbol{\beta} \in \mathbb{R}^J$ representing the relative importance of the subpopulations for the learning objective. We further make the stochastic assumption that samples from each arm k (including the control) and subpopulation i are drawn i.i.d. from an unknown probability distribution $\nu_{k,i}$ on $\mathbb{R}$, whose mean we will denote by $\mu_{k,i}$. The quality of arm k is $\mu_k \coloneqq \sum_{i=1}^J \beta_i \mu_{k,i}$, the combination of the means of the arms in the different populations. We now define formally the ABC-S problem.

Definition 2.1. For $\boldsymbol{\beta} \in \mathbb{R}^J$, we define the ABC-S problem as the correct identification of the set

$$\mathcal{S}_\beta(\boldsymbol{\mu}) \coloneqq \left\{ k \in \{1, \dots, K\} \;\middle|\; \sum_{i=1}^J \beta_i \mu_{k,i} > \sum_{i=1}^J \beta_i \mu_{0,i} \right\}.$$

At every time step t, the algorithm selects an arm A_t based on previous choices and outcomes and observes or selects (except when explicitly specified) the population type I_t. Upon the selection of the arm A_t a reward X_t is obtained. This defines a sigma-field generated by the observations up to time t denoted $\mathcal{F}_t = \sigma(A_1, I_1, X_1, \dots, A_t, I_t, X_t)$. The number of times arm k was selected for subpopulation i at time t is denoted $N_{k,i}(t) \coloneqq \sum_{s=1}^t \mathbb{1}(A_s = k, I_s = i)$ and the number of draws of arm k, $N_k(t) \coloneqq \sum_{s=1}^t \mathbb{1}(A_s = k)$. We define the gap with the control arm and arm k, $\Delta_k \coloneqq \mu_0 - \mu_k$.

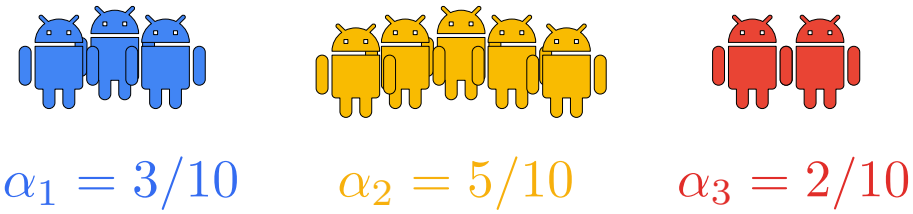


Figure 2.3: Example with three subpopulations and $\boldsymbol{\alpha} = (0.3, 0.5, 0.2) \in \Sigma_3$.

Modes of interaction We consider four modes of interaction of the learner with the bandit, as specified in Figure 2.4 below. In any of the three passive modes of interaction (described in Figures 2.4b to 2.4d), we assume that the subpopulation i represents a known proportion α_i of the total population, and hence that the sequence of subpopulations is drawn i.i.d. from the fixed and discrete distribution $I_t \sim \boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_J)$ with $\boldsymbol{\alpha} \in \Sigma_J \coloneqq \{x \in [0,1]^J \mid \sum_i x_i = 1\}$ the J-dimensional simplex. An example of a population with 3 different subpopulations is illustrated in Figure 2.3. $\boldsymbol{\alpha}$ is an exogenous parameter and can differ from $\boldsymbol{\beta}$ which is inherent to the learning objective and is also assumed to be known. Although it is most natural in many applications to consider that $\boldsymbol{\beta} = \boldsymbol{\alpha}$ (it is even necessary in the oblivious mode to make the estimation of the μ_k's feasible), Example 2 below describes a concrete scenario in which $\boldsymbol{\beta}$ has negative components.

Example 1. Let us consider a company that sells cap to the subpopulations from Figure 2.3. It might be the case that users from different subpopulations behave differently when offered the

	1. See $I_t \sim \boldsymbol{\alpha}$	1. Pick A_t	1. Pick A_t
1. Pick A_t and I_t	2. Pick A_t	2. See $I_t \sim \boldsymbol{\alpha}$	2. Do *not* see $I_t \sim \boldsymbol{\alpha}$
2. See $X_t \sim \nu_{A_t,I_t}$	3. See $X_t \sim \nu_{A_t,I_t}$	3. See $X_t \sim \nu_{A_t,I_t}$	3. See $X_t \sim \nu_{A_t,I_t}$
(a) *Active* mode	(b) *Proportional* mode	(c) *Agnostic* mode	(d) *Oblivious* mode.

Figure 2.4: Modes of Interaction between Learner and Bandit in each round. In Active mode the learner determines the subpopulation, while in the right three passive modes it is sampled from $\boldsymbol{\alpha}$.

same product as illustrated in Figure 2.5. For this reason, it seems natural from the company perspective to consider the performance of a product as the weighted sum over the different subpopulations. While natural proportion $\boldsymbol{\alpha}$ might exist, it could be the case that for some reason the company is really interested in targeting a specific subpopulation. This is when allowing $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$ in the learning objective becomes attractive.

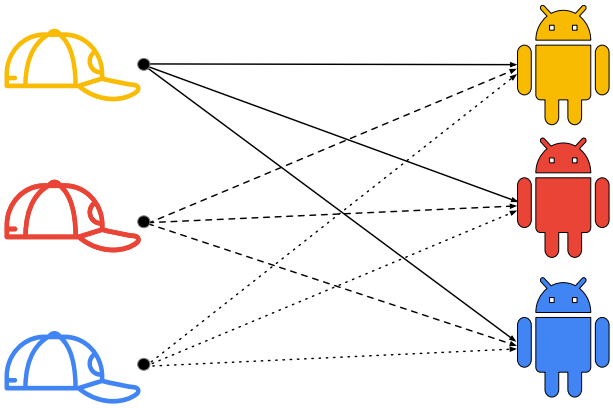


Figure 2.5: A company selling caps to a population with $J = 3$ subpopulations.

Example 2 ([Kaufmann and Koolen, 2018, Largest Profit Identification problem])**.** Consider a company choosing among K product designs the model to mass produce. Each candidate design k has an (equilibrium) sales price $\mu_{k,1}$ and production cost $\mu_{k,2}$. The goal is to find the model k with the largest profit $\mu_{k,1} - \mu_{k,2}$. Prices and costs are currently unknown, but can be adaptively sampled. Sampling the "price" subpopulation $i = 1$ is typically implemented by performing user preference studies, taking questionnaires, etc. Samples from the "cost" subpopulation $i = 2$ involve rating manufacturing facilities, forecasting material and labor costs etc. The importance vector is here $\boldsymbol{\beta} = (1, -1)$ and $\boldsymbol{\alpha}$ has to be set by the learner.

The distributions $(\nu_{a,i})_{a,i}$ are assumed to belong the same one-parameter exponential family, $\mathcal{P} \coloneqq \{(\nu_\theta)_\theta : d\nu_\theta/d\xi = \exp(\theta x - b(\theta))\}$, with ξ a reference measure on $\mathbb{R}$ and $b : \Theta \subset \mathbb{R} \mapsto \mathbb{R}$. Every probability distribution ν_θ in $\mathcal{P}$ is entirely defined by its mean $\dot{b}(\theta)$ [Cappé et al., 2013]. We may hence identify any bandit instance with its matrix of means $\boldsymbol{\mu} \in \mathbb{R}^{(K+1)\times J}$. In addition, the Kullback-Leibler divergence between two distributions ν_θ and $\nu_{\theta'} \in \mathcal{P}$ may be written in the following Bregman form:

$$d(\mu, \mu') = \mathrm{KL}(\nu_\theta, \nu_{\theta'}) = b(\theta') - b(\theta) - \dot{b}(\theta)(\theta' - \theta) \ ,$$

where $\mu = \dot{b}(\theta)$ and $\mu' = \dot{b}(\theta')$ correspond to the means of the two distributions ν_θ and $\nu_{\theta'}$. We also use the notation $\mathrm{kl}(p, q)$ to denote the KL divergence of two Bernoulli distributions of parameter p and q.

We define $\mathcal{L} \coloneqq \{\boldsymbol{\mu} : \forall k \in \{1, \dots, K\} \cup \{0\}, \forall i \in \{1, \dots, J\}, \nu_{k,i} \in \mathcal{P} \text{ and } \mu_0 \neq \mu_k\}$ the set of identifiable instances where no arm has the same weighted mean as the control. Our objective is to solve the ABC-S problem for any bandit instance $\boldsymbol{\mu} \in \mathcal{L}$. At every time step, the policies we consider output a risk assessment $\hat{\delta}_t$ together with a recommendation $\hat{\mathcal{S}}_t$. We focus on *safely calibrated* policies, that are defined below.

Definition 2.2 (Safely Calibrated Policies)**.** A policy is said to be safely calibrated when satisfying

$$\forall \boldsymbol{\mu} \in \mathcal{L},\ \forall \delta \in (0,1), \quad \mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \geq 1 : \hat{\mathcal{S}}_t \neq \mathcal{S}_\beta(\boldsymbol{\mu}) \ \cap\ \hat{\delta}_t \leq \delta\right) \leq \delta\ . \tag{2.1}$$

Finally, when fixing a level of risk δ, we consider the stopping time associated to the filtration $\mathcal{F}_t$, $\tau_\delta = \inf\{t \geq 0,\ \hat{\delta}_t \leq \delta\}$. The objective is then to minimize the expected number of rounds necessary to obtain a level of risk of at most δ. Contrary to usual δ-correct algorithms if stopped before τ_δ, the strategy still provides guarantees on the output set following Equation 2.1. In particular, safely calibrated policies have a sampling rule that does not depend on any pre-specified δ, and as such they are δ-correct for any δ.

2.3.2 General Form of the Sample Complexity

Depending on the mode of interaction from Figure 2.4, the learner has a set of sampling constraints to satisfy, here denoted $\mathcal{C}$ and precisely defined in the next section. We define $\mathrm{Alt}_\beta(\boldsymbol{\mu})$, the different problem instances where the set of arms better than the control differs from that of the instance $\boldsymbol{\mu}$. Formally, $\mathrm{Alt}_\beta(\boldsymbol{\mu}) \coloneqq \{\boldsymbol{\lambda} \in \mathcal{L} \mid \mathcal{S}_\beta(\boldsymbol{\lambda}) \neq \mathcal{S}_\beta(\boldsymbol{\mu})\}$. This allows us to bound the sample complexity.

Theorem 2.3. *Let $\delta \in (0,1)$ and $\boldsymbol{\beta} \in \mathbb{R}^J$. For any strategy satisfying Equation 2.1 and any $\boldsymbol{\mu} \in \mathcal{L}$, the expected number of rounds for the ABC-S problem for the agnostic, proportional and active mode satisfies:*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T^\star(\boldsymbol{\mu})\,\mathrm{kl}(\delta, 1-\delta) \quad \text{and} \quad \liminf_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\ln(1/\delta)} \geq T^\star(\boldsymbol{\mu})\ . \tag{2.2}$$

where recalling that $\lambda_k = \sum_{i=1}^J \beta_i \lambda_{k,i}$,

$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{w} \in \mathcal{C}} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}_\beta(\boldsymbol{\mu})} \sum_{k=0}^{K} \sum_{i=1}^{J} w_{k,i} d(\mu_{k,i}, \lambda_{k,i}) \tag{2.3}$$

$$= \sup_{\boldsymbol{w} \in \mathcal{C}} \min_{b \neq 0} \inf_{\boldsymbol{\lambda} \in \mathcal{L} : \lambda_0 = \lambda_b} \sum_{k \in \{0,b\}} \sum_{i=1}^{J} w_{k,i} d(\mu_{k,i}, \lambda_{k,i})\ . \tag{2.4}$$

Proof. Using the transportation lemma from [Kaufmann et al., 2016] and recalling that $N_{k,i}(t)$ is the number of draws of arm k in subpopulation i up to time t, we have for any

safely calibrated policies

$$\forall \boldsymbol{\lambda} \in \mathrm{Alt}_\beta(\boldsymbol{\mu}),\ \sum_{i=1}^{J}\sum_{k=0}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_{k,i}(\tau_\delta)] d(\mu_{k,i}, \lambda_{k,i}) \geq \mathrm{kl}(\delta, 1-\delta)\ .$$

Therefore,

$$\begin{aligned}
\mathrm{kl}(\delta, 1-\delta) &\leq \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}_\beta(\boldsymbol{\mu})} \sum_{k=0}^{K}\sum_{i=1}^{J} \mathbb{E}_{\boldsymbol{\mu}}[N_{k,i}(\tau_\delta)] d(\mu_{k,i}, \lambda_{k,i}) \\
&= \mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}_\beta(\boldsymbol{\mu})} \sum_{k=0}^{K}\sum_{i=1}^{J} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_{k,i}(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]} d(\mu_{k,i}, \lambda_{k,i}) \\
&\leq \mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \sup_{\boldsymbol{w}\in\mathcal{C}} \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}_\beta(\boldsymbol{\mu})} \sum_{k=0}^{K}\sum_{i=1}^{J} w_{k,i} d(\mu_{k,i}, \lambda_{k,i})\ .
\end{aligned}$$

In the last inequality, we used the fact that the normalized expected numbers of draws satisfy the set of constraints defined by $\mathcal{C} \subset \Sigma_{(K+1)J}$. Using $\mathrm{kl}(\delta, 1-\delta) \sim \ln(1/\delta)$ when δ tends to 0 gives the first result.

We denote $\Lambda_J(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \coloneqq \sum_{k=0}^{K}\sum_{i=1}^{J} w_{k,i} d(\mu_{k,i}, \lambda_{k,i})$. To obtain the second result, we will simplify the expression of $T^\star(\boldsymbol{\mu})^{-1}$. Using that the KL divergences and the weights are positive, for $\boldsymbol{\lambda}$ to be in the alternative, one of the two following conditions need to be met: (1) there exists $k \in \mathcal{S}_\beta(\boldsymbol{\mu})$ such that $\lambda_k < \lambda_0$ or (2) there exists $k \in \mathcal{S}_\beta^-(\boldsymbol{\mu}) \coloneqq \{k \in \{1, \dots, K\} \mid \mu_k < \mu_0\}$ such that $\lambda_k > \lambda_0$.
For this reason, one has

$$\inf_{\boldsymbol{\lambda}\in\mathrm{Alt}_\beta(\boldsymbol{\mu})} \Lambda_J(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \min\left(\min_{k\in\mathcal{S}_\beta(\boldsymbol{\mu})} \inf_{\boldsymbol{\lambda}:\lambda_k<\lambda_0} \Lambda_J(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\mu}),\ \min_{k\in\mathcal{S}_\beta^-(\boldsymbol{\mu})} \inf_{\boldsymbol{\lambda}:\lambda_k>\lambda_0} \Lambda_J(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \right)\ .$$

We obtain the desired result by remarking that the inner optimization programs $\inf_{\boldsymbol{\lambda}}$ are each achieved on the boundary (the constraint being satisfied with equality) where they coincide, and that $\{1, \dots, K\} = \mathcal{S}_\beta(\boldsymbol{\mu}) \cup \mathcal{S}_\beta^-(\boldsymbol{\mu})$. □

2.3.3 Influence of the Mode of Interaction

We consider the four different modes governing the sampling rule as outlined in Figure 2.4.

2.3.3.1 Agnostic Mode

In the *agnostic* mode (Fig. 2.4c) an arm is first selected, after which the subpopulation type is observed. Mathematically, this brings the equality $\mathbb{E}_{\boldsymbol{\mu}}[N_{k,i}(T)] = \alpha_i \mathbb{E}_{\boldsymbol{\mu}}[N_k(T)]$ established in Lemma 2.4 and the independence constraint on the weights $\boldsymbol{w} \in \mathcal{C}_{\text{agnostic}} \coloneqq \{w_{k,i} = \alpha_i u_k : (u_0, \dots, u_K) \in \Sigma_{K+1}\}$.

Lemma 2.4. *For any agnostic policy where A_t is chosen knowing $\mathcal{F}_{t-1}$ but independently from I_t, when defining $N_{k,j}(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = k \cap I_s = j)$ and $N_k(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = k)$, then*

$$\forall k \in \{0, \dots, K\}, \forall j \in \{1, \dots, J\}, \forall t \geq 1, \quad \mathbb{E}_{\boldsymbol{\mu}}[N_{k,j}(t)] = \alpha_j \mathbb{E}_{\boldsymbol{\mu}}[N_k(t)]$$

Proof.

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mu}}[N_{k,j}(t)] &= \sum_{s=1}^{t} \mathbb{P}(A_s = k \cap I_s = j) = \sum_{s=1}^{t} \mathbb{P}_{\boldsymbol{\mu}}(A_s = k | I_s = j)\mathbb{P}(I_s = j) \\ &= \sum_{s=1}^{t} \alpha_j \mathbb{P}_{\boldsymbol{\mu}}(A_s = k | I_s = j) = \sum_{s=1}^{t} \alpha_j \mathbb{P}_{\boldsymbol{\mu}}(A_s = k) \\ &= \alpha_j \mathbb{E}_{\boldsymbol{\mu}}[N_k(t)] \ ,\end{aligned}$$

where in the fourth equality, we have used that the action A_t is selected independently from the population indicator I_t. □

2.3.3.2 Proportional Mode.

In the *proportional* mode (Fig. 2.4b), A_t is chosen based on $\mathcal{F}_{t-1}$ and the current subpopulation I_t. Here, the constraint is that the total number of pulls of the different arms in the subpopulation i should respect the frequency of this subpopulation, i.e. $\sum_k \mathbb{E}_{\boldsymbol{\mu}}[N_{k,i}(T)] = \alpha_i T$. This induces a marginal constraint on the weights of the form $\boldsymbol{w} \in \mathcal{C}_{\text{prop}} \coloneqq \{\boldsymbol{w} \in \Sigma_{(K+1)J} \mid \forall i \leq J, \ \sum_k w_{k,i} = \alpha_i\}$. This result is established in Lemma 2.5 reported below.

Lemma 2.5. *For any proportional policy where A_t is chosen knowing $\mathcal{F}_{t-1}$ and I_t, when defining $N_{k,j}(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = k \cap I_s = j)$ and $N_k(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = k)$, then*

$$\forall j \in \{1, \dots, J\}, \forall t \geq 1, \quad \sum_{k=0}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_{k,j}(t)] = \alpha_j t \ .$$

Proof.

$$\begin{aligned}\sum_{k=0}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_{k,j}(t)] &= \sum_{s=1}^{t} \sum_{k=0}^{K} \mathbb{E}_{\boldsymbol{\mu}}\left[\mathbb{1}(I_s = j)\mathbb{1}(A_s = k)\right] = \sum_{s=1}^{t} \mathbb{E}_{\boldsymbol{\mu}}\left[\mathbb{1}(I_s = j) \sum_{k=0}^{K} \mathbb{1}(A_s = k)\right] \\ &= \sum_{s=1}^{t} \mathbb{P}_{\boldsymbol{\mu}}(I_s = j) = \alpha_j t \ .\end{aligned}$$

□

2.3.3.3 Active Mode

In the *active* mode (Fig. 2.4a), the learner has an additional degree of freedom and can ask for any subpopulation type at any round. In that case, $\boldsymbol{w} \in \mathcal{C}_{\text{active}} \coloneqq \Sigma_{(K+1)J}$ is unconstrained.

By remarking that $\mathcal{C}_{\text{agnostic}} \subset \mathcal{C}_{\text{prop}} \subset \mathcal{C}_{\text{active}}$, and given the optimization program (Equation (2.3)) solved to obtain the characteristic time, one immediately gets

$$\forall \boldsymbol{\mu} \in \mathcal{L}, \quad T^{\star}_{\text{active}}(\boldsymbol{\mu}) \leq T^{\star}_{\text{proportional}}(\boldsymbol{\mu}) \leq T^{\star}_{\text{agnostic}}(\boldsymbol{\mu}) \ . \tag{2.5}$$

Hence, as expected, the more control/information on the subpopulation the learner has, the faster he is able to identify the set of arms that are better than the control.

2.3.3.4 Oblivious Mode

To compare with the *oblivious* mode, in which the subpopulation information is not even observed, we have to assume that $\boldsymbol{\alpha} = \boldsymbol{\beta}$. In that case, the arm rewards follow a mixture distribution: $X_t|A_t = k \sim \sum_{i=1}^{J} \alpha_i \nu_{k,i}$. We properly define the characteristic time of an oblivious safely calibrated policy in the following proposition.

Proposition 2.6. *Let $\delta \in (0,1)$ and $\boldsymbol{\beta} \in \mathbb{R}^J$. For any oblivious strategy satisfying Equation 2.1 and any $\boldsymbol{\mu} \in \mathcal{L}$, the expected number of rounds for the ABC-S problem for the oblivious mode satisfies:*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T^\star_{\text{oblivious}}(\boldsymbol{\mu})\,\mathrm{kl}(\delta, 1-\delta) \quad \text{and} \quad \liminf_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\ln(1/\delta)} \geq T^\star_{\text{oblivious}}(\boldsymbol{\mu})\ .$$

where

$$T^\star_{\text{oblivious}}(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{w} \in \Sigma_{K+1}} \inf_{\nu' \in \mathrm{Alt}(\nu)} \sum_{k=0}^{K} w_k \mathrm{KL}\left(\sum_{i=1}^{J} \alpha_i \nu_{k,i}, \sum_{i=1}^{J} \alpha_i \nu'_{k,i}\right)\ . \tag{2.6}$$

Furthermore,

$$\forall \boldsymbol{\mu} \in \mathcal{L}, \quad T^\star_{\text{oblivious}}(\boldsymbol{\mu}) \geq T^\star_{\text{agnostic}}(\boldsymbol{\mu})\ .$$

Proof. While with observable subpopulations the distributions are entirely characterized by their means, this is no longer the case with mixture distributions. In particular, this requires defining a different alternative.

$$\mathrm{Alt}(\nu) \coloneqq \left\{ \nu' \;\middle|\; \forall k, \nu'_k = \sum_{i=1}^{J} \alpha_i \nu'_{k,i} \text{ with } \nu'_{k,i} \in \mathcal{P} \text{ and } \mathcal{S}_{\boldsymbol{\beta}}(\nu') \neq \mathcal{S}_{\boldsymbol{\beta}}(\nu) \right\}\ .$$

Using the transportation lemma from [Kaufmann et al., 2016], we have for any safely calibrated oblivious policy

$$\forall \nu' \in \mathrm{Alt}(\nu),\ \sum_{k=0}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_k(\tau_\delta)] \mathrm{KL}\left(\sum_{i=1}^{J} \alpha_i \nu_{k,i}, \sum_{i=1}^{J} \alpha_i \nu'_{k,i}\right) \geq \mathrm{kl}(\delta, 1-\delta)\ .$$

Therefore,

$$\begin{aligned}
\mathrm{kl}(\delta, 1-\delta) &\leq \inf_{\nu' \in \mathrm{Alt}(\nu)} \sum_{k=0}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_k(\tau_\delta)] \mathrm{KL}\left(\sum_{i=1}^{J} \alpha_i \nu_{k,i}, \sum_{i=1}^{J} \alpha_i \nu'_{k,i}\right) \\
&= \mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \inf_{\nu' \in \mathrm{Alt}(\nu)} \sum_{k=0}^{K} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_k(\tau_\delta)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]} \mathrm{KL}\left(\sum_{i=1}^{J} \alpha_i \nu_{k,i}, \sum_{i=1}^{J} \alpha_i \nu'_{k,i}\right) \\
&\leq \mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \sup_{\boldsymbol{w} \in \Sigma_{K+1}} \inf_{\nu' \in \mathrm{Alt}(\nu)} \sum_{k=0}^{K} w_k \mathrm{KL}\left(\sum_{i=1}^{J} \alpha_i \nu_{k,i}, \sum_{i=1}^{J} \alpha_i \nu'_{k,i}\right)\ .
\end{aligned}$$

Using $\mathrm{kl}(\delta, 1-\delta) \sim \ln(1/\delta)$ when δ tends to 0 gives the first result. Using the joint convexity of the KL divergence one gets

$$\mathrm{KL}\left(\sum_{i=1}^{J} \alpha_i \nu_{k,i}, \sum_{i=1}^{J} \alpha_i \nu'_{k,i}\right) \leq \sum_{i=1}^{J} \alpha_i \mathrm{KL}(\nu_{k,i}, \nu'_{k,i})\ .$$

Assuming that the mean of $\nu'_{k,i}$ equals $\lambda_{k,i}$ and recalling that for distributions in $\mathcal{P}$, one has $\mathrm{KL}(\nu_{k,i}, \nu'_{k,i}) = d(\mu_{k,i}, \lambda_{k,i})$, we deduce,

$$\begin{aligned} T^\star_{\text{oblivious}}(\boldsymbol{\mu})^{-1} &= \sup_{w \in \Sigma_{K+1}} \inf_{\nu' \in \text{Alt}(\nu)} \sum_{k=0}^{K} w_k \mathrm{KL}\left(\sum_{i=1}^{J} \alpha_i \nu_{k,i}, \sum_{i=1}^{J} \alpha_i \nu'_{k,i}\right) \\ &\leq \sup_{w \in \Sigma_{K+1}} \inf_{\boldsymbol{\lambda} \in \text{Alt}(\boldsymbol{\mu})} \sum_{k=0}^{K} \sum_{i=1}^{J} \alpha_i w_k d(\mu_{k,i}, \lambda_{k,i}) \\ &= T^\star_{\text{agnostic}}(\boldsymbol{\mu})^{-1} \, . \end{aligned}$$

□

This completes the picture of the ordering of the characteristic times by showing that, when $\boldsymbol{\alpha} = \boldsymbol{\beta}$,

$$\forall \boldsymbol{\mu} \in \mathcal{L}, \quad T^\star_{\text{active}}(\boldsymbol{\mu}) \leq T^\star_{\text{proportional}}(\boldsymbol{\mu}) \leq T^\star_{\text{agnostic}}(\boldsymbol{\mu}) \leq T^\star_{\text{oblivious}}(\boldsymbol{\mu}) \, . \tag{2.7}$$

Note that although we provide, in Section 2.4, algorithms to numerically compute the first three complexites, evaluating $T^\star_{\text{oblivious}}(\boldsymbol{\mu})$ would be much harder, as the mixture distributions can no more be parameterized by their mean only. Our current techniques do not yield a general-purpose practical algorithm that is asymptotically optimal in the *oblivious* mode for the ABC-S problem. In the Bernoulli case, however, as mixtures of Bernoulli distributions are Bernoulli distribution, one can use the single-population Bernoulli approach discussed in the next paragraph. For Gaussian distributions, one can use a suboptimal approach based on the observation that location mixtures of Gaussians with bounded means are sub-Gaussian (see Appendix 2.E.1 for details).

2.3.4 Single Population and Relationship with Best Arm Identification

In order to illustrate the nature of the ABC-S problem, we make a detour through the single population case, that is, when $J = 1$. Given two weights w_a, w_b and two means μ_a, μ_b, we introduce the minimum weighted transportation cost for moving the means to a common position.

$$d_{\text{mid}}(w_a, \mu_a, w_b, \mu_b) \;\coloneqq\; \inf_v w_a d(\mu_a, v) + w_b d(\mu_b, v) = w_a d(\mu_a, v^\star_{a,b}) + w_b d(\mu_b, v^\star_{a,b}) \tag{2.8}$$

where $v^*_{a,b}$, the optimal common location, is the weighted average, i.e.

$$v^*_{a,b} = \frac{w_a}{w_a + w_b}\mu_a + \frac{w_b}{w_a + w_b}\mu_b \, .$$

Constructing an Instance in the Alternative When identifying all the arms better than a control, there are two different ways to obtain a close-by bandit model $\boldsymbol{\lambda}$ in the alternative. The first option consists in taking an arm which does not belong to $\mathcal{S}_\beta(\boldsymbol{\mu})$ and to augment its mean on the alternative model such that it becomes above the control (or to reduce the mean of the control). Otherwise, it is possible to take an arm that is better than the control in the bandit model $\boldsymbol{\mu}$ and to shrink its mean such that it becomes lower than the control on the alternative (or augment the control). Note that the infimum over the alternative has the same expression in the two cases (see proof of Proposition 2.7).

There is a priori no link between a BAI problem and an ABC one. In particular, in the BAI problem there are only $K+1$ possible choices for the best arm while when looking for $\mathcal{S}_\beta(\boldsymbol{\mu})$ there are up to 2^K different sets to consider. Yet, the next proposition shows that the characteristic time $T^\star$ of any ABC problem with $J=1$ subpopulation shares strong similarities with that of BAI problems.

Proposition 2.7 (Characteristic time with a single population)**.** *Let $\delta \in (0,1)$, $\boldsymbol{\mu} \in \mathcal{L}$ and assume that $J=1$. For any strategy satisfying Equation 2.1, Equation 2.2 holds with*

$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{w}\in\Sigma_{K+1}} \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}_\beta(\boldsymbol{\mu})} \sum_{k=0}^{K} w_k d(\mu_k,\lambda_k) = \sup_{\boldsymbol{w}\in\Sigma_{K+1}} \min_{b\neq 0} d_{\mathrm{mid}}(w_0,\mu_0,w_b,\mu_b)\ .$$

Proof. In the particular case when $J=1$, the expression of the characteristic time can be simplified. The first part of the proof can be obtained using similar argument than for Theorem 2.3. The missing part is the simplification of the expression of $T^\star(\boldsymbol{\mu})$.

We denote $\Lambda_1(\boldsymbol{w},\boldsymbol{\lambda},\boldsymbol{\mu}) \coloneqq \sum_{k=0}^K w_k d(\mu_k,\lambda_k)$. Following the reasoning from the proof of Theorem 2.3 one of the two following conditions needs to be met: (1) there exists $k\in\mathcal{S}_\beta(\boldsymbol{\mu})$ such that $\lambda_k<\lambda_0$. (2) there exists $k\in\mathcal{S}_\beta^-(\boldsymbol{\mu}) \coloneqq \{k\in\{1,\dots,K\} \mid \mu_k<\mu_0\}$ such that $\lambda_k>\lambda_0$.
For this reason, one has

$$\inf_{\boldsymbol{\lambda}\in\mathrm{Alt}_\beta(\boldsymbol{\mu})} \Lambda_1(\boldsymbol{w},\boldsymbol{\lambda},\boldsymbol{\mu}) = \min\left(\min_{k\in\mathcal{S}_\beta(\boldsymbol{\mu})} \inf_{\boldsymbol{\lambda}:\lambda_k<\lambda_0} \Lambda_1(\boldsymbol{w},\boldsymbol{\lambda},\boldsymbol{\mu}), \min_{k\in\mathcal{S}_\beta^-(\boldsymbol{\mu})} \inf_{\boldsymbol{\lambda}:\lambda_k>\lambda_0} \Lambda_1(\boldsymbol{w},\boldsymbol{\lambda},\boldsymbol{\mu}) \right)\ .$$

In this simpler case, it is possible to obtain an explicit formula for this infimum. We start from

$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{w}\in\Sigma_{K+1}} \min\left(\min_{k\in\mathcal{S}_\beta(\boldsymbol{\mu})} \inf_{\boldsymbol{\lambda}:\lambda_k<\lambda_0} \Lambda_1(\boldsymbol{w},\boldsymbol{\lambda},\boldsymbol{\mu}), \min_{k\in\mathcal{S}_\beta^-(\boldsymbol{\mu})} \inf_{\boldsymbol{\lambda}:\lambda_k>\lambda_0} \Lambda_1(\boldsymbol{w},\boldsymbol{\lambda},\boldsymbol{\mu}) \right)\ .$$

Let us focus on the case, $\lambda_k<\lambda_0$ and fix an index $k\in\mathcal{S}_\beta(\boldsymbol{\mu})$. Λ_1 is always smaller when all the λ_b for $b\neq 0$ and $b\neq k$ coincides with μ_b. This gives,

$$\min_{k\in\mathcal{S}_\beta(\boldsymbol{\mu})} \inf_{\boldsymbol{\lambda}:\lambda_k<\lambda_0} \Lambda_1(\boldsymbol{w},\boldsymbol{\lambda},\boldsymbol{\mu}) = \min_{k\in\mathcal{S}_\beta(\boldsymbol{\mu})} \inf_{\boldsymbol{\lambda}:\lambda_k\leq\lambda_0} w_0 d(\mu_0,\lambda_0) + w_k d(\mu_k,\lambda_k)\ .$$

We consider the Lagrangian function, $L(\lambda_0,\lambda_k,q) = w_0 d(\mu_0,\lambda_0) + w_k d(\mu_k,\lambda_k) + q(\lambda_k-\lambda_0)$. Differentiating with respect to λ_0 and λ_k brings the condition

$$\lambda_0^\star = \lambda_k^\star = \lambda_{k,0}^\star = \mathrm{argmin}_\lambda w_0 d(\mu_0,\lambda) + w_k d(\mu_k,\lambda) = \frac{w_0}{w_0+w_k}\mu_0 + \frac{w_k}{w_0+w_k}\mu_k\ .$$

Recalling, $d_{\mathrm{mid}}(w_0,\mu_0,w_k,\mu_k) \coloneqq \inf_v w_0 d(\mu_0,v) + w_k d(\mu_k,v)$ one has,

$$\min_{k\in\mathcal{S}_\beta(\boldsymbol{\mu})} \inf_{\boldsymbol{\lambda}:\lambda_k<\lambda_0} \Lambda_1(\boldsymbol{w},\boldsymbol{\lambda},\boldsymbol{\mu}) = \min_{k\in\mathcal{S}_\beta(\boldsymbol{\mu})} d_{\mathrm{mid}}(w_0,\mu_0,w_k,\mu_k)\ . \tag{2.9}$$

Solving the optimization program for $k\in\mathcal{S}_\beta^-(\boldsymbol{\mu})$ and under the constraint $\lambda_k>\lambda_0$, gives the exact same set of constraints and optimal solution, i.e.

$$\min_{k\in\mathcal{S}_\beta^-(\boldsymbol{\mu})} \inf_{\boldsymbol{\lambda}:\lambda_k>\lambda_0} \Lambda_1(\boldsymbol{w},\boldsymbol{\lambda},\boldsymbol{\mu}) = \min_{k\in\mathcal{S}_\beta^-(\boldsymbol{\mu})} d_{\mathrm{mid}}(w_0,\mu_0,w_k,\mu_k)\ . \tag{2.10}$$

Bringing Equation 2.9 and Equation 2.10 together and remarking that $\{1,\dots,K\} = \mathcal{S}_\beta(\boldsymbol{\mu}) \cup \mathcal{S}_\beta^-(\boldsymbol{\mu})$ gives the announced result. □

Note that the expression of the sample complexity is really close to the one in the BAI setting ([Garivier and Kaufmann, 2016, Lemma 3] or Theorem 1.8 from Section 1.2.3) except that we consider all the indices different from the control here instead of the indices different from the best arm.

2.3.5 The Gaussian Case

In this section, we consider the Gaussian case which is of interest as the characteristic time admits a more explicit expression, making it possible to further investigate the differences between the various modes of interaction. We will state our results for the heteroscedastic case, in particular to get a closed-form proxy for the Bernoulli case, where each variance is a function of the (unknown) mean.

A/B Testing. When $K = 1$ (one arm and the control arm), we are considering a standard A/B test with subpopulations and one can prove the following result (established in Appendix 2.A).

Proposition 2.8. *For any $\boldsymbol{\mu} \in \mathcal{L}$ with $K = 1$ and $\nu_{k,i} = \mathcal{N}(\mu_{k,j}, \sigma^2_{k,j})$, recalling that $\Delta_1 = \sum_{i=1}^J \beta_i(\mu_{0,i} - \mu_{1,i})$ one has*

1. *In the agnostic case, the characteristic time and the optimal weights satisfy*

$$T^\star_{\text{agnostic}}(\boldsymbol{\mu}) = \frac{2\left(\sqrt{\sum_{i=1}^J \frac{\beta_i^2\sigma_{0,i}^2}{\alpha_i}} + \sqrt{\sum_{i=1}^J \frac{\beta_i^2\sigma_{1,i}^2}{\alpha_i}}\right)^2}{\Delta_1^2}$$

$$\text{and } \forall i \le J, \forall k \in \{0,1\},\ w^\star_{k,i} = \frac{\alpha_i\sqrt{\sum_{i=1}^J \frac{\beta_i^2\sigma_{k,i}^2}{\alpha_i}}}{\sqrt{\sum_{i=1}^J \frac{\beta_i^2\sigma_{0,i}^2}{\alpha_i}} + \sqrt{\sum_{i=1}^J \frac{\beta_i^2\sigma_{1,i}^2}{\alpha_i}}}\,.$$

2. *In the proportional case, the characteristic time and the optimal weights satisfy*

$$T^\star_{\text{prop}}(\boldsymbol{\mu}) = \frac{2\sum_{i=1}^J \frac{\beta_i^2}{\alpha_i}(\sigma_{0,i} + \sigma_{1,i})^2}{\Delta_1^2}$$

$$\text{and } \forall i \le J, \forall k \in \{0,1\},\ w^\star_{k,i} = \frac{\alpha_i\sigma_{k,i}}{\sigma_{0,i} + \sigma_{1,i}}\,.$$

3. *In the active case, the characteristic time and the optimal weights satisfy*

$$T^\star_{\text{active}}(\boldsymbol{\mu}) = \frac{2\left(\sum_{i=1}^J |\beta_i|(\sigma_{0,i} + \sigma_{1,i})\right)^2}{\Delta_1^2}$$

$$\text{and } \forall i \le J, \forall k \in \{0,1\},\ w^\star_{k,i} = \frac{|\beta_i|\sigma_{k,i}}{\sum_{i=1}^J |\beta_i|(\sigma_{0,i} + \sigma_{1,i})}\,.$$

The optimal allocations in the *agnostic* and *proportional* cases are constrained by the proportion of the different subpopulations $\boldsymbol{\alpha}$, whereas, for the *active* mode, the optimal weights only depend on $\boldsymbol{\beta}$. In general, the optimal weights also depend on the subpopulation variances, as is well-known in stratified sampling estimation. Note however, that when (a) the subpopulations

all have a common variance σ^2 and (b) $\boldsymbol{\beta} = \boldsymbol{\alpha}$, then the optimal allocations and the characteristic times are equal for the *agnostic*, the *proportional* and the *active* modes. In that case, $w^\star_{k,i} = \alpha_i/2$, which also corresponds to the well-known result in Gaussian A/B testing [Kaufmann et al., 2016]. We have more generally observed that whenever the subpopulations have approximately the same variances, the *agnostic* and *proportional* modes yield very similar performances.

Weight Computation in the Homoscedastic Case Even in scenarios where all subpopulation variances are equal to σ^2, the *active* mode remains very attractive in the cases where $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$. The following proposition shows that in that case, the optimal weights for the ABC-S problem can be computed efficiently.

Proposition 2.9 (Efficient computation of the optimal weights in the Gaussian case)**.** *Assume Gaussian distributions with a known variance σ^2, and let*

$$(u_0^\star, \dots, u_K^\star) = \mathrm{argmax}_{u \in \Sigma_{K+1}} \min_{b \neq 0} \frac{\Delta_b^2}{2\left(\frac{1}{u_0} + \frac{1}{u_b}\right)} \ .$$

The optimal weights for the active mode satisfy

$$\forall k \in \{0, \dots, K\},\ \forall i \leq J,\ w^\star_{k,i} = u^\star_k \frac{|\beta_i|}{\sum_{i=1}^J |\beta_i|} \ .$$

If, in addition $\boldsymbol{\alpha} = \boldsymbol{\beta}$, the above also holds for the agnostic and the proportional modes.

The interesting part of Proposition 2.9 is that computing $(u_0^\star, \dots, u_K^\star)$ can be done efficiently using Theorem 5 from [Garivier and Kaufmann, 2016]. The optimal weights of the ABC-S problem can be deduced from $u^\star$ without any further calculation.

2.4 Algorithms

To obtain our algorithms, we instantiate the Track-and-Stop algorithm template introduced and analyzed in [Garivier and Kaufmann, 2016] for the BAI setting and extend it to our ABC-S problem. We refer the reader to Section 1.2.3 for a brief overview of the tools used for the BAI task.

The Sampling Rule. The high level overview of the algorithm is as follows. We are given the number of arms K and subpopulations J, the exponential family, the mode of interaction, the subpopulation importance coefficients $\boldsymbol{\beta}$ and, for passive modes, their natural frequencies $\boldsymbol{\alpha}$. The algorithm then proceeds in rounds $t = 1, 2, \dots$ Each round t, it calculates the empirical frequencies $\hat{\boldsymbol{\mu}}_t \in \mathbb{R}^{(K+1)\times J}$ given by

$$\hat{\mu}_{k,i}(t) = \frac{1}{N_{k,i}(t)} \sum_{s=1}^{t} X_s \mathbf{1}\left\{A_s = k, I_s = i\right\} \ .$$

It then computes (a suitable approximation of) the maximiser (i.e. the oracle policy) $\boldsymbol{w}_t = \boldsymbol{w}^*(\hat{\boldsymbol{\mu}}_t) \in \Sigma_{(K+1)\times J}$ of problem Equation (2.2). In the active mode, we "D-track" $\boldsymbol{w}_t$, i.e. we sample $(A_t, I_t) \in \mathrm{argmax}_{k,i} N_{k,i}(t-1) - t\boldsymbol{w}_t(k,i)$. In the proportional mode, the subpopulation I_t is given and we "D-track" the conditional distribution of $\boldsymbol{w}_t$ on arms given the subpopulation,

i.e. $A_t \in \operatorname{argmax}_k N_{k,I_t}(t-1) - t\alpha_{I_t}\boldsymbol{w}_t(k|I_t)$, where $\boldsymbol{w}_t(k,i) = \alpha_i \boldsymbol{w}_t(k|i)$. In the agnostic mode we "D-track" the marginal distribution of $\boldsymbol{w}_t$ on arms, i.e. $A_t \in \operatorname{argmax}_k N_k(t-1) - t\boldsymbol{w}_t(k)$. For each mode, this sampling strategy ensures that $N_{k,i}(t) \approx t\boldsymbol{w}_t(k,i) \approx tw^*_{k,i}(\boldsymbol{\mu})$, thus driving down the reported level of confidence as quickly as possible given the lower bound from Theorem 2.3.

The Recommendation. Concluding each round, we recommend

$$\hat{\mathcal{S}}(t) = \Big\{k \in \{1,\dots,K\} \;\Big|\; \sum_{i=1}^{J} \beta_i \hat{\mu}_{k,i}(t) > \sum_{i=1}^{J} \beta_i \hat{\mu}_{0,i}(t)\Big\} \tag{2.11}$$

at confidence level $\hat{\delta}(t) = \min\{\delta \in (0,1) | Z(t) \geq \beta(t,\delta)\}$ obtained by inverting the threshold $\beta(t,\delta)$ at the GLR statistic

$$Z(t) \;=\; \min_{b\neq 0} \inf_{\boldsymbol{\lambda}\in\mathcal{L}:\lambda_0=\lambda_b} \sum_{a\in\{0,b\}} \sum_{i=1}^{J} N_{a,i}(t) d(\hat{\mu}_{a,i}(t), \lambda_{a,i}) \;. \tag{2.12}$$

At first sight the expression of $Z(t)$ may seem coming out of nowhere. In Appendix 2.E.2, we explain in the single population case why it is quite natural to consider this quantity and the relation with its counterpart in the BAI setting from Equation (1.8).

The Threshold. For obtaining a valid threshold, we can rely on existing works that can be applied to our setting. For the sharpest theoretically supported thresholds we refer to [Kaufmann and Koolen, 2018]. Namely, an ABC-S problem with K-arms and J-subpopulations has 2^K answers, and its *rank* [Kaufmann and Koolen, 2018, Definition 22] is $2J$, as can be read off from Equation (2.4). By [Kaufmann and Koolen, 2018, Proposition 23] we have validity for $\beta(t,\delta) = 6J\ln\ln t + \ln\frac{1}{\delta} + K + 2J\cdot O(\ln\ln\frac{1}{\delta})$. In practice, we follow [Garivier and Kaufmann, 2016] and use instead the heavily stylized $\ln((1+\ln t)/\delta)$ that omits several union bounds.

2.4.1 Implementation Details

In this section we go into more details on the algorithm for each mode. Let us start with some notation. Let $\beta(t,\delta)$ be a threshold function. We denote the inverse of $\beta(t,\delta)$ in its second argument by

$$\beta^{-1}(t,Z) \;=\; \min\{\delta \in (0,1) | Z \geq \beta(t,\delta)\}\,.$$

We extend the definition of the GLR statistic to sample frequencies $\boldsymbol{w}$ and bandit $\boldsymbol{\mu}$ by

$$Z(\boldsymbol{w},\boldsymbol{\mu}) \;:=\; \min_{b\neq 0} \inf_{\boldsymbol{\lambda}\in\mathcal{L}:\lambda_0=\lambda_b} \sum_{a\in\{0,b\}} \sum_{i=1}^{J} w_{a,i} d(\mu_{a,i}, \lambda_{a,i}) \;,$$

so that the original definition Equation (2.12) is $Z(t) = Z(\boldsymbol{N}(t)/t, \hat{\boldsymbol{\mu}}(t))$. For any $\boldsymbol{\mu}$, we denote by $\nabla_{\boldsymbol{w}} Z(\boldsymbol{w},\boldsymbol{\mu})$ any sub-gradient of $\boldsymbol{w} \mapsto Z(\boldsymbol{w},\boldsymbol{\mu})$. We can obtain one such a sub-gradient by letting $(b,\boldsymbol{\lambda})$ be any minimiser of $Z(\boldsymbol{w},\boldsymbol{\mu})$, and constructing the vector with entry (a,i) given by

$$(a,i) \;\mapsto\; \begin{cases} d(\mu_{a,i},\lambda_{a,i}) & \text{if } a \in \{0,b\} \\ 0 & \text{otherwise}\,. \end{cases}$$

Our algorithms will make use of an online learning method (called $\mathcal{A}$ below) for linear losses defined on the simplex. This online learning task is known as the Hedge or Experts setting in

the literature. We will make use of AdaHedge [De Rooij et al., 2014], as it adapts automatically to the range of the losses and does not require tuning. Our methods for the active, proportional and agnostic modes are displayed as Algorithms 4, 5 and 6. Each algorithm consists of a Forced Exploration part, which serves to ensure that the empirical estimate of the bandit model converges, i.e. $\hat{\boldsymbol{\mu}}(t) \to \boldsymbol{\mu}$. By forcing exploration sub-linearly often, the main term in the sample complexity is unaffected asymptotically. Each algorithm further makes use of online learning to compute $\boldsymbol{w}^*(\boldsymbol{\mu})$. In the notation of this section, we have

$$\boldsymbol{w}^*(\boldsymbol{\mu}) \;=\; \operatorname*{argmax}_{\boldsymbol{w}\in\mathcal{C}} \; Z(\boldsymbol{w}, \boldsymbol{\mu}) \;.$$

Our approach to learning $\boldsymbol{w}^*(\boldsymbol{\mu})$ is to perform gradient steps on the plug-in loss function $\boldsymbol{w} \mapsto -Z(\boldsymbol{w}, \hat{\boldsymbol{\mu}}(t))$. It is in the convex domain $\mathcal{C} \subseteq \Sigma_{(K+1)\times J}$ that we see the main difference between the three modes. Recall from Section 2.3.3 that in the active mode $\boldsymbol{w}$ is not constrained further, in the proportional mode the subpopulation marginal of $\boldsymbol{w}$ must equal $\boldsymbol{\alpha}$, i.e. $\langle \mathbf{1}, \boldsymbol{w}\rangle = \boldsymbol{\alpha}$, and in the agnostic mode $\boldsymbol{w}$ must be the independent product $\boldsymbol{w} = \boldsymbol{v}\boldsymbol{\alpha}$ of some arm marginal $\boldsymbol{v} \in \Sigma_{K+1}$ and the subpopulation frequencies $\boldsymbol{\alpha}$. We hence need to design online learners for each of the three sets of constraints. In the active case, we have one learner $\mathcal{A}$ that learns the full joint $\boldsymbol{w}^*(a, j)$ directly, in the proportional case we use one learner $\mathcal{A}_j$ for each subpopulation $j \in \{1, \dots, J\}$ to learn the conditional distribution $\boldsymbol{w}^*(a|j)$, and in the agnostic case we again use one learner to learn the common marginal $\boldsymbol{w}^*(a)$. This difference is reflected in the loss function used in each mode, and hence in the gradient that is fed to each learner. In the active case we use the full $(K+1) \times J$ gradients

$$\boldsymbol{\ell}_t^{\text{active}} \;:=\; -\nabla_{\boldsymbol{w}} Z(\boldsymbol{w}_t, \hat{\boldsymbol{\mu}}(t)) \;.$$

In the proportional case we have $\boldsymbol{w}(a, i) = \boldsymbol{w}(a|i)\alpha_i$, and by the chain rule we hence have gradients

$$\boldsymbol{\ell}_t^{i,\text{proportional}} \;:=\; -\nabla_{\boldsymbol{w}(a|i)} Z(\boldsymbol{w}_t, \hat{\boldsymbol{\mu}}(t)) \;=\; -\alpha_i \nabla_{\boldsymbol{w}} Z(\boldsymbol{w}_t, \hat{\boldsymbol{\mu}}(t))\boldsymbol{e}_i \;,$$

where $\boldsymbol{e}_i$ is the i-th canonical vector from $\mathbb{R}^J$. Finally, in the agnostic case we have $\boldsymbol{w}(a, i) = w(a)\alpha_i$, and again by the chain rule we have

$$\boldsymbol{\ell}_t^{\text{agnostic}} \;:=\; -\nabla_{\boldsymbol{w}(a)} Z(\boldsymbol{w}_t, \hat{\boldsymbol{\mu}}(t)) \;=\; -\nabla_{\boldsymbol{w}} Z(\boldsymbol{w}_t, \hat{\boldsymbol{\mu}}(t))\boldsymbol{\alpha} \;.$$

In the following, we report the pseudo code of the algorithm in the active mode. The pseudo code for the agnostic and the proportional modes is deferred to Appendix 2.D with a discussion on the run time of the algorithms.

2.4.2 Theoretical Guarantees

We now establish the asymptotic optimality of our approach with the following theorem proved in Appendix 2.C.

Theorem 2.10. *For every mode, Subpopulation Track-and-Stop is safely calibrated (Equation 2.1). Moreover, Subpopulation Track-and-Stop is asymptotically optimal and matches the lower bound from Theorem 2.3, in the sense that*

$$\text{for every bandit } \boldsymbol{\mu} \in \mathcal{L},\; \lim_{\delta\to 0} \frac{\mathbb{E}[\tau_\delta]}{\ln(1/\delta)} = T^\star(\boldsymbol{\mu}) \;.$$

Input: K arms, $\beta(t,\delta)$ threshold, $\mathcal{A}$ online learner for $(K+1)\times J$ experts.
for $t = 1, 2, \ldots$ **do**
  if *any pair* (a,i) *has* $N_{a,i}(t-1) \leq \sqrt{t}$ **then**
    Pick A_t, I_t any such pair
  else
    Get $\boldsymbol{w}_t$ from online learner $\mathcal{A}$
    Pick $(A_t, I_t) \in \operatorname{argmin}_{a,i} N_{a,i}(t-1) - t w_t(a,i)$
  Send loss vector $\boldsymbol{\ell}_t = -\nabla_{\boldsymbol{w}} Z(\boldsymbol{w}_t, \hat{\boldsymbol{\mu}}(t))$ to $\mathcal{A}$
  Obtain sample X_t from ν_{A_t,I_t}
  Recommend $\hat{\mathcal{S}}(t)$ from Equation (2.11) at confidence $\delta_t = \beta^{-1}(t, Z(\boldsymbol{N}(t)/t, \hat{\boldsymbol{\mu}}(t)))$

Algorithm 4: Algorithm for Active Mode.

2.5 Experiments

We conduct numerical experiments[1] to evaluate the proposed algorithms, focusing on Bernoulli bandit models, which are ubiquitous in practical applications. In our experiments, in addition to our T-a-S algorithms with the various interaction modes, we include two more sampling rules for comparison: (1) uniform sampling as a baseline, and (2) the experimentally efficient *Best Challenger* (BC) heuristic inspired by [Garivier and Kaufmann, 2016], adapted to the ABC problem and denoted BC-ABC in the sequel. BC for the BAI problem samples in every round the empirical best arm $\hat{a}_t$ or its best challenger, i.e. the arm $\hat{c}_t \neq \hat{a}_t$ at which the GLR statistic (Equation 1.8) reaches its minimum. Our BC-ABC adaptation samples in every round the control arm or the arm that yields the minimum GLR statistic $Z(t)$ in the agnostic interaction mode (since $Z(t)$ is subpopulation independent). For clearer comparison between the sampling strategies, all algorithms use the Chernoff stopping criterion to determine either when to stop or output the risk assessment at a given time. We also opted for sampling rules independent from the confidence parameter δ, because we are aiming for safely calibrated policies.

We first illustrate the fact that the T-a-S algorithm provides a correct –but rather conservative– assessment of the risk of its decision whatever the time it is stopped at. To do so, we generated 1000 bandit instances uniformly at random from [0, 1] with $K = 2$ arms. For each instance, we recorded the first time a certain risk assessment level is reached and the correctness of the algorithm's recommendation at that point. We map to each risk assessment level the proportion of errors across all instances. We chose two stopping rates that are not supported by theory but are recommended in practice [Garivier and Kaufmann, 2016]. Figure 2.6 (Left) illustrates the isotonic curve fitted on our observations and suggests that even the most lenient stopping threshold $\ln((\ln(t)+1)/\delta)$ results in much lower empirical probability of error than the risk assessment. In the following, we use the stopping threshold $\ln((\ln(t)+1)/\delta)$.

In our second experiment, we generated 3000 Bernoulli bandit instances with $K = 2$ and a random number of subpopulations J between 2 and 10. Each subpopulation-arm's mean $\mu_{a,i}$ is drawn uniformly at random from $[0,1]$, and the subpopulation frequency vector $\boldsymbol{\alpha}$ is drawn from a Dirichlet(10) distribution. Table 2.1 reports the average stopping time of each algorithm across all bandit instances. On average, the T-a-S algorithms at all modes stop at similar times, and all adaptive sampling methods terminate faster than uniform sampling.

[1] Code at https://gitlab.com/ckatsimerou/abc_s_public

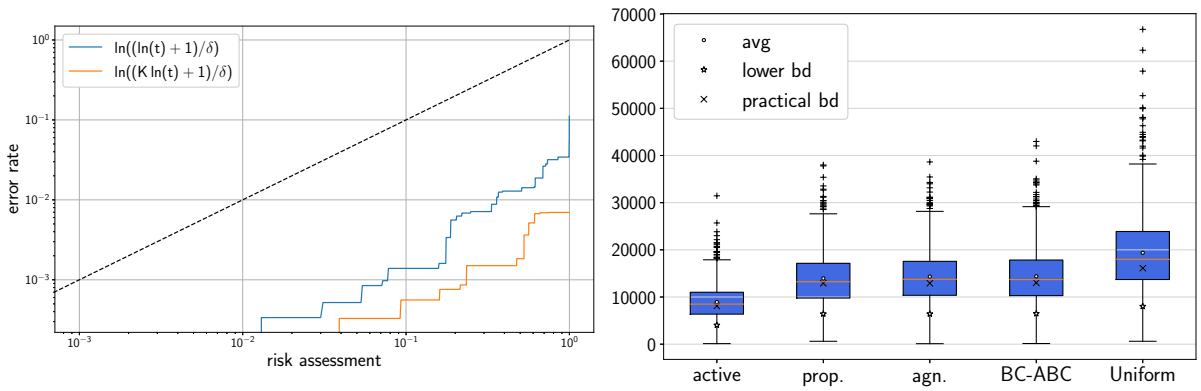


Figure 2.6: (Left) Risk assessment calibration on a log-log scale. (Right) Stopping time boxplot for $\boldsymbol{\mu} = [0.1\ 0.4\ 0.3; 0.2\ 0.5\ 0.2; 0.5\ 0.1\ 0.1] \in [0,1]^{(K+1)\times J}$ when $\boldsymbol{\beta} = [1/3, 1/3, 1/3], \boldsymbol{\alpha} = [0.4, 0.5, 0.1]$ with Bernoulli distributions.

Table 2.1: Average stopping time. Description in text.

T-a-S (active)	T-a-S (proportional)	T-a-S (agnostic)	BC-ABC	Uniform
14871	15231	15444	15279	21586

To better understand the role of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, we ran the algorithms on a specific model (see Figure 2.6, Right) with $\boldsymbol{\alpha} \neq \boldsymbol{\beta}$. In this case, the optimal proportions are constrained by the frequencies of the subpopulation for passive interaction modes. The expected number of samples needed to identify the ABC-S solution is lower for the active policy, which has an additional degree of freedom in its sampling strategy. The *proportional* interaction mode and the *agnostic* interaction modes perform similarly. As expected, all the proposed strategies outperform the uniform sampling rule. We contrast the stopping time with the lower bound $\mathrm{kl}(\delta, 1-\delta)T^*(\boldsymbol{\mu})$, and with a more practical version, which indicates, approximately, the first time at which the GLR statistic crosses the threshold, i.e. solving $t = \ln((\ln(t)+1)/\delta)T^*(\boldsymbol{\mu})$, as was done in [Degenne et al., 2019]. All adaptive algorithms perform well on this instance, with their average runtime being very close to their respective practical bound.

2.6 Conclusion

In this chapter, we considered the pure exploration task of identifying all the arms that are better than a control arm in the presence of subpopulations (ABC-S). We designed asymptotically optimal policies for this problem under different assumptions on the mode of interaction between the learner and the bandit. We observed that the *active* mode, in which the learner decides which subpopulation he samples, may significantly reduce decision times. On the other hand, the other modes, in which the learner has to respect the natural proportions of the different subpopulations (i.e., in *proportional* and *agnostic* modes) produce more modest effects, except when the subpopulations differ significantly in variances. Finally, we proposed a natural way to provide anytime decisions with risk guarantees in the Track-and-Stop framework.

Appendix

Appendix 2.A Optimal Allocations in the Gaussian Case for $K = 1$

Lemma 2.11. *When $K = 1$ with Gaussian distributions such that $\nu_{a,i} = \mathcal{N}(\mu_{a,i}, \sigma_{a,i}^2)$ the following holds*

$$\inf_{\boldsymbol{\lambda}:\lambda_0=\lambda_1} \sum_{i=1}^{J} w_{0,i} d(\mu_{0,i}, \lambda_{0,i}) + \sum_{i=1}^{J} w_{1,i} d(\mu_{1,i}, \lambda_{1,i}) = \frac{\Delta_1^2}{2\sum_{i=1}^{J} \beta_i^2 \left(\frac{\sigma_{0,i}^2}{w_{0,i}} + \frac{\sigma_{1,i}^2}{w_{1,i}}\right)} \,.$$

Proof. One has for $b \in \{0, 1\}$,

$$d(\mu_{b,i}, \lambda_{b,i}) = \frac{(\lambda_{b,i} - \mu_{b,i})^2}{2\sigma_{b,i}^2} \,.$$

Using the result from Theorem 2.3 for the case $K = 1$, the following holds

$$\inf_{\boldsymbol{\lambda} \in \mathrm{Alt}_{\boldsymbol{\beta}}(\boldsymbol{\mu})} \sum_{a=0}^{1} \sum_{i=1}^{J} w_{a,i} d(\mu_{a,i}, \lambda_{a,i}) = \min_{\boldsymbol{\lambda} \in \mathcal{L}:\lambda_0=\lambda_1} \sum_{a=0}^{1} \sum_{i=1}^{J} w_{a,i} d(\mu_{a,i}, \lambda_{a,i}) \,.$$

We introduce

$$L(\lambda_0, \lambda_1, q) = \sum_{i=1}^{J} w_{0,i} \frac{(\lambda_{0,i} - \mu_{0,i})^2}{2\sigma_{0,i}^2} + \sum_{i=1}^{J} w_{1,i} \frac{(\lambda_{1,i} - \mu_{1,i})^2}{2\sigma_{1,i}^2} + q \left(\sum_{i=1}^{J} \beta_i (\lambda_{0,i} - \lambda_{1,i}) \right) \,.$$

One has,

$$\min_{\boldsymbol{\lambda} \in \mathcal{L}:\lambda_0=\lambda_1} \sum_{a=0}^{1} \sum_{i=1}^{J} w_{a,i} d(\mu_{a,i}, \lambda_{a,i}) = \sup_{q \in \mathbb{R}} \inf_{\boldsymbol{\lambda} \in \mathcal{L}} L(\lambda_0, \lambda_1, q) \,.$$

Differentiating with respect to $\lambda_{0,i}$ and $\lambda_{1,i}$ brings the conditions

$$\lambda_{0,i} = \mu_{0,i} - \frac{q\beta_i\sigma_{0,i}^2}{w_{0,i}} \quad \text{and} \quad \lambda_{1,i} = \mu_{1,i} + \frac{q\beta_i\sigma_{1,i}^2}{w_{1,i}} \,.$$

Plugging these values back in L gives the function

$$f(q) = -\frac{q^2}{2} \sum_{i=1}^{J} \beta_i^2 \left(\frac{\sigma_{0,i}^2}{w_{0,i}} + \frac{\sigma_{1,i}^2}{w_{1,i}} \right) + q \sum_{i=1}^{J} \beta_i (\mu_{0,i} - \mu_{1,i}) \,.$$

Easy calculations show that the maximum of the function f is attained for

$$q^\star = \frac{\sum_{i=1}^{J} \beta_i (\mu_{0,i} - \mu_{1,i})}{\sum_{i=1}^{J} \beta_i^2 \left(\frac{\sigma_{0,i}^2}{w_{0,i}} + \frac{\sigma_{1,i}^2}{w_{1,i}} \right)} \,.$$

Plugging this value back in the expression of f,

$$f(q^\star) = \frac{\Delta_1^2}{2\sum_{i=1}^{J} \beta_i^2 \left(\frac{\sigma_{0,i}^2}{w_{0,i}} + \frac{\sigma_{1,i}^2}{w_{1,i}} \right)} \,.$$

□

Proposition 2.8. *For any $\boldsymbol{\mu} \in \mathcal{L}$ with $K = 1$ and $\nu_{k,i} = \mathcal{N}(\mu_{k,j}, \sigma_{k,j}^2)$, recalling that $\Delta_1 = \sum_{i=1}^J \beta_i(\mu_{0,i} - \mu_{1,i})$ one has*

1. *In the agnostic case, the characteristic time and the optimal weights satisfy*

$$T^\star_{\text{agnostic}}(\boldsymbol{\mu}) = \frac{2\left(\sqrt{\sum_{i=1}^J \frac{\beta_i^2\sigma_{0,i}^2}{\alpha_i}} + \sqrt{\sum_{i=1}^J \frac{\beta_i^2\sigma_{1,i}^2}{\alpha_i}}\right)^2}{\Delta_1^2}$$

$$\text{and } \forall i \le J, \forall k \in \{0,1\},\ w^\star_{k,i} = \frac{\alpha_i\sqrt{\sum_{i=1}^J \frac{\beta_i^2\sigma_{k,i}^2}{\alpha_i}}}{\sqrt{\sum_{i=1}^J \frac{\beta_i^2\sigma_{0,i}^2}{\alpha_i}} + \sqrt{\sum_{i=1}^J \frac{\beta_i^2\sigma_{1,i}^2}{\alpha_i}}}\,.$$

2. *In the proportional case, the characteristic time and the optimal weights satisfy*

$$T^\star_{\text{prop}}(\boldsymbol{\mu}) = \frac{2\sum_{i=1}^J \frac{\beta_i^2}{\alpha_i}(\sigma_{0,i} + \sigma_{1,i})^2}{\Delta_1^2}$$

$$\text{and } \forall i \le J, \forall k \in \{0,1\},\ w^\star_{k,i} = \frac{\alpha_i\sigma_{k,i}}{\sigma_{0,i} + \sigma_{1,i}}\,.$$

3. *In the active case, the characteristic time and the optimal weights satisfy*

$$T^\star_{\text{active}}(\boldsymbol{\mu}) = \frac{2\left(\sum_{i=1}^J |\beta_i|(\sigma_{0,i} + \sigma_{1,i})\right)^2}{\Delta_1^2}$$

$$\text{and } \forall i \le J, \forall k \in \{0,1\},\ w^\star_{k,i} = \frac{|\beta_i|\sigma_{k,i}}{\sum_{i=1}^J |\beta_i|(\sigma_{0,i} + \sigma_{1,i})}\,.$$

Proof.

Agnostic mode From the Lemma 2.4, we have

$$\begin{aligned}
T^\star_{\text{agnostic}}(\boldsymbol{\mu})^{-1} &= \sup_{\boldsymbol{w}\in\mathcal{C}_{\text{agnostic}}} \inf_{\boldsymbol{\lambda}\in \text{Alt}_{\boldsymbol{\beta}}(\boldsymbol{\mu})} \sum_{a=0}^{1}\sum_{i=1}^{J} w_{a,i} d(\mu_{a,i}, \lambda_{a,i}) \\
&= \sup_{\boldsymbol{w}\in\mathcal{C}_{\text{agnostic}}} \inf_{\boldsymbol{\lambda}:\lambda_0=\lambda_1} \sum_{a=0}^{1}\sum_{i=1}^{J} w_{a,i} d(\mu_{a,i}, \lambda_{a,i}) \quad \text{(Theorem 2.3)} \\
&= \sup_{\boldsymbol{w}\in\mathcal{C}_{\text{agnostic}}} \frac{\Delta_1^2}{2\sum_{i=1}^J \beta_i^2\left(\frac{\sigma_{0,i}^2}{w_{0,i}} + \frac{\sigma_{1,i}^2}{w_{1,i}}\right)} \quad \text{(Lemma 2.11)}\,.
\end{aligned}$$

$\boldsymbol{w} \in \mathcal{C}_{\text{agnostic}}$ implies $w_{a,i} = \alpha_i u_a$ with $(u_0, \dots, u_K) \in \Sigma_{K+1}$. For this reason,

$$\boldsymbol{w}^\star = \text{argmin}_{\boldsymbol{u}:u_0+u_1=1} \sum_{i=1}^J \frac{\beta_i^2}{\alpha_i}\left(\frac{\sigma_{0,i}^2}{u_0} + \frac{\sigma_{1,i}^2}{u_1}\right)\,.$$

We let $c_a \coloneqq \sum_{i=1}^J \frac{\beta_i^2\sigma_{a,i}^2}{\alpha_i}$ for $a \in \{0,1\}$. Plugging $u_1 = 1 - u_0$ in the previous expression and

differentiating with respect to u_0 brings the condition

$$u_0^2 + 2u_0 \frac{c_0}{c_1 - c_0} - \frac{c_0}{c_1 - c_0} .$$

Solving this polynomial and using that $\boldsymbol{u} \in \Sigma_2$ gives the unique solution

$$u_0^\star = \frac{\sqrt{c_0}}{\sqrt{c_0} + \sqrt{c_1}} .$$

Implying,

$$w_{0,i}^\star = \alpha_i \frac{\sqrt{\sum_{i=1}^J \frac{\beta_i^2 \sigma_{0,i}^2}{\alpha_i}}}{\sqrt{\sum_{i=1}^J \frac{\beta_i^2 \sigma_{0,i}^2}{\alpha_i}} + \sqrt{\sum_{i=1}^J \frac{\beta_i^2 \sigma_{1,i}^2}{\alpha_i}}} \quad \text{and} \quad w_{1,i}^\star = \alpha_i \frac{\sqrt{\sum_{i=1}^J \frac{\beta_i^2 \sigma_{1,i}^2}{\alpha_i}}}{\sqrt{\sum_{i=1}^J \frac{\beta_i^2 \sigma_{0,i}^2}{\alpha_i}} + \sqrt{\sum_{i=1}^J \frac{\beta_i^2 \sigma_{1,i}^2}{\alpha_i}}} .$$

With those values,

$$T_{\text{agnostic}}^\star(\boldsymbol{\mu}) = \frac{2\left(\sqrt{\sum_{i=1}^J \frac{\beta_i^2 \sigma_{0,i}^2}{\alpha_i}} + \sqrt{\sum_{i=1}^J \frac{\beta_i^2 \sigma_{1,i}^2}{\alpha_i}}\right)^2}{\Delta_1^2} .$$

Proportional mode Following the same line of proof, gives

$$T_{\text{prop}}^\star(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{w} \in \mathcal{C}_{\text{prop}}} \frac{\Delta_1^2}{2\sum_{i=1}^J \beta_i^2 \left(\frac{\sigma_{0,i}^2}{w_{0,i}} + \frac{\sigma_{1,i}^2}{w_{1,i}}\right)} .$$

The main difference is now on the constraints on the weights. In the proportional mode, following Lemma 2.5, $\forall i \leq J, \sum_{a=0}^1 w_{a,i} = \alpha_i$. We consider the Lagrangian function:

$$L(w_0, w_1, q_1, \ldots, q_J) = \sum_{i=1}^J \beta_i^2 \left(\frac{\sigma_{0,i}^2}{w_{0,i}} + \frac{\sigma_{1,i}^2}{w_{1,i}}\right) + \sum_{i=1}^J q_i \left(\sum_{a \in \{0,1\}} w_{a,i} - \alpha_i\right) .$$

Differentiating with respect to $w_{0,i}$ and $w_{1,i}$ gives the constraints:

$$\frac{-\beta_i^2 \sigma_{0,i}^2}{w_{0,i}^2} + q_i = 0 \quad \text{and} \quad \frac{-\beta_i^2 \sigma_{1,i}^2}{w_{1,i}^2} + q_i = 0 .$$

From which we can deduce

$$\frac{w_{0,i}}{\sigma_{0,i}} = \frac{w_{1,i}}{\sigma_{1,i}} .$$

From $w_{0,i} + w_{1,i} = \alpha_i$, we deduce,

$$q_i^\star = \frac{\beta_i^2 (\sigma_{0,i} + \sigma_{1,i})^2}{\alpha_i^2} .$$

Plugging this value in the first constraint gives

$$w_{0,i}^\star = \alpha_i \frac{\sigma_{0,i}}{\sigma_{0,i} + \sigma_{1,i}} \quad \text{and} \quad w_{1,i}^\star = \alpha_i \frac{\sigma_{1,i}}{\sigma_{0,i} + \sigma_{1,i}} .$$

Using those weights,

$$T_{\text{prop}}^\star(\boldsymbol{\mu}) = \frac{2\sum_{i=1}^J \frac{\beta_i^2}{\alpha_i} (\sigma_{0,i} + \sigma_{1,i})^2}{\Delta_1^2} .$$

Active mode Following the proof of Proposition 2.8, one has

$$T^{\star}_{\text{active}}(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{w}\in\Sigma_{(K+1)J}} \frac{\Delta_1^2}{2\sum_{i=1}^{J}\beta_i^2\left(\frac{\sigma_{0,i}^2}{w_{0,i}}+\frac{\sigma_{1,i}^2}{w_{1,i}}\right)} . \tag{2.13}$$

Using the constraint $\boldsymbol{w} \in \Sigma_{(K+1)J}$, one gets

$$w_{1,J} = 1 - \sum_{a=\{0,1\}}\sum_{i=1}^{J-1} w_{a,i} - w_{0,J} . \tag{2.14}$$

We need to minimize the function (where $w_{1,J}$ has been replaced by the expression from Equation 2.14)

$$f(\boldsymbol{w}) = \sum_{a=\{0,1\}}\sum_{i=1}^{J-1}\beta_i^2\frac{\sigma_{a,j}^2}{w_{a,j}} + \beta_J^2\frac{\sigma_{0,J}^2}{w_{0,J}} + \beta_J^2\frac{\sigma_{1,J}^2}{1-\sum_{a=0}^{1}\sum_{i=1}^{J-1}w_{a,i}-w_{0,J}} .$$

For $i \leq J-1$, taking the derivative with respect to $w_{0,i}$ and $w_{1,i}$ gives the following constraints

$$\beta_i^2\sigma_{0,i}^2\left(1-\sum_{a=0}^{1}\sum_{i=1}^{J-1}w_{a,i}-w_{0,J}\right)^2 = \beta_J^2\sigma_{1,J}^2w_{0,i}^2 ,$$

$$\beta_j^2\sigma_{1,i}^2\left(1-\sum_{a=0}^{1}\sum_{i=1}^{J-1}w_{a,i}-w_{0,J}\right)^2 = \beta_J^2\sigma_{1,J}^2w_{1,i}^2 .$$

From which we deduce

$$\forall i \leq J-1, \quad \frac{w_{0,i}}{\sigma_{0,i}} = \frac{w_{1,i}}{\sigma_{1,i}} . \tag{2.15}$$

Differentiating with respect to $w_{1,J}$ gives

$$\sigma_{0,J}\left(1-\sum_{a=0}^{1}\sum_{i=1}^{J-1}w_{a,i}-w_{0,J}\right) = \sigma_{1,J}w_{0,J} .$$

Rearranging and using Equation 2.15 gives,

$$w_{0,J} = \frac{\sigma_{0,J}}{\sigma_{0,J}+\sigma_{1,J}} - \sum_{i=1}^{J-1}\frac{w_{0,i}}{\sigma_{0,i}}\frac{\sigma_{0,i}+\sigma_{1,i}}{\sigma_{0,J}+\sigma_{1,J}}\sigma_{0,J} . \tag{2.16}$$

Using Equation 2.15 and Equation 2.16, we define the function

$$g(w_{0,1},\dots,w_{0,J-1}) = \sum_{i=1}^{J-1}\beta_i^2\frac{\sigma_{0,i}}{w_{0,i}}(\sigma_{0,i}+\sigma_{1,i}) + \frac{\beta_J^2(\sigma_{0,J}+\sigma_{1,J})^2}{1-\sum_{i=1}^{J-1}\frac{w_{0,i}(\sigma_{0,i}+\sigma_{1,i})}{\sigma_{0,i}}} .$$

Differentiating with respect to $w_{0,i}$ for $i \leq J-1$ brings

$$\forall i \leq J-1, \quad \frac{|\beta_i|\sigma_{0,i}}{\sigma_{0,J}+\sigma_{1,J}}\left(1-\sum_{i=1}^{J-1}\frac{w_{0,i}}{\sigma_{0,i}}(\sigma_{0,i}+\sigma_{1,i})\right) = |\beta_J|w_{0,i} . \tag{2.17}$$

Multiplying both sides of this equation by $(\sigma_{0,i} + \sigma_{1,i})/\sigma_{0,i}$ and summing for $i \leq J-1$,

$$\sum_{i=1}^{J-1} \frac{w_{0,i}}{\sigma_{0,i}} (\sigma_{0,i} + \sigma_{1,i}) = \frac{\sum_{i=1}^{J-1} |\beta_i|(\sigma_{0,i} + \sigma_{1,i})}{\sum_{i=1}^{J} |\beta_i|(\sigma_{0,i} + \sigma_{1,i})} .$$

Plugging this value back in Equation 2.17 one has:

$$\forall i \leq J-1, \quad w_{0,i} = \frac{|\beta_i|\sigma_{0,i}}{\sum_{i=1}^{J} |\beta_i|(\sigma_{0,i} + \sigma_{1,i})} .$$

From Equation 2.15, we deduce,

$$\forall i \leq J-1, \quad w_{1,i} = \frac{|\beta_i|\sigma_{1,i}}{\sum_{i=1}^{J} |\beta_i|(\sigma_{0,i} + \sigma_{1,i})} .$$

We obtain the value of $w_{0,J}$ using Equation 2.16 and that of $w_{1,J}$ using Equation 2.14. Plugging those weights in the expression given by Equation 2.13 yields the characteristic time. □

Appendix 2.B Optimal Allocation in the Gaussian Case with $K > 1$

Proposition 2.9 (Efficient computation of the optimal weights in the Gaussian case)**.** *Assume Gaussian distributions with a known variance σ^2, and let*

$$(u_0^\star, \dots, u_K^\star) = \operatorname{argmax}_{u \in \Sigma_{K+1}} \min_{b \neq 0} \frac{\Delta_b^2}{2\left(\frac{1}{u_0} + \frac{1}{u_b}\right)} .$$

The optimal weights for the active mode satisfy

$$\forall k \in \{0, \dots, K\},\ \forall i \leq J,\ w_{k,i}^\star = u_k^\star \frac{|\beta_i|}{\sum_{i=1}^{J} |\beta_i|} .$$

If, in addition $\boldsymbol{\alpha} = \boldsymbol{\beta}$, the above also holds for the agnostic and the proportional modes.

Proof. From Lemma 2.11, when the distribution are Gaussian with a known variance σ^2 one has

$$T^\star_{\text{active}}(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{w} \in \Sigma_{(K+1)J}} \min_{b \neq 0} \frac{\Delta_1^2}{2\sum_{i=1}^{J} \beta_i^2 \left(\frac{\sigma^2}{w_{0,i}} + \frac{\sigma^2}{w_{b,i}}\right)} .$$

Using the same continuity argument than in [Garivier and Kaufmann, 2016], we know that the supremum of $\boldsymbol{w}$ is attained and is indeed a maximum. Let

$$\Lambda_b(v, w) := \frac{\Delta_b^2}{2\sum_{i=1}^{J} \beta_i^2 \left(\frac{\sigma^2}{v_i} + \frac{\sigma^2}{w_i}\right)} .$$

Then

$$\begin{aligned}\max_{\boldsymbol{w}\in\Sigma_{(K+1)J}} \min_{b\neq 0} \Lambda_b(w_0,w_b) &= \max_{\substack{u\in\Sigma_{K+1}\\ \forall a,\sum_{i=1}^J w_{a,i}=u_a}} \min_{b\neq 0} \Lambda_b(w_0,w_b)\\ &= \max_{u\in\Sigma_{K+1}} \max_{\substack{w\in\Sigma_{(K+1)J}\\ \forall a,\sum_i w_{a,i}=u_a}} \min_{b\neq 0} \Lambda_b(w_0,w_b)\\ &\leq \max_{u\in\Sigma_{K+1}} \min_{b\neq 0} \max_{\substack{w\in\Sigma_{(K+1)J}\\ \forall a,\sum_i w_{a,i}=u_a}} \Lambda_b(w_0,w_b) \quad \text{(Max-min inequality)}\,.\end{aligned}$$

Let $b \neq 0$,

$$\max_{\substack{w\in\Sigma_{(K+1)J}\\ \forall a,\sum_i w_{a,i}=u_a}} \Lambda_b(w_0,w_b) = \max_{\substack{w\in\Sigma_{(K+1)J}\\ \forall a,\sum_i w_{a,i}=u_a}} \frac{\Delta_b^2}{2\sum_{i=1}^J \beta_i^2\left(\frac{\sigma^2}{w_{0,i}}+\frac{\sigma^2}{w_{b,i}}\right)}\,.$$

Equivalently, we are interested in

$$\min_{\substack{w\in\Sigma_{(K+1)J}\\ \forall a,\sum_i w_{a,i}=u_a}} \sum_{i=1}^J \beta_i^2\left(\frac{\sigma^2}{w_{0,i}}+\frac{\sigma^2}{w_{b,i}}\right)\,.$$

We introduce the associated Lagrangian function

$$f(w,q) = \sum_{i=1}^J \beta_i^2\left(\frac{1}{w_{0,i}}+\frac{1}{w_{b,i}}\right) + \sum_{a=0}^K q_a\left(\sum_{i=1}^J w_{a,i}-u_a\right)\,.$$

Taking the derivative with respect to $w_{0,i}$ and $w_{b,i}$ for the different values of i yields

$$w_{0,i} = \frac{|\beta_i|}{\sqrt{q_0}} \quad \text{and} \quad w_{b,i} = \frac{|\beta_i|}{\sqrt{q_b}}\,.$$

Summing over i implies that

$$\sqrt{q_0} = \frac{\sum_{i=1}^J |\beta_i|}{u_0} \quad \text{and} \quad \sqrt{q_b} = \frac{\sum_{i=1}^J |\beta_i|}{u_b}\,,$$

and plugging the above in the expression of the weights yields

$$w_{0,i} = \frac{|\beta_i|}{\sum_{i=1}^J |\beta_i|}u_0 \quad \text{and} \quad w_{b,i} = \frac{|\beta_i|}{\sum_{i=1}^J |\beta_i|}u_b\,.$$

In particular,

$$\max_{\substack{w\in\Sigma_{(K+1)J}\\ \forall a,\sum_i w_{a,i}=u_a}} \frac{\Delta_b^2}{2\sum_{i=1}^J \beta_i^2\left(\frac{\sigma^2}{w_{0,i}}+\frac{\sigma^2}{w_{b,i}}\right)} = \frac{\Delta_b^2}{2\sigma^2\left(\sum_{i=1}^J |\beta_i|\right)^2\left(\frac{1}{u_0}+\frac{1}{u_b}\right)}\,, \tag{2.18}$$

yielding

$$\max_{\boldsymbol{w}\in\Sigma_{(K+1)J}} \min_{b\neq 0} \frac{\Delta_b^2}{2\sum_{i=1}^J \beta_i^2\left(\frac{\sigma^2}{w_{0,i}}+\frac{\sigma^2}{w_{b,i}}\right)} \leq \frac{1}{2\sigma^2\left(\sum_{i=1}^J |\beta_i|\right)^2} \max_{u\in\Sigma_{(K+1)J}} \min_{b\neq 0} \frac{\Delta_b^2}{\left(\frac{1}{u_0}+\frac{1}{u_b}\right)}\,.$$

On the other hand, letting $w_{a,i} = \frac{|\beta_i|}{\sum_{i=1}^J |\beta_i|} u_a$ with $\sum_a u_a = 1$ we have

$$\frac{1}{2\sigma^2 \left(\sum_{i=1}^J |\beta_i|\right)^2} \max_{u \in \Sigma_{(K+1)J}} \min_{b \neq 0} \frac{\Delta_b^2}{\left(\frac{1}{u_0} + \frac{1}{u_b}\right)} \leq \max_{\boldsymbol{w} \in \Sigma_{(K+1)J}} \min_{b \neq 0} \frac{\Delta_b^2}{2\sum_{i=1}^J \beta_i^2 \left(\frac{\sigma^2}{w_{0,i}} + \frac{\sigma^2}{w_{b,i}}\right)} ,$$

showing that the two optimization programs are equivalent and that when denoting

$$(u_0^\star, \dots, u_K^\star) = \operatorname{argmax}_{u \in \Sigma_{(K+1)J}} \min_{b \neq 0} \frac{\Delta_b^2}{\left(\frac{1}{u_0} + \frac{1}{u_b}\right)} ,$$

one has

$$\forall a \in \{0, \dots, K\},\ \forall i \leq J,\ w_{a,i}^\star = u_a^\star \frac{|\beta_i|}{\sum_{i=1}^J |\beta_i|} .$$

This corresponds to the optimal allocation strategy in the *active* mode. Recalling that when $\alpha = \beta$, the optimal weights for the *active* mode satisfy both $\mathcal{C}_{\text{prop}}$ and $\mathcal{C}_{\text{agnostic}}$ completes the proof. □

Appendix 2.C Proof of Theorem 2.10

In this section we show that T-a-S with C-tracking [Garivier and Kaufmann, 2016] and a certain threshold $\beta(t, \delta)$ is safely calibrated and asymptotically optimal. This is an important sanity check to validate our approach theoretically. Note that for the experimental validation we have explored a practically appealing variant of this algorithm: we employ an iterative scheme to approximate $\boldsymbol{w}^*(\hat{\boldsymbol{\mu}}(t))$, use D-tracking, and stylise the threshold.

Safe calibration follows from the definition of the recommendation rule (we report the answer $\mathcal{S}_\beta(\hat{\boldsymbol{\mu}}(t))$ at the empirical estimate $\hat{\boldsymbol{\mu}}(t)$ of the bandit instance), together with the computation of the risk assessment $\hat{\delta}_t$. It does not depend on the sampling rule. Our confidence level $\hat{\delta}_t$ is obtained by inverting the threshold $\beta(t, \delta)$ at the GLR statistic Equation (2.12). Safe calibration then follows from an anytime-valid GLR deviation inequality with boundary $\beta(t, \delta)$. We refer to [Kaufmann and Koolen, 2018, Proposition 23] for a boundary that is, in case of the ABC-S problem, of order $\ln \frac{1}{\delta} + K + 2J \cdot O(\ln \ln \frac{t}{\delta})$.

It remains to argue that the T-a-S sampling rule converges to the oracle weights. The original T-a-S proof for the BAI problem is due to [Garivier and Kaufmann, 2016, Theorem 14]. An upgrade to any single-answer problem, including our ABC-S, is due to [Degenne and Koolen, 2019]. For active mode, their theorem applies directly, while for agnostic mode it applies with the pair (I_t, X_t) regarded as the observation. We get:

Theorem 2.12 ([Degenne and Koolen, 2019, Theorems 7 and 10])**.** *For all ABC-S instances $\boldsymbol{\mu} \in \mathcal{L}$ in active mode and agnostic mode, Track-and-Stop with C-tracking and stopping threshold $\beta(t, \delta) = \ln(t^2/\delta) + O(1)$ is δ-correct with asymptotically optimal sample complexity.*

In *proportional* mode, we have the additional constraint that the learner chooses its arm in response to seeing (but not controlling) the subpopulation I_t. Still, the tracking convergence result [Degenne and Koolen, 2019, Lemma 6] goes through, upon observing that the empirical distribution of I_t converges to $\boldsymbol{\alpha}$ by the law of the large numbers, and hence our conditional tracking (see "sampling rule" in Section 2.4) adds the right conditional to the right marginal.

All in all, the computed joint weights converge to the joint $\boldsymbol{w}^*_{\text{prop}}(\boldsymbol{\mu})$, and tracking makes the sampling proportions also converge there.

We conclude with a remark on our use of D-tracking. Recall that D-tracking is the idea of advancing $N_k(t)$ towards t times the most current oracle weights, i.e. $tw^*_k(\hat{\boldsymbol{\mu}}(t))$, while C-tracking makes $N_k(t)$ advance towards the sum of encountered oracle weights, i.e. $\sum_{s=1}^t w^*_k(\hat{\boldsymbol{\mu}}(s))$. As argued in [Degenne et al., 2019, Appendix E], D-tracking can fail to make $N_k(t)/t$ converge to $w^*_k(\boldsymbol{\mu})$. However, this requires that the maximiser of the lower bound problem is not unique at $\boldsymbol{\mu}$ (as we are maximising a concave function, the set of maximisers is always convex). Here we argue that such a situation does not occur for the ABC-S problem. To see why, note that the lower bound objective, as a function of $\boldsymbol{w}$, is strictly concave. It suffices to show this for the *active* mode problem, as the problems for the other modes are further constrained maximisation problems of the same objective.

Lemma 2.13. *Fix a bandit instance $\boldsymbol{\mu} \in \mathcal{L}$. Let $\lambda \mapsto d(\mu_{k,j}, \lambda)$ be a strongly convex function for each arm k and subpopulation j. Then for the ABC-S problem with $\boldsymbol{\beta}$ such that $\beta_j \neq 0$ for all j, the oracle weights $\boldsymbol{w}^*(\boldsymbol{\mu})$ are unique.*

Proof. Let $\boldsymbol{w}^*(\boldsymbol{\mu})$ be any oracle weights at $\boldsymbol{\mu}$. We will show the lower bound objective Equation (2.4) is strictly concave as a function of $\boldsymbol{w}$ around $\boldsymbol{w}^*(\boldsymbol{\mu})$, so that $\boldsymbol{w}^*(\boldsymbol{\mu})$ was in fact unique. For each $k > 0$, let $\boldsymbol{\lambda}^k$ be the minimiser in $\text{Alt}^k(\boldsymbol{\mu})$ of the weighted divergence in Equation (2.4).

We perform a second-order Taylor expansion of the inner objective around $\boldsymbol{\lambda}^k$, which is a good approximation near $\boldsymbol{\lambda}^k$ (which is, after all, what matters when reasoning about $\boldsymbol{w}$ near $\boldsymbol{w}^*(\boldsymbol{\mu})$). To this end, let us abbreviate the divergences, and their first and second derivatives in their second argument by $d^k_{aj} \coloneqq d(\mu_{a,j}, \lambda^k_{a,j})$, $g^k_{aj} \coloneqq d'(\mu_{a,j}, \lambda^k_{a,j})$ and $h^k_{aj} \coloneqq d''(\mu_{a,j}, \lambda^k_{a,j})$, which all depend on $\boldsymbol{\lambda}^k$. A second-order Taylor expansion of the inner objective of Equation (2.4) around $\boldsymbol{\lambda}^k$ yields

$$\inf_{\boldsymbol{\lambda} \in \text{Alt}_k} \sum_{a,j} w_{a,j} d(\mu_{a,j}, \lambda_{a,j}) \approx \sum_{a \in \{0,k\}, j} w_{a,j} \left(d^k_{aj} - \frac{(g^k_{aj})^2}{2h^k_{aj}} \right) + \frac{\left(\sum_j \beta_j \left(\frac{g^k_{0j}}{h^k_{0j}} - \frac{g^k_{kj}}{h^k_{kj}} \right) \right)^2}{2 \sum_{a \in \{0,k\}, j} \frac{\beta_j^2}{w_{a,j} h^k_{aj}}}$$

where the optimiser is given by

$$\lambda_{a,j} = \lambda^k_{a,j} - \frac{g^k_{aj}}{h^k_{aj}} + \frac{\beta_j (\delta_{a=0} - \delta_{a=k})}{w_{a,j} h^k_{aj}} \frac{\sum_j \beta_j \left(\frac{g^k_{0j}}{h^k_{0j}} - \frac{g^k_{kj}}{h^k_{kj}} \right)}{\sum_{a \in \{0,k\}, j} \frac{\beta_j^2}{w_{a,j} h^k_{aj}}} .$$

Due to the last term, each of these is a strictly concave function of $w_{a,j}$ for $a \in \{0, k\}$ and all $j \leq J$ (here we use $\beta_j \neq 0$ and strong convexity $h^k_{aj} > 0$).
Now we still need to consider the $\max_{\boldsymbol{w} \in \Sigma_{(K+1) \times J}} \min_{k>0}$ problem. Let's convexify this for the inside finite min, and min-max swap to get a problem of the form $\min_{\boldsymbol{q} \in \Sigma_K} \max_{\boldsymbol{w} \in \Sigma_{(K+1) \times J}}$. Fixing the minimax outer strategy for $\boldsymbol{q}$, we find that $\boldsymbol{w}$ is the maximiser of the strictly concave function

$$\boldsymbol{w} \mapsto \sum_{k>0} q_k \left(\sum_{a \in \{0,k\}, j} w_{a,j} \left(d^k_{aj} - \frac{(g^k_{aj})^2}{2h^k_{aj}} \right) + \frac{\left(\sum_j \beta_j \left(\frac{g^k_{0j}}{h^k_{0j}} - \frac{g^k_{kj}}{h^k_{kj}} \right) \right)^2}{2 \sum_{a \in \{0,k\}, j} \frac{\beta_j^2}{w_{a,j} h^k_{aj}}} \right)$$

To complete the argument, we argue that $q_k > 0$ for all $k > 0$, or, equivalently, that at $\boldsymbol{w}^*$ the $\min_{k>0}$ are all equalised. As if not, we can move mass from $w_{k,j}$ for the higher $k > 0$ to $w_{k',j}$ for the lower k' and increase the objective value. This then proves that $\boldsymbol{w}^*(\boldsymbol{\mu})$ is unique, as the objective function is bounded above by a strictly concave function itself maximised at $\boldsymbol{w} = \boldsymbol{w}^*(\boldsymbol{\mu})$. □

Appendix 2.D Algorithm Details

Input: K arms, $\beta(t, \delta)$ threshold, J online learners $\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(J)}$ for $(K+1)$ experts each.
for $t = 1, 2, \ldots$ **do**
  See $I_t \sim \boldsymbol{\alpha}$
  if *any arm* k *has* $N_{k,I_t}(t-1) \leq \sqrt{\sum_{k=1}^{K} N_{k,I_t}(t-1)}$ **then**
    Pick A_t any such arm
  else
    Get $\boldsymbol{w}_t^{(i)}$ from each online learner $\mathcal{A}^{(i)}$
    Pick $A_t \in \operatorname{argmin}_k N_{k,I_t}(t-1) - t w_t(k, I_t)$
  Obtain sample X_t from ν_{A_t,I_t}
  For $i \leq J$ send $\boldsymbol{\ell}_t^{(i)} = -\alpha_i \nabla_{\boldsymbol{w}} Z\left([\alpha_1 \boldsymbol{w}_t^{(1)} \cdots \alpha_J \boldsymbol{w}_t^{(J)}], \hat{\boldsymbol{\mu}}(t)\right) \boldsymbol{e}_i$ to $\mathcal{A}^{(i)}$
  Recommend $\hat{\mathcal{S}}(t)$ from Equation (2.11) at confidence $\delta_t = \beta^{-1}(t, Z(\boldsymbol{N}(t)/t, \hat{\boldsymbol{\mu}}(t)))$

Algorithm 5: Algorithm for Proportional Mode.

Input: K arms, $\beta(t, \delta)$ threshold, Online learner $\mathcal{A}$ for $(K+1)$ experts.
for $t = 1, 2, \ldots$ **do**
  if *any arm* a *has* $\sum_{j=1}^{J} N_{a,j}(t-1) \leq \sqrt{t}$ **then**
    Pick A_t any such arm
  else
    Get $\boldsymbol{w}_t$ from online learner $\mathcal{A}$
    Pick $A_t \in \operatorname{argmin}_a \sum_{j=1}^{J} N_{a,j}(t-1) - t w_t(a)$
  See $I_t \sim \boldsymbol{\alpha}$
  Obtain sample X_t from ν_{A_t,I_t}
  Send loss vector $\boldsymbol{\ell}_t = -\nabla_{\boldsymbol{w}} \Lambda\left(\boldsymbol{w}_t \boldsymbol{\alpha}^\top, \hat{\boldsymbol{\mu}}(t)\right) \boldsymbol{\alpha}$ to $\mathcal{A}$
  Recommend $\hat{\mathcal{S}}(t)$ from Equation (2.11) at confidence $\delta_t = \beta^{-1}(t, Z(\boldsymbol{N}(t)/t, \hat{\boldsymbol{\mu}}(t)))$

Algorithm 6: Algorithm for Agnostic Mode.

Run Time In each of the three modes, our algorithms evaluate Λ for the confidence in the recommendation, compute one sub-gradient of Λ for the loss function, and spend $O(K \times J)$ time bookkeeping. Evaluation and sub-gradient computation for Λ boil down to solving a convex minimisation problem with an equality constraint. We use Newton's method with backtracking line search to find the minimiser given b. Each Newton iteration takes $O(J^2)$ time (recall that only 2 arms are involved), and we never needed more than 40. Doing this K times for the explicit minimum over b yields a total per iteration run time of $O(KJ^2)$.

Appendix 2.E Miscellaneous

2.E.1 Sub-gaussianity and Mixtures

Except in the case of Bernoulli distributions, where the mixture is also a Bernoulli distribution, finding a strategy that matches $T^\star_{\text{oblivious}}(\boldsymbol{\mu})^{-1}$ is a hard task. However, one may use the following lemma to treat the mixture in a sub-optimal way, based on the fact that it exhibits sub-gaussian behavior.

Lemma 2.14 (Sub-gaussianity of mixture)**.** *For each $\mu \in \mathbb{R}$, assume that ν_μ is a distribution on $\mathbb{R}$ with mean $\mathbb{E}_{X\sim\nu_\mu}[X] = \mu$ that is σ-subgaussian, meaning that $\mathbb{E}_{X\sim\nu_\mu}\left[e^{\lambda(X-\mu)}\right] \leq e^{\sigma^2\lambda^2/2}$ for any $\lambda \in \mathbb{R}$. Further let $\alpha(\mu)$ be a prior on μ with mean m that is itself η-subgaussian, meaning that $\mathbb{E}_{\mu\sim\alpha}\left[e^{\lambda(\mu-m)}\right] \leq e^{\lambda^2\eta^2/2}$. Then the mixture distribution $Q = \mathbb{E}_{\mu\sim\alpha}[\nu_\mu]$ is $\sqrt{\sigma^2+\eta^2}$-subgaussian.*

Proof. The mixture distribution obviously has mean $\mathbb{E}_{X\sim Q}[X] = m$ and

$$\begin{aligned}\mathbb{E}_{X\sim Q}\left[e^{\lambda(X-m)}\right] &= \mathbb{E}_{\mu\sim\alpha}\left[e^{\lambda(\mu-m)}\mathbb{E}_{X\sim\nu_\mu}\left[e^{\lambda(X-\mu)}\right]\right] \leq \mathbb{E}_{\mu\sim\alpha}\left[e^{\lambda(\mu-m)}\right]e^{\sigma^2\lambda^2/2} \\ &\leq e^{(\sigma^2+\eta^2)\lambda^2/2}.\end{aligned}$$

□

In particular, if α is supported on $[\pm M]$, then α is M-subgaussian, and hence Q is $\sqrt{\sigma^2+M^2}$-subgaussian.

2.E.2 Generalized Likelihood Ratio Statistic

Let us focus on the best arm identification setting for which $J = 1$. In the BAI setting, the stopping rule from Equation (1.9) features $\max_{a\leq K}\min_{b\in\{1,\dots,K\}\setminus\{a\}} Z_{a,b}(t)$ with $Z_{a,b}(t)$ defined in Equation (1.8). Interestingly, when $\hat{\mu}_a(t) \geq \hat{\mu}_b(t)$, a closed form expression for $Z_{a,b}(t)$ can be obtained

$$Z_{a,b}(t) = N_a(t)d(\hat{\mu}_a(t), \hat{\mu}_{a,b}(t)) + N_b(t)d(\hat{\mu}_b(t), \hat{\mu}_{a,b}(t))$$

where

$$\hat{\mu}_{a,b}(t) \coloneqq \frac{N_a(t)}{N_a(t)+N_b(t)}\hat{\mu}_a(t) + \frac{N_b(t)}{N_a(t)+N_b(t)}\hat{\mu}_b(t)$$

Let us denote $\hat{a}_t$ the arm with the highest empirical mean at time t.
We have $\min_{b\in\{1,\dots,K\}\setminus\{a\}} Z_{a,b}(t)$ is positive only if $\hat{\mu}_a(t) \geq \hat{\mu}_b(t)$ for all $b \neq a$. Which gives

$$Z'(t) \coloneqq \max_{a\leq K}\min_{b\in\{1,\dots,K\}\setminus\{a\}} Z_{a,b}(t) = \min_{b\in\{1,\dots,K\}\setminus\{\hat{a}_t\}} Z_{\hat{a}_t,b}(t)$$

Recalling the expression of d_{mid} from Equation (2.8)

$$
\begin{aligned}
Z'(t) &= \min_{b\in\{1,\dots,K\}\setminus\{\hat{a}_t\}} Z_{\hat{a}_t,b}(t) \\
&= t \min_{b\in\{1,\dots,K\}\setminus\{\hat{a}_t\}} d_{\mathrm{mid}}\left(\frac{N_{\hat{a}_t}(t)}{t}, \hat{\mu}_{\hat{a}_t}(t), \frac{N_b(t)}{t}, \hat{\mu}_b(t)\right) \\
&= t \min_{\boldsymbol{\lambda}\in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \sum_{a=1}^{K} \frac{N_a(t)}{t} d\left(\hat{\mu}_a(t), \lambda_a\right) \\
&= \min_{b\in\{1,\dots,K\}\setminus\{\hat{a}_t\}} \inf_{\boldsymbol{\lambda}\in\mathcal{L}:\lambda_{\hat{a}_t}=\lambda_b} \sum_{a\in\{\hat{a}_t,b\}} N_a(t) d\left(\hat{\mu}_a(t), \lambda_a\right)
\end{aligned}
$$

In the ABC-S problem, instead of the best empirical arm $\hat{a}_t$, the comparison is around the control arm. We recall the expression $Z(t)$ that is used.

$$
Z(t) \;=\; \min_{b\neq 0} \inf_{\boldsymbol{\lambda}\in\mathcal{L}:\lambda_0=\lambda_b} \sum_{a\in\{0,b\}} \sum_{i=1}^{J} N_{a,i}(t) d(\hat{\mu}_{a,i}(t), \lambda_{a,i}) \; .
$$

Comparing the two expression justifies why it is natural to consider $Z(t)$.

3 | On Limited-Memory Subsampling Strategies

In the previous chapter, we have discussed a pure exploration task where we were trying to identify the different arms better than a control. For the remaining part of the thesis, we switch to the regret minimization setting. In this chapter, we propose an alternative to traditional upper-confidence bounds based algorithms where subsampling is used for comparing the means of the different arms. In stationary environment, we design a policy that is asymptotically optimal without knowing the distribution of the arms. We also propose a variant of this algorithm that achieves optimal regret guarantees in any abruptly changing environment. The results from this chapter are based on [Baudry et al., 2021].

Outline

3.1 Introduction

3.1.1 Setting

In this chapter, we consider the K-armed stochastic bandit model that was presented in Section 1.2. We recall that the learner aims at maximizing his expected sum of rewards and needs to sequentially adapt his decision strategy in light of the information gained up to now. In this model, over-confident policies are provably suboptimal and a proper trade-off between exploitation and exploration has to be found. In its standard formulation the multi-armed bandit model postulates that the distributions of the rewards obtained when drawing the different arms remain constant over time. However, in some scenarios the stationary assumption is not realistic. In clinical trials, the disease to defeat may mutate and the initially optimal treatment could become suboptimal compared to another candidate [Gorre et al., 2001]. In strategic pricing problems, the price maximizing the profit of a given asset can evolve with the introduction of a new product on the market [Eliashberg and Jeuland, 1986]. For online recommendation systems, the preferences of the users are likely to evolve [Wu et al., 2018] and collected data becomes progressively obsolete.

During the past ten years, several works have considered non-stationary variants of the multi-armed bandit model, proposing methods that can be grouped into two main categories: they either actively try to detect modifications in the distribution of the arms with change-point detection algorithms [Liu et al., 2018, Cao et al., 2019, Auer et al., 2019, Chen et al., 2019, Besson et al., 2020] or they passively forget past information [Garivier and Moulines, 2011] but also [Raj and Kalyani, 2017, Trovo et al., 2020]. To some extent, all of these methods require some knowledge on the distribution to obtain theoretical guarantees.

To balance exploration and exploitation, the algorithms mentioned so far are based on one of the two standard building blocks introduced in the bandit literature: *Upper Confidence Bound* (UCB) constructions [Auer et al., 2002a] or *Thompson Sampling* (TS) [Thompson, 1933]. However, there has been a recent surge of interest for alternative non-parametric bandit strategies with for example [Kveton et al., 2019a] but also [Kveton et al., 2019b] and [Riou and Honda, 2020]. Instead of using prior information on the reward distributions as in Thompson sampling or of building tailored upper-confidence bounds [Cappé et al., 2013] those methods only use the empirical distribution of the data. These algorithms are non-parametric in the sense that the *exact same* implementation can be used with different probability distributions, while still achieving optimal regret guarantees. An interesting class of non-parametric algorithms are based on subsampling.

3.1.2 Subsampling Algorithms

A striking characteristic of the vast majority of UCB-based techniques is that except the exploration bonus only the sample means are used. Getting away from the computation of upper-confidence bounds seems necessary if we aim at building asymptotically optimal and non-parametric algorithms. Indeed, either the UCB are tailored to a specific distribution (i.e. a different implementation is required when using another distribution) and asymptotically optimality can be obtained [Cappé et al., 2013] or we approximate the distributions with a more general class of probability distributions (e.g. subgaussian for Bernoulli distributions) at the cost of loosing the optimality. The following question is central in this chapter.

Is it possible to maintain the asymptotic optimality while using the same algorithm's implementation for a broad class of probability distributions?

We answer positively to this question using a subsampling bandit algorithm. The idea of subsampling algorithms is to modify the way the empirical means are computed for ensuring enough exploration. When comparing two arms looking at their empirical means is too conservative and greedy policy are known to be suboptimal. One way for comparing two arms using their empirical means while ensuring enough exploration is to use subsampling. Assume arm 1 has been pulled n_1 times and arm 2 has been pulled n_2 times with $n_1 > n_2$. The idea of subsampling is to compute the empirical mean of arm 1 using only n_2 samples. Depending on how those n_2 samples will be collected gives rise to different algorithms that we discuss now.

From a high level perspective subsampling algorithms all rely on the same two components. **(1)** *subsampling*: the arms that have been pulled a lot are randomized by sampling only a fraction of their history. **(2)** *duels*: the arms are pulled based on the outcomes of duels between the different pairs of arms. Note that the term *duel*, used throughout the chapter, refers to the algorithmic principle of comparing the arms two by two, based on their subsamples. It is totally unrelated to the dueling bandit framework introduced by [Yue and Joachims, 2009].

The first subsampling algorithm called Best Empirical Sampled Average (BESA) was proposed in [Baransi et al., 2014]. When $K = 2$, the n_2 samples from arm 1 are sampled randomly without replacement from the history of arm 1. The arm 1 competes with its empirical mean based on a subsample of size n_2 while the arm 2 uses its entire history of size n_2. The arm with the largest empirical mean is declared winner and is pulled. By introducing, this subsampling layer, the comparison between the arms is fairer. When $K > 2$, [Baransi et al., 2014] propose a tournament structure with a divide-an-conquer like algorithm. The arms are separated in two different groups, we compute the winner of each group and at the next stage the winners from each group compete against each other. The winner of the tournament is then pulled.

Following BESA, [Chan, 2020] introduces SubSample Mean Comparison (SSMC). When $K > 2$, it departs from the tournament structure and introduces the notion of *leader*, that will be the arm competing with all the other *challengers*. We define this precisely in the chapter. Let us explain the duel between the arm 1 and the arm 2 assuming that 1 is the leader. With SSMC several subsamples are used for the leader. The algorithm looks at all possible blocks of n_2 consecutive samples from the arm 1. If there is a block for which the empirical mean based its samples is smaller than the empirical mean of the arm 2 (based on its entire history), then the arm 2 is the winner.

Building on the tools introduced in [Chan, 2020], [Baudry et al., 2020] introduced the Subsampling Duelling Algorithm (SDA) framework. They propose several sampling strategies and prove their asymptotical optimality when they are based on a randomized sampler and the distributions belong to the same one-parameter exponential family. They also present a deterministic sampling strategy Last Block sampling (without analyzing it) that will be the core of this chapter.

3.1.3 Scope and Contributions

In this chapter, we build on the Last-Block Subsampling Duelling Algorithm (LB-SDA) introduced by [Baudry et al., 2020] but for which no theoretical guarantees were provided. This approach is of interest because of its simplicity and its computational efficiency compared to other strategies based on randomized subsampling. We first prove that for stationary environments LB-SDA is asymptotically optimal in one-parameter exponential family models and therefore matches the guarantees obtained by [Baudry et al., 2020] for randomized subsampling schemes. The main technical challenge is to devise an alternative to the *diversity* condition used in their work, which was specifically designed for randomized subsampling schemes.

Furthermore, we show that, without additional changes, these guarantees still hold for a variant of the algorithm using a *limited memory* of the observations of each arm. We prove that storing $\Omega\left((\log T)^2\right)$ observations instead of T is sufficient to ensure the asymptotic guarantees, making the algorithm more tractable for larger time horizons. To the best of our knowledge, we are the first to propose an asymptotically optimal subsampling algorithm with polylogarithmic storage of rewards under general assumptions.

Building a subsampling algorithm based on the most recent observations makes it an ideal candidate for a passively forgetting policy. Our third contribution is to propose a natural extension of the LB-SDA strategy to non-stationary environments. By limiting the extent of the time window in which subsampling is allowed to occur, one obtains a passively forgetting non-parametric bandit algorithm, which we refer to as Sliding Window Last Block Subsampling Duelling Algorithm (SW-LB-SDA). To analyze the performance of this algorithm, we assume an abruptly changing environment in which the reward distributions change at unknown time instants called *breakpoints*. We show that SW-LB-SDA guarantees a regret of order $\mathcal{O}(\sqrt{\Gamma_T T \log(T)})$ for any abruptly changing environment with at most Γ_T breakpoints, thus matching the lower bound from [Garivier and Moulines, 2011], up to logarithmic factors. The only required assumption is that, during each stationary phase, the reward distributions belong to the same one-parameter exponential family for all arms. Due to its non-parametric nature, this algorithm can thus be used in many scenarios of interest beyond the standard bounded-rewards / change-in-the-mean framework. We discuss some of these scenarios in Section 3.5, where we validate numerically the potential of the approach by comparing it with a variety of state-of-the-art algorithms for non-stationary bandits.

Structure of the Chapter. The chapter is organized as follows. In Section 3.2, we introduce the mathematical model of the problem. In Section 3.3, we analyze LB-SDA in any stationary environment and propose Memory-Limited LB-SDA an adaptation of LB-SDA that enjoys a significant reduction of the storage cost. In Section 3.4, we add additional mechanisms for using LB-SDA in any abruptly changing environment and obtain order optimal regret bounds. Finally, in Section 3.5, extensive numerical simulations highlight the merits of this approach, particularly when the changes are not only affecting the means of the rewards. All the missing elements for the results presented in this chapter are reported in appendix.

3.2 Preliminaries

The algorithms to be presented below are designed for the stochastic K-armed bandit model that was presented in Section 1.2. We recall in this section the two variants that will be considered in this chapter: stationary and abruptly changing environments.

Stationary Environments. When the environment is stationary, the K arms are characterized by the reward distributions $(\nu_k)_{k\leq K}$ and their associated means $(\mu_k)_{k\leq K}$, we recall that $\mu^\star = \max_{k\in\{1,\dots,K\}} \mu_k$ denotes the highest expected reward. We denote by $(Y_{k,s})_{s\in\mathbb{N}}$ the i.i.d. sequence of rewards from arm k. Following [Chan, 2020], our algorithm operates in successive rounds, whose length varies between 1 and K time steps. At each round r, the *leader* denoted $\ell(r)$ is defined and $(K-1)$ duels with the remaining arms called *challengers* are performed. Denoting by $N_k(r)$ the number of pulls of arm k up to the round r the leader is the arm that has been pulled the most. Namely,

Definition 3.1. At round r the leader is the arm satisfying

$$\ell(r) = \text{argmax}_{k\in\{1,\dots,K\}} N_k(r) \; . \tag{3.1}$$

When several arms are candidate for the maximum number of pulls, the one with the largest sum of rewards is chosen. If this is still not sufficient to obtain a unique arm, the leader is chosen at random among the arms maximizing both criteria. At round r, a subset $\mathcal{A}_r \subset \{1,\dots,K\}$ is selected by the learner based on the outcomes of the duels against $\ell(r)$. Next, all arms in $\mathcal{A}_r$ are drawn, yielding $Y_{k,N_k(r)}$ for $k \in \mathcal{A}_r$, where $N_k(r) = \sum_{s=1}^{r} \mathbb{1}(k \in \mathcal{A}_s)$. In these stationary environments when considering the exponential family bandit model any asymptotically optimal strategy will satisfy the lower bound from Definition 1.7.

Abruptly changing environments. In Section 3.4, we consider abruptly changing environments. The number of breakpoints up to time T, denoted Γ_T, is defined by

$$\Gamma_T = \sum_{t=1}^{T-1} \mathbb{1}\{\exists k \in \{1,\dots,K\}, \nu_{k,t} \neq \nu_{k,t+1}\} \; .$$

The time instants $(t_1,\dots,t_{\Gamma_T})$ associated to these breakpoints define $\Gamma_T + 1$ stationary phases where the reward distributions are fixed. Note that in this model, the change do not need to affect all arms simultaneously. In such environments, letting $\mu_t^\star = \max_{k\in\{1,\dots,K\}} \mu_{k,t}$ denote the best arm at time t, the performance of a policy is measured through the *dynamic regret* defined as

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

The notion of leader in those non-stationary environments is slightly different and we explain how to extend it in Section 3.4. We recall that in the non-stationary case, the lower bound for the dynamic regret takes a different form: for any strategy, there exists an abruptly changing instance such that $\mathbb{E}[\mathcal{R}_T] = \Omega(\sqrt{T\Gamma_T})$ [Garivier and Moulines, 2011, Seznec et al., 2020]. Even if the non-stationarity could be characterized by the more general variation budget introduced by [Besbes et al., 2014] and defined in 1.3, we focus in this chapter on abruptly changing environments.

3.3 LB-SDA in Stationary Environments

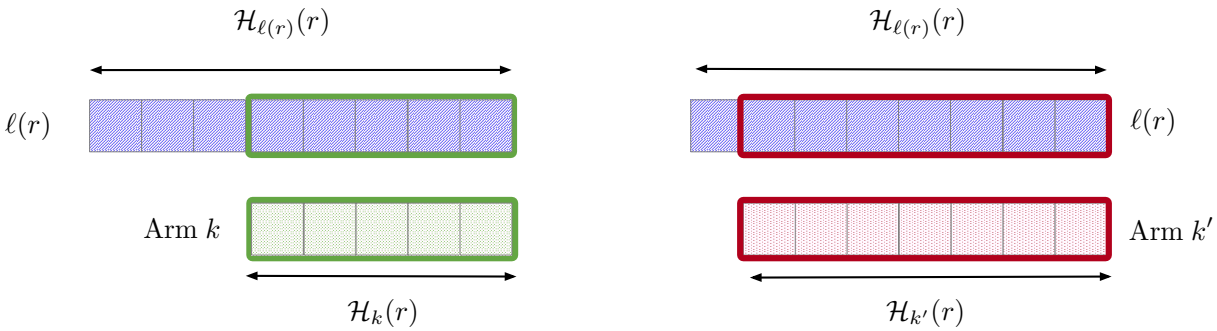
In this section we detail the subsampling strategy used in the LB-DSA algorithm and obtain asymptotically optimal regret guarantees for its performance. In Section 3.3.4, we consider the variant of LB-SDA in which the memory available to the algorithm is strongly limited.

3.3.1 Last Block Sampling

Compared to the algorithms analyzed in [Baudry et al., 2020] where the sampler is randomized, we consider a **deterministic sampler**. At round r, the duel between arm $k \neq \ell(r)$ and the leader consists in comparing the average reward from arm k with the average reward computed only from the last $N_k(r)$ observations of the leader. The challenger k thus wins its duel if

$$\bar{Y}_{k,N_k(r)} \geq \bar{Y}_{\ell(r),N_{\ell(r)}(r)-N_k(r)+1:N_{\ell(r)}(r)} \ , \tag{3.2}$$

where $\bar{Y}_{k,i:j} = \frac{1}{j-i+1}\sum_{n=i}^{j} Y_{k,n}$ denotes the average computed on the $j-i+1$ observations of arm k between its i-th and j-th pull, and $\bar{Y}_{k,n}$ is a shortcut for $\bar{Y}_{k,1:n}$. We denote $\mathcal{H}_k(r)$ the history available for arm k at round r. We illustrate the duel procedure on Figure 3.1.



Duel between k and $\ell(r)$ at round r | Duel between k' and $\ell(r)$ at round r

Figure 3.1: Illustration of the Last Block sampling at roud r for the duel between the arm k and the leader $\ell(r) \neq k$ and the duel between the leader and the arm k'. For the leader only the data in the green box (respectively red box) is kept for the duel against k (respectively k'), whereas arms k and k' use their entire history. Note that the number of datapoints kept by the leader in its duel with k is equal to $|\mathcal{H}_k(r)|$.

At each round, the set $\mathcal{A}_{r+1}$ includes all of the challengers that have defeated the leader, according to Equation (3.2), as well as under-explored arms for which $N_k(r) \leq \sqrt{\log(r)}$. If $\mathcal{A}_{r+1}$ is empty, only the leader is pulled. Combining these elements gives LB-SDA detailed below.

[Baransi et al., 2014] propose interesting arguments explaining why subsampling methods work. Essentially, if the sampler allows enough *diversity* in the duels, the probability of repeatedly selecting a suboptimal arm is small. On the sampler side, this condition is satisfied when out of a large number of duels between two arms there is a reasonable amount of them with non-overlapping subsamples. We prove that last block sampling satisfies such property. The second requirement concerns the distribution of the arms, and has been formulated by [Baransi et al., 2014] who introduced the *balance function* of a family of distributions. In particular, [Chan, 2020] shows that introducing an asymptotically negligible sampling obligation of $\sqrt{\log r}$ is enough to make subsampling suitable when the arms come from the same one-parameter exponential family of distributions. Namely, if each arm has at least $\sqrt{\log r}$ samples at round r, the *diversity* of duels

Input: K arms, horizon T
Initialization: $t \leftarrow 1$, $r \leftarrow 1$, $\forall k \in \{1, ..., K\}, N_k \leftarrow 0$
while $t < T$ **do**
  $\mathcal{A} \leftarrow \{\}$, $\ell \leftarrow \text{leader}(N, Y)$
  if $r = 1$ **then**
    $\mathcal{A} \leftarrow \{1, \dots, K\}$ (Draw each arm once)
  else
    for $k \neq \ell \in \{1, ..., K\}$ **do**
      if $N_k \leq \sqrt{\log(r)}$ or $\bar{Y}_{k,N_k} \geq \bar{Y}_{\ell,N_\ell - N_k + 1:N_\ell}$ **then**
        $\mathcal{A} \leftarrow \mathcal{A} \cup \{k\}$
      if $|\mathcal{A}| = 0$ **then**
        $\mathcal{A} \leftarrow \{\ell\}$
  for $k \in \mathcal{A}$ **do**
    Pull arm k, observe reward Y_{k,N_k+1}, $N_k \leftarrow N_k + 1$, $t \leftarrow t + 1$
  $r \leftarrow r + 1$

Algorithm 7: LB-SDA

will guarantee each arm to be pulled enough. This exploration rate does not have to be tuned and is not detrimental in practice : for an horizon of, say, $T = 10^6$ it only forces each arm to be sampled at least 4 times.

3.3.2 Regret Analysis of LB-SDA

We consider that the arms come from the same one-parameter exponential family of distributions as described in Equation (1.2) with $\mathcal{P} = \{(\nu_\theta)_\theta : d\nu_\theta/d\xi = \exp(\theta x - b(\theta))\}$. This assumption is standard in literature and covers a broad range of bandits applications. The exact knowledge of the family of distributions of the arms (e.g Bernoulli, Gaussian with known variance, Poisson, etc.) can be used to calibrate algorithms like Thompson Sampling [Kaufmann et al., 2012], KL-UCB [Cappé et al., 2013] or IMED [Honda and Takemura, 2015] in order to reach asymptotic optimality. Recently, subsampling algorithms like SSMC [Chan, 2020] and RB-SDA [Baudry et al., 2020] have been proved to be optimal *without* knowing exactly $\mathcal{P}$. This means that the same algorithm can run on Bernoulli or Gaussian distributions and achieve optimality. We first prove that LB-SDA matches these theoretical guarantees. As before, we denote $\text{kl}(\mu, \mu')$ the Kullback-Leibler divergence between two distributions of mean μ and μ' in the exponential family $\mathcal{P}$.

Theorem 3.2 (Asymptotic optimality of LB-SDA)**.** *Let $\mathcal{E}$ be the exponential family bandit model and $\nu = (\nu_1, \dots, \nu_K) \in \mathcal{E}$ with respective means $(\mu_1, \dots, \mu_K)$ where the K arms belong to the same one-parameter exponential family of distributions. The regret of LB-SDA satisfies, for all $\epsilon > 0$,*

$$\mathcal{R}_\nu(T) \leq \sum_{k:\mu_k < \mu^\star} \frac{(1+\epsilon)\Delta_k}{\text{kl}(\mu_k, \mu^\star)} \log(T) + C(\nu, \epsilon) \ ,$$

where $C(\nu, \epsilon)$ is a problem-dependent constant.

Proof. We assume without loss of generality that there is a unique optimal arm denoted $k^\star$. The analysis of [Chan, 2020] and [Baudry et al., 2020] shows that for any SDA algorithm the number of pulls of a suboptimal arm may be bounded as follow.

Lemma 3.3 (Lemma 4.1 in [Baudry et al., 2020])**.** *For any suboptimal arm $k \neq k^\star$, the expected number of pulls of k is upper bounded by*

$$\mathbb{E}[N_k(T)] \leq \frac{1+\epsilon}{\mathrm{kl}(\mu_k, \mu^\star)} \log(T) + C_k(\nu, \epsilon) + 32 \sum_{r=1}^{T} \mathbb{P}(N_{k^\star}(r) \leq (\log r)^2) \ ,$$

where $C_k(\nu, \epsilon)$ is a problem-dependent constant.

The next step consists in upper bounding the probability that the best arm is not pulled "enough" during a run of the algorithm. This part is more challenging and relies on the notion of *diversity* in the subsamples provided by the subsampling algorithm. This notion was introduced by [Baransi et al., 2014] to analyze the Best Empirical Sampled Average (BESA) algorithm. Intuitively, random block sampling [Baudry et al., 2020] or sampling without replacement [Baransi et al., 2014] explore different part of the history thus bringing diversity in the duels. Unfortunately, this property is not satisfied by deterministic samplers. Nonetheless, with a careful examination of the relation implied by the deterministic nature of last-block subsampling it is possible to prove that the number of pulls of the optimal arm is large enough with high probability.

Lemma 3.4. *The probability that the optimal arm is not pulled enough by LB-SDA can be upper bounded as follows*

$$\sum_{r=1}^{+\infty} \mathbb{P}\left(N_{k^\star}(r) \leq (\log r)^2\right) \leq C_{k^\star}(\nu) \ ,$$

for some constant $C_{k^\star}(\nu)$.

□

Plugging the result of Lemma 3.4 in Lemma 3.3 gives the asymptotic optimality of LB-SDA (Theorem 3.2). We prove Lemma 3.4 in the following section.

3.3.3 Proof of Lemma 3.4

We recall that $\bar{Y}_{k,i}$ denotes the mean of the i first rewards of arm k and that for a set $\mathcal{B}$, $\bar{Y}_{k,\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{s \in \mathcal{S}} Y_{k,s}$. When $|\mathcal{B}| = i$, $\bar{Y}_{k,\mathcal{B}}$ and $\bar{Y}_{k,i}$ have the same distribution. Before establishing our main result for LB-SDA, we introduce the balance function of an arm, which was first defined in [Baransi et al., 2014]. Assume that the K arms are characterized by the reward distributions $(\nu_1, \ldots, \nu_K)$. Assume that there is a unique optimal arm denoted $k^\star$.

Definition 3.5 (Balance function)**.** Let $\nu_{k,j}$ denote the distribution of the sum of j independent variables drawn from ν_k, and $F_{\nu_{k,j}}$ the corresponding CDF. The balance function of arm k is

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

If we draw one sample from a distribution $\nu_{k^\star,j}$ and M independent samples from another distribution $\nu_{k,j}$, the balance function $\alpha_k(M, j)$ quantifies the probability that each sample from

$\nu_{k,j}$ is larger than the sample from $\nu_{k^\star,j}$. The index j represents itself the fact that these variables are built as the sum of j independent random variables from the same distribution (respectively $\nu_{k^\star}$ and ν_k). This function has been studied in detail in [Baudry et al., 2020] (Appendix G and H), and we will use its properties to prove the following result.

Lemma 3.4. *The probability that the optimal arm is not pulled enough by LB-SDA can be upper bounded as follows*

$$\sum_{r=1}^{+\infty} \mathbb{P}\left(N_{k^\star}(r) \leq (\log r)^2\right) \leq C_{k^\star}(\nu) \ ,$$

for some constant $C_{k^\star}(\nu)$.

Proof. Let us fix a round $r \geq 1$ and assume $N_{k^\star}(r) \leq (\log r)^2$. The main problem with the last block sampling is that if both the leader and a given challenger are not played for some time, the index used in their duels remain the same due to the deterministic nature of the sampler. As a consequence this challenger is never played as long as the leader remains the same. If this situation occur too often, this would limit the diversity for the duels played by the optimal arm $k^\star$ against suboptimal leaders. We show that this is not possible by proving that the leader will be played a large number of times, which necessarily brings some diversity. To measure this, we define the quantity of duels won by the leader at the different rounds as

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

where we added 1 to consider the first round where every arm is pulled once. We recall the sampling obligation rule introduced in Section 3.3. and that we use to consider rounds where the optimal arm has enough samples. At any round r each arm with less than $f(r) = \sqrt{\log r}$ samples is pulled. We focus on rounds where we are sure that arm $k^\star$ has been pulled "enough", and compute the probability that it has lost a lot of duels after this moment. In particular, we consider a_r as the smallest round satisfying $f(a_r) \geq f(r) - 1$, ensuring $N_{k^\star}(a_r) \geq \lfloor f(r) - 1 \rfloor$. This round is exactly $\lceil f^{-1}(f(r) - 1) \rceil$, that can be computed as

$$\begin{aligned} f^{-1}(f(r) - 1) &= \exp\left((f(r) - 1)^2\right) = \exp\left(f(r)^2 + 1 - 2f(r)\right) \\ &= r \times \exp(-2f(r) + 1) \ . \end{aligned}$$

This means that for any $\gamma \in (0, 1)$, if r is large enough to satisfy $f(r) \geq \frac{1 - \log \gamma}{2}$ then $a_r \leq \gamma r$. For the rest of the proof we consider the number of duels lost by the arm $k^\star$ after the round a_r against unique subsamples of a suboptimal leader. The number of duels won by the leader between the rounds a_r and r is equal to $W_r - W_{a_r}$. Out of those duels, at most $(\log r)^2$ of them can concern the optimal arm $k^\star$ because $N_{k^\star}(r) \leq (\log r)^2$. Consequently, there is at least $W_r - W_{a_r} - (\log r)^2$ duels won by a suboptimal leader ($k \neq k^\star$) between rounds a_r and r. Using Lemma 3.6 stated below and $W_{a_r} \leq a_r$ one has,

$$\begin{aligned} W_r - W_{a_r} - (\log r)^2 &\geq \frac{r}{K} - a_r - (\log r)^2 \\ &\geq \frac{r}{K} - \gamma r - (\log r)^2 \ . \end{aligned}$$

Lemma 3.6. *With $W_r = 1 + \sum_{s=1}^{r-1} \mathbb{1}(\mathcal{A}_{s+1} = \{\ell(s)\})$, for any round r under LB-SDA it holds that*

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

To simplify the expression we just write that for any $\beta \in (0,1)$ there exists a constant $r(\beta, K)$ satisfying $\forall r \geq r(\beta, K)$,

$$W_r - W_{a_r} - (\log r)^2 \geq \beta \frac{r}{K} \ . \tag{3.3}$$

Under $N_{k^\star}(r) \leq (\log r)^2$, we show that there exists some $j \in \{1, ..., \lfloor (\log r)^2 \rfloor\}$ such that a fraction $1/(\log r)^2$ of the $\beta r/K$ duels counted above have been played with j samples for $k^\star$. Let us denote $\widetilde{W}_r \coloneqq W_r - W_{a_r} - (\log r)^2$ and show this by contradiction. Out of those duels, we denote $\widetilde{W}_{r,j}$ the indices of duels played when $k^\star$ uses j samples and $|\widetilde{W}_{r,j}|$ the number of duels in $\widetilde{W}_{r,j}$. Note that the rounds in $\widetilde{W}_{r,j}$ are necessarily consecutive because the number of samples in the entire history of $k^\star$ can only increase when performing a duel with a suboptimal leader. In particular, if there is a round s in $\widetilde{W}_{r,j}$ where $k^\star$ wins against a suboptimal leader then in the next round the history of $k^\star$ will be of length $j+1$. No round after s will belong to $\widetilde{W}_{r,j}$ because the following duels in $\widetilde{W}_r$ will contain at least $j+1$ samples for $k^\star$. If we assume that for all $j \leq \lfloor (\log r)^2 \rfloor$, there is strictly less than $\frac{\beta}{(\log r)^2} \frac{r}{K}$ duels played when $k^\star$ has j samples. The following would hold,

$$W_r - W_{a_r} - (\log r)^2 = \widetilde{W}_r = \sum_{j=1}^{\lfloor (\log r)^2 \rfloor} |\widetilde{W}_{r,j}| < \sum_{j=1}^{\lfloor (\log r)^2 \rfloor} \frac{\beta}{(\log r)^2} \frac{r}{K} < \beta \frac{r}{K} \ .$$

There is a contradiction with Equation (3.3). This means that there exists a value $j \leq \lfloor (\log r)^2 \rfloor$ and $\beta r/((\log r)^2 K)$ duels such that $k^\star$ competes using its same block of observations of size j.

Furthermore, using a similar reasoning we are sure that a fraction $1/(K-1)$ of these duels is played against the same leader $k \in \{2, \dots, K\}$. We would now like to obtain duels with non-overlapping blocks. Even if the blocks are all consecutive, waiting for j steps is enough to ensure that the blocks used by the leader are not overlapping. Taking a fraction $1/j$ of the duels from the previous subset is hence enough to guarantee this. We illustrate this on Figure 3.2

Combining all previous arguments, we can conclude that for any $\beta \in (0,1)$ there exists a constant $r(\beta, K)$ such that for any round $r > r(\beta, K)$, under the event $\{N_{k^\star}(r) \leq (\log r)^2\}$, there exists some $k \in \{2, \dots, K\}$ and some $j \in \{\lfloor \sqrt{\log r} - 1 \rfloor, \lfloor (\log r)^2 \rfloor\}$ such that arm $k^\star$ lost at least $M = \beta \frac{r}{K(K-1)(\log r)^2 j}$ duels against non-overlapping blocks of arm k. For all those duels, k is the leader and $k^\star$ use the same block of j observations.

We denote $\mathcal{B}_1, \dots, \mathcal{B}_M$ the M non-overlapping blocks of observations for arm k containing j observations each. We also denote $\mathcal{B}^\star$ the block containing the j observations from arm $k^\star$ that are used for all those duels. We have then established that for all $r \geq r(\beta, K)$,

$$\{N_{k^\star}(r) \leq (\log r)^2\} \subset \bigcup_{k \neq k^\star} \bigcup_{j=\lfloor f(r)-1 \rfloor}^{\lfloor (\log r)^2 \rfloor} \bigcap_{m=1}^{M} \{\bar{Y}_{k^\star, \mathcal{B}^\star} < \bar{Y}_{k, \mathcal{B}_m}\} \ .$$

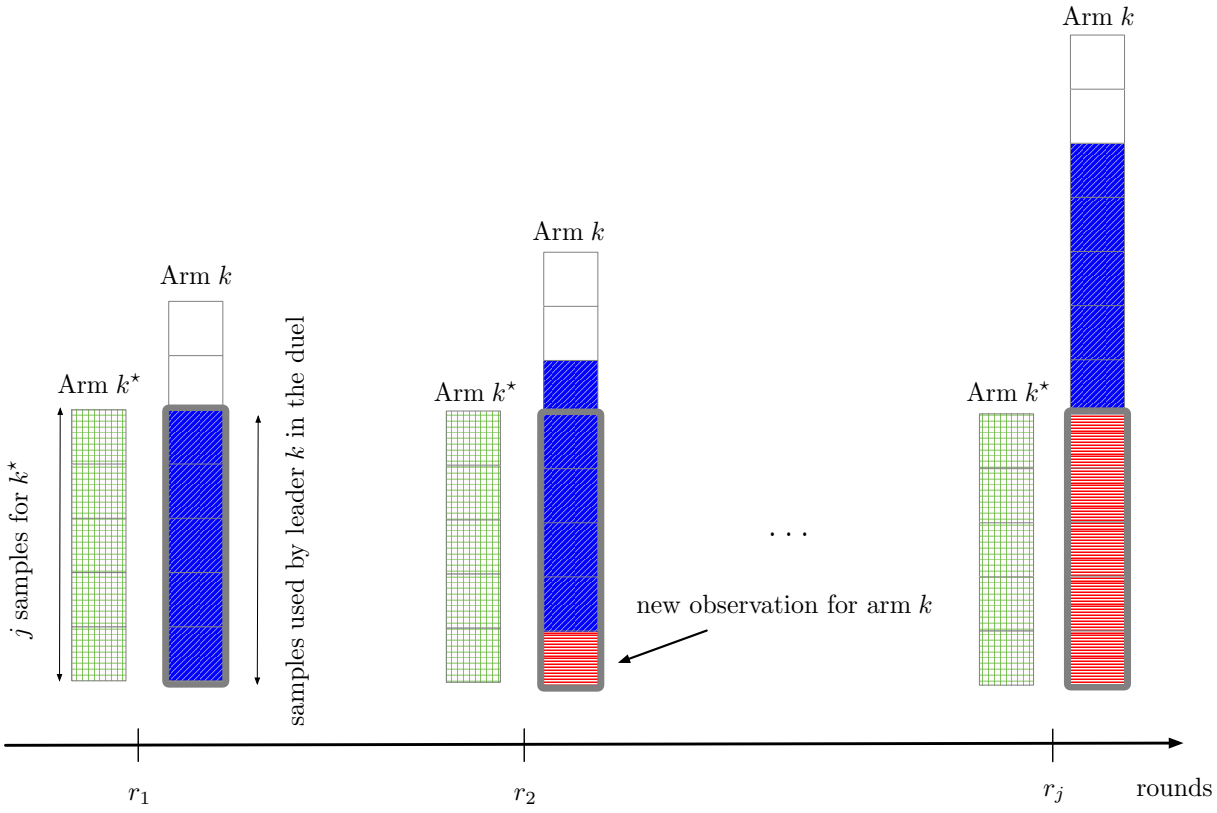


Figure 3.2: Let $(r_1, \dots, r_j)$ be j consecutive duels between the leader k and $k^\star$ where $k^\star$ keep the same j samples. Waiting j steps is enough for obtaining non-overlapping blocks (the red and the blue blocks here).

For this reason, using the notation from Definition 3.5 and introducing $Z^\star \sim \nu_{k^\star,j}$ and an i.i.d sequence $Z_1, \dots, Z_M \sim \nu_{k,j}$, when $r \geq r(\beta, K)$ one has

$$\mathbb{P}\left(N_{k^\star}(r) \leq (\log r)^2\right) \leq \sum_{k \neq k^\star} \sum_{j=\lfloor f(r)-1 \rfloor}^{\lfloor (\log r)^2 \rfloor} \mathbb{P}\left(Z^\star < \min_{i=\{1,\dots,M\}} Z_i\right) .$$

We can then rewrite $\mathbb{P}\left(Z^\star < \min_{i=\{1,\dots,M\}} Z_i\right)$ as follows

$$\begin{aligned}
\mathbb{P}\left(Z^\star < \min_{i=\{1,\dots,M\}} Z_i\right) &= \mathbb{E}_{\substack{Z^\star \sim \nu_{k^\star,j} \\ Z_1,\dots,Z_M \sim \nu_{k,j}}} \left[\prod_{i=1}^{M} \mathbb{1}(Z^\star < Z_i)\right] \\
&= \mathbb{E}_{Z^\star \sim \nu_{k^\star,j}} \left[\mathbb{E}_{Z_1,\dots,Z_M \sim \nu_{k,j}} \left[\prod_{i=1}^{M} \mathbb{1}(Z^\star < Z_i) \middle| Z^\star\right]\right] \\
&= \mathbb{E}_{Z^\star \sim \nu_{k^\star,j}} \left[(1 - F_{\nu_{k,j}}(Z^\star))^M\right] .
\end{aligned}$$

The term in the last equality corresponds exactly to the balance function $\alpha_k(M, j)$ from Definition 3.5, with $M = \beta \frac{r}{K(K-1)(\log r)^2 j}$. Hence, we have the following upper bound

$$\sum_{r=1}^{T} \mathbb{P}\left(N_{k^\star}(r) \leq (\log r)^2\right) \leq r(\beta, K) + \sum_{k \neq k^\star} \sum_{r=r(\beta,K)}^{T} \sum_{j=\lfloor \log(r)-1 \rfloor}^{\lfloor (\log r)^2 \rfloor} \alpha_k(M, j) .$$

Remark 3.1. *The fact that the duels concern non-overlapping blocks of arm k is necessary to obtain independent samples. It is also important that those duels are based on exactly j observations in order to introduce the balance function.*

We conclude the proof using the following lemma which is proved in appendix.

Lemma 3.7. *If the arms k and $k^\star$ come from the same one-parameter exponential family of distributions it holds that*

$$\sum_{r=r(\beta,K)}^{T} \sum_{j=\lfloor \log(r)-1 \rfloor}^{\lfloor (\log r)^2 \rfloor} \alpha_k \left(\beta \frac{r}{K(K-1)(\log r)^2 j}, j \right) = O(1) \, .$$

3.3.4 Memory-Limited LB-SDA

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

The fact that LB-SDA is asymptotically optimal means that, when T is large, the arm with the largest mean is most often the leader with all of its challengers having a number of pulls that is of order $O(\log T)$ only. With duels based on the last block, this would mean in particular that only the last $O(\log T)$ observations from the optimal arm should be stored and that previous observations will *never* be used again in practice. Based on this intuition, one might think that keeping only $\log(T)/(\mu^\star - \mu_k)^2$ observations is enough for LB-SDA. However, this could only be done with the knowledge of the gaps that are unknown. We propose instead to limit the storage memory of each arm at round r to a value of the form

$$m_r \coloneqq \max \left(M, \left\lceil C(\log r)^2 \right\rceil \right) \, ,$$

where $C > 0$ and $M \in \mathbb{N}$. M ensures that a minimum number of samples are stored during the first few rounds. Following the definition of [Agrawal and Goyal, 2012], we then define the set of *saturated arms* at a round r as

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

The only modification of LB-SDA is the following: at each round r, if a saturated arm is pulled then the newly collected observation replaces the oldest observation in its history.

3.3.4.1 The Algorithm

Before giving the algorithm, we introduce additional notations that are used in the statement of the algorithm. The stored history for the arm k at round r is denoted $\mathcal{H}_k(r)$. At round r when comparing the leader $\ell(r)$ and the arm $k \neq \ell(r)$ the last block of the history of $\ell(r)$ is used and is denoted $\mathcal{S}(\mathcal{H}_k(r), \mathcal{H}_\ell(r))$. In particular, when both arms are saturated their entire history

of length m_r is used for the duel. The Last Block Subsampling Duelling Algorithm with Limited Memory is reported in Algorithm 8

Input: K arms, horizon T, m_r storage limitation
Initialization: $t \leftarrow 1$, $r = 1$ $\forall k \in \{1, ..., K\}, N_k \leftarrow 0$, $\mathcal{H}_k = \{\}$
while $t < T$ **do**
  $\mathcal{A} \leftarrow \{\}$, $\ell \leftarrow \text{leader}(N, t)$
  if $r = 1$ **then**
    $\mathcal{A} \leftarrow \{1, \dots, K\}$ (Draw each arm once)
  else
    for $k \neq \ell \in \{1, ..., K\}$ **do**
      if $N_k \leq \sqrt{\log r}$ *or* $\bar{Y}_{k,\mathcal{H}_k} > \bar{Y}_{\ell,\mathcal{S}(\mathcal{H}_k,\mathcal{H}_\ell)}$ **then**
        $\mathcal{A} \leftarrow \mathcal{A} \cup \{k\}$
      if $|\mathcal{A}| = 0$ **then**
        $\mathcal{A} \leftarrow \{l\}$
  for $k \in \mathcal{A}$ **do**
    if $\text{card}(\mathcal{H}_k) \geq m_r$ **then**
      $\text{pop}(\mathcal{H}_k)$ // Removing the oldest observation
    Pull arm k, observe reward Y_{k,N_k+1}, $N_k \leftarrow N_k + 1$, $t \leftarrow t + 1$
    $\mathcal{H}_k = \mathcal{H}_k \cup \{Y_{k,N_k+1}\}$ // Append the new observation
  $r \leftarrow r + 1$

Algorithm 8: LB-SDA with Limited Memory

3.3.4.2 Theoretical Guarantees

The following result shows that LB-SDA with Limited Memory keeps the same asymptotical performance as LB-SDA under general assumptions on m_r.

Theorem 3.8 (Asymptotic optimality of LB-SDA with Limited Memory)**.** *Let $\mathcal{E}$ be the exponential family bandit model and $\nu = (\nu_1, \dots, \nu_K) \in \mathcal{E}$ with means $(\mu_1, \dots, \mu_K)$ where the K arms belong to the same one-parameter exponential family of distributions. If $m_r / \log(r) \to \infty$, the regret of LB-SDA with Limited Memory satisfies, for all $\epsilon > 0$,*

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

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

The proof of this theorem is reported in Appendix 3.B, which provides precise estimates of the dependence of $C'(\nu, \epsilon, \mathcal{M})$ with respect to the parameters, and in particular, with respect to the sequence $\mathcal{M}$. Note that LB-SDA-LM remains an anytime algorithm because the storage constraint does not depend on the time horizon T but only on the current round.

3.3.5 Storage and Computational Cost

To the best of our knowledge, LB-SDA-LM is the only subsampling bandit algorithm that does not require to store the full history of rewards. We report in Table 3.1 estimates of the computational cost of LB-SDA-LM and its competitors.

Table 3.1: Storage and computational cost at round T for existing subsampling algorithms.

Algorithm	Storage	Computational cost Best-Worst case
BESA [Baransi et al., 2014]	$O(T)$	$O((\log T)^2)$
SSMC [Chan, 2020]	$O(T)$	$O(1)$-$O(T)$
RB-SDA [Baudry et al., 2020]	$O(T)$	$O(\log T)$
LB-SDA **(this chapter)**	$O(T)$	$O(1)$-$O(\log T)$
LB-SDA-LM **(this chapter)**	$O((\log T)^2)$	$O(1)$-$O(\log T)$

The computational cost can be broken into two parts: (a) the subsampling cost and (b) the computation of the means of the samples. We assume that drawing a sample of size n without replacement has $O(n)$ cost and that computing the mean of this subsample costs another $O(n)$. Furthermore, at round T, each challenger to the best arm has about $O(\log T)$ samples. This gives an estimated cost of $O\left((\log T)^2\right)$ for BESA [Baransi et al., 2014]. For RB-SDA [Baudry et al., 2020] the estimated cost is $O(\log(T))$, because the sampling cost for random block sampling is $O(1)$ and only the sample mean has to be recomputed at each round.

For the three deterministic algorithms (namely SSMC [Chan, 2020], LB-SDA, LB-SDA-LM), when the leader arm wins all its duels, its sample mean can be updated sequentially at cost $O(1)$. This is the *best case* in terms of computational cost. However, when a challenger arm is pulled, SSMC requires a full screening of the leader's history, with $O(T)$ cost, while LB-SDA and LB-SDA-LM only need the computation of the mean of the last $O(\log T)$ samples from the leader.

3.4 LB-SDA in Non-Stationary Environments

In stationary environments, LB-SDA achieves optimal regret rates, even when its decisions are constrained to use at most $O((\log T)^2)$ observations. One might think that this argument itself is sufficient to address non-stationary scenarios as the duels are performed mostly using recent observations. However, the latter is only true for the best arm and in the case where an arm that has been bad for a long period of time suddenly becomes the best arm, adapting to the change would still be prohibitively slow. For this reason, LB-SDA has to be equipped with an additional mechanism to perform well in non-stationary environments.

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

We keep a *round-based* structure for the algorithm, where, at each round r, duels between arms are performed and the algorithm subsequently selects the subset of arms $\mathcal{A}_r$ that will be pulled. In contrast to Section 3.3.4, where a constraint on storage related to the number of pulls

was added, here, we use a sliding window of length τ to limit the historical data available to the algorithm to that of the last τ rounds.

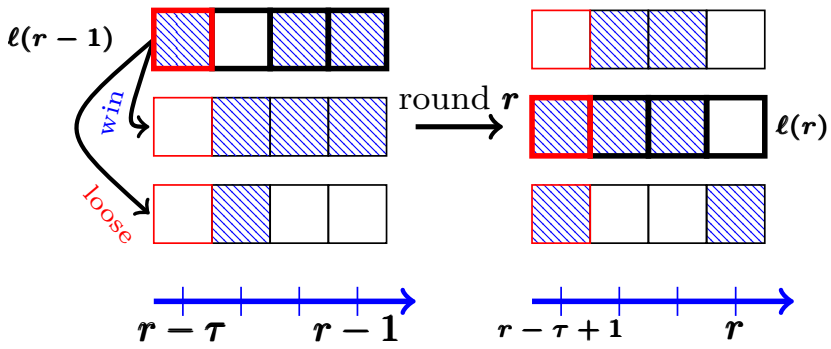


Figure 3.3: Illustration of a *passive leadership takeover* with a sliding window $\tau = 4$ when the standard definition of leader is used. The bold rectangle correspond to the leader. A blue square is added when an arm has an observation for the corresponding round and the red square correspond to the information that will be lost at the end of the round due to the sliding window.

Modified Leader Definition. The introduction of a sliding window requires a new definition for the *leader*. By analogy with the stationary case, the leader could be defined as the arm that has been pulled the most during the τ last rounds. However, with the inclusion of the sliding window, a new phenomenon, which we call *passive leadership takeover*, can occur. Let us define $N_k^\tau(r) = \sum_{s=r-\tau}^{r-1} \mathbb{1}(k \in \mathcal{A}_{s+1})$, the number of times arm k has been pulled during the last τ rounds and consider a situation with 3 arms $\{1, 2, 3\}$. Assume that the leader is arm 1 and at a round $(r-1)$ we have $N_1^\tau(r-1) = N_2^\tau(r-1)$. If the leader has been pulled τ rounds away and wins its duel against arm 2 but looses against arm 3, only arm 3 will be pulled at round r. Consequently, at round r, arm 2 will have a strictly larger number of pulls than arm 1 without having actually defeated the leader. This situation, illustrated on Figure 3.3, is not desirable as it can lead to spurious leadership changes. We fix this by imposing that any arm has to defeat the current leader to become the leader itself. Define,

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

Then for any $r \in \mathbb{N}$, the leader is defined as follows.

Definition 3.9 (Leader in non-stationary environments)**.** The leader at round $r+1$ is defined as

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

This modified definition of the leader ensures that an arm can become the leader only after earning at least τ/K samples and winning a duel against the current leader, or if the leader loses a lot of duels and its number of samples falls under a fixed threshold. Thanks to this definition it holds that $N_{\ell^\tau(r)}^\tau(r) \geq \min(r, \tau)/(2K)$. More details are given in Appendix 3.C.

Additional Diversity Flag. As in the vanilla LB-SDA, we use a sampling obligation to ensure that each arm has a minimal number of samples. However, in contrast to the stationary case, this very limited number of forced samples may not be sufficient to guarantee an adequate variety of duels due to the forgetting window. To this end, the sampling obligation is coupled with a *diversity flag*.

Definition 3.10 (Diversity flag)**.** The diversity flag for arm k at round r is a binary variable that is equal to 1 only when for the last $\lceil (K-1)(\log\tau)^2 \rceil$ rounds the three following conditions are satisfied:

1. some arm $k' \neq k$ has been leader during all these rounds
2. k' has not been pulled
3. k has not been pulled and satisfy $N_k^\tau(r) \leq (\log\tau)^2$.

In practice, there is a very low probability that these conditions are met simultaneously but this additional mechanism is required for the theoretical analysis. Note that the diversity flags have no impact on the computational cost of the algorithm as they require only to store the number of rounds since the last draw of the different arms (which can be updated recursively) as well as the last leader takeover. Arms that raise their diversity flag are automatically added to the set of pulled arms. Bringing these parts together, gives the pseudo-code of SW-LB-SDA in Algorithm 9.

3.4.2 Regret Analysis in Abruptly Changing Environments

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

Theorem 3.11 (Asymptotic optimality of SW-LB-SDA)**.** *If the time horizon T and number of breakpoint Γ_T are known, choosing $\tau = O(\sqrt{T\log(T)/\Gamma_T})$ ensures that the dynamic regret of SW-LB-SDA satisfies*

$$\mathcal{R}_\nu(T) = O\left(\sqrt{T\Gamma_T \log T}\right) .$$

To prove this result we only need to assume that, during each stationary period, the rewards come from the same one-parameter exponential family of distributions. In contrast, current state-of-the-art algorithms for non-stationary bandits typically require the assumption that the rewards are *bounded* to obtain similar guarantees. Hence, this result is of particular interest for tasks involving unbounded reward distributions that can be discrete (e.g Poisson) or continuous (e.g Gaussian, Exponential). SW-LB-SDA can also be used for general bounded rewards with the same performance guarantees by using the *binarization trick* [Agrawal and Goyal, 2013a]. Note however, that the knowledge of the horizon T and the estimated number of change point Γ_T is still required to obtain optimal rates, which is an interesting direction for future works on this approach [Auer et al., 2019, Besson et al., 2020]. We provide a high-level outline of the analysis behind Theorem 3.11 and the complete proof is given in Appendix 3.C.

Input: K arms, horizon T, τ length of sliding window
Initialization: $t \leftarrow 1$, $r \leftarrow 1$, $\forall k \in \{1, ..., K\}, N_k \leftarrow 0$, $N_k^\tau \leftarrow 0$
while $t < T$ **do**
  $\mathcal{A} \leftarrow \{\}$, $\ell \leftarrow \text{leader}(N, Y, \tau)$
  if $r = 1$ **then**
    $\mathcal{A} \leftarrow \{1, \dots, K\}$ (Draw each arm once)
  else
    for $k \neq \ell \in \{1, ..., K\}$ **do**
      if $N_k^\tau \leq \sqrt{\log(\tau)}$ *or* $D_k^\tau(r) = 1$ **then**
        $\mathcal{A} \leftarrow \mathcal{A} \cup \{k\}$
      else
        $\widehat{\mu}_k^\tau = \bar{Y}_{k, N_k - N_k^\tau + 1 : N_k}$
        $N = \min(N_k^\tau, N_\ell^\tau)$
        $\widehat{\mu}_{\ell,k}^\tau = \bar{Y}_{N_\ell - N + 1 : N_\ell}$
        if $\widehat{\mu}_k^\tau \geq \widehat{\mu}_{\ell,k}^\tau$ **then**
          $\mathcal{A} \leftarrow \mathcal{A} \cup \{k\}$
    if $|\mathcal{A}| = 0$ **then**
      $\mathcal{A} \leftarrow \{\ell\}$
  for $k \in \mathcal{A}$ **do**
    Pull arm k, observe reward $Y_{k, N_k + 1}$
    Update $N_k \leftarrow N_k + 1$, $N_k^\tau \leftarrow N_k^\tau + 1$, $t \leftarrow t + 1$
  for $k \in \{1, ..., K\}$ **do**
    if $k \in \mathcal{A}_{r - \tau + 1}$ **then**
      $N_k^\tau \leftarrow N_k^\tau - 1$
  $r \leftarrow r + 1$

Algorithm 9: SW-LB-SDA

Regret decomposition For the $\Gamma_T + 1$ stationary phases $[t_\phi, t_{\phi+1} - 1]$ with $\phi \in \{1, \dots, \Gamma_T\}$, we define r_ϕ as the first round where an observation from the phase ϕ was pulled. We further define $N_k^\phi = \sum_{r=r_\phi - 1}^{r_{\phi+1}-2} \mathbb{1}(k \in \mathcal{A}_{r+1})$ the number of pulls of an arm k during a phase ϕ. Introducing the gaps $\Delta_k^\phi = \mu_{t_\phi}^* - \mu_{t_\phi, k}$ and denoting the optimal arm $k_\phi^\star$, we can rewrite the regret as

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

Note that the quantities t_ϕ, r_ϕ and Δ_k^ϕ for the different stationary phases ϕ are only required for the theoretical analysis and the algorithm has no access to those values.

Remark 3.2. *We highlight that the sequence $(r_\phi)_{\phi \geq 1}$ is a random variable that depends on the trajectory of the algorithm. However, we show in Appendix 3.C that this causes no additional difficulty for upper bounding the regret.*

We introduce $\delta_\phi = t_{\phi+1} - t_\phi$ the length of a phase ϕ and give a sketch of proof for Theorem 3.11.

Sketch of Proof. Combining elements from the proofs of [Garivier and Moulines, 2011] and that of Theorem 3.2, we first provide an upper bound on $\mathbb{E}[N_k^\phi]$ for any suboptimal arm k during the phase ϕ as

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

We define $A_k^{\phi,\tau} = b_k^\phi \log(\tau)$ for some constant $b_k^\phi > 0$, along with the terms $c_{k,1}^{\phi,\tau}$, $c_{k,2}^{\phi,\tau}$ and $c_{k,3}^{\phi,\tau}$, which all represents a different technical aspect of the regret decomposition of SW-LB-SDA. Before interpreting them we start with their formal definition.

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

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

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

Bounding Individual Terms. The three terms have intuitive interpretation and summarize well the technical contributions behind Theorem 3.11. To some extent they all rely on the notion of *saturated* arms defined in Section 3.3.4 and that we refine in Appendix 3.C for the problems considered in this section (mainly by properly tuning $A_k^{\phi,\tau}$ in the theoretical analysis).

First, $\mathbb{E}[c_{k,1}^{\phi,\tau}]$ is an upper bound on the expectation of the number of times a *saturated suboptimal arm* can defeat the *optimal* leader (i.e $\ell^\tau(r) = k_\phi^*$). To prove this result we establish a new concentration inequality for Last-Block Sampling in the context of SW-LB-SDA.

The second term $\mathbb{E}[c_{k,2}^{\phi,\tau}]$ controls the probability that the *diversity flag* is activated when the optimal arm k_ϕ^* is the leader. We prove that if this event happen, then k_ϕ^* has necessarily lost at least one duel against a saturated *sub-optimal* arm, and that this event has only a low probability.

The term $\mathbb{E}[c_{k,3}^{\phi,\tau}]$ is the most difficult to handle, the main challenge is to upper bound the probability that the *optimal arm* is *not saturated* after a large number of rounds. We provide the complete analysis of each of these terms and a full description of all the technical results that led to Theorem 3.11 in Appendix 3.C .

3.5 Experiments

3.5.1 Limiting the Storage in Stationary Environments.

In our first experiment[1] reported on Figure 3.4, we compare LB-SDA and LB-SDA-LM on a stationary instance with $K = 2$ arms with Bernoulli distributions for a horizon $T = 10000$. We add natural competitors (Thompson Sampling [Thompson, 1933], kl-UCB [Cappé et al., 2013]), that know ahead of the experiment that the reward distributions are Bernoulli and are tuned

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

accordingly. The arms satisfy $(\mu_1, \mu_2) = (0.05, 0.15)$ with a gap $\Delta = 0.1$. We run LB-SDA-LM with a memory limit $m_r = \log(r)^2 + 50$, which gives a storage ranging from 50 to 150 samples (much smaller than the horizon $T = 10000$). The regret is averaged on 2000 independent replications and the upper and lower quartiles are reported. In this setup LB-SDA-LM performs similarly to KL-UCB, and the impact of limiting the memory is mild, when compared to LB-SDA. This illustrates that even with relatively small gaps (here 0.1), a substantial reduction of the storage can be done with only minor loss of performance with LB-SDA-LM.

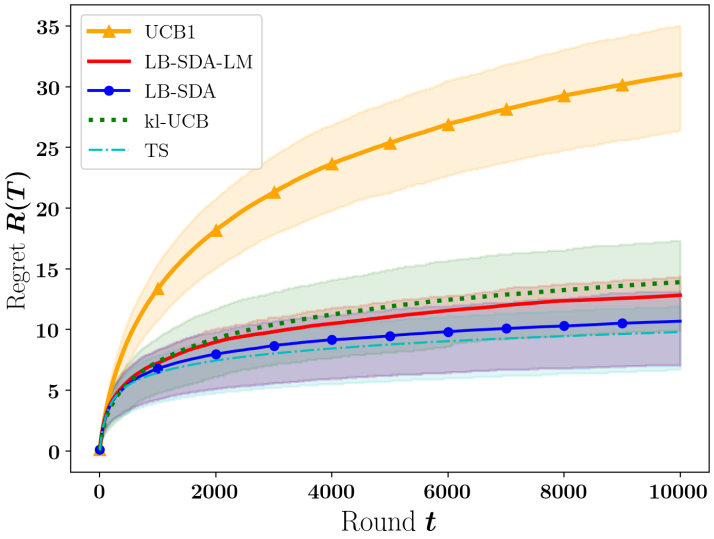


Figure 3.4: Cost of storage limitation on a Bernoulli instance. The reported regret are averaged over 2000 independent replications.

3.5.2 Empirical Performance in Abruptly Changing Environments.

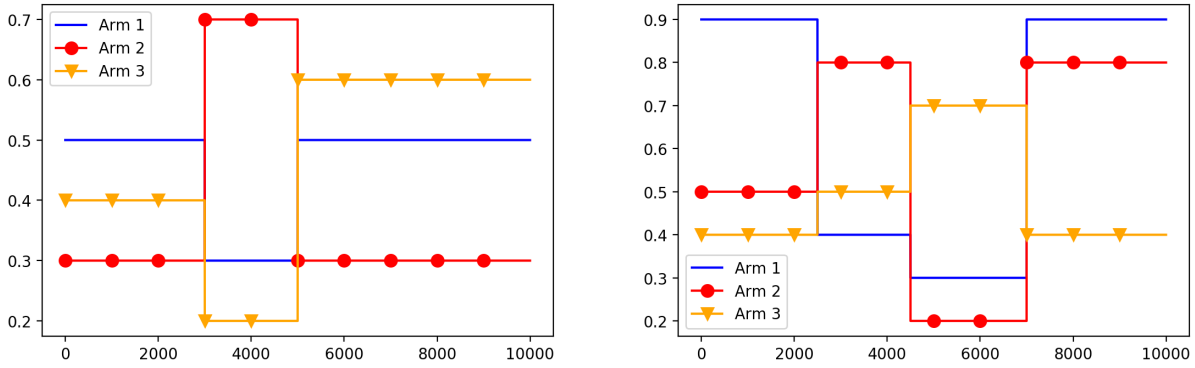


Figure 3.5: Evolution of the means: Left, Bernoulli arms (used in Fig. 3.6); Right, Gaussian arms (used in Figs. 3.7 and 3.8).

In the second experiment, we compare different state-of-the-art algorithms on a problem with $K = 3$ Bernoulli-distributed arms. The means of the distributions are represented on the left hand side of Figure 3.5 and the performance averaged on 2000 independent replications are reported on Figure 3.6. Two changepoint detection algorithms, CUSUM [Liu et al., 2018] and

M-UCB [Cao et al., 2019] are compared with progressively forgetting policies based on upper confidence bound, SW-klUCB and D-klUCB adapted from [Garivier and Moulines, 2011], or Thompson sampling, DTS [Raj and Kalyani, 2017] and SW-TS [Trovo et al., 2020]. We also add EXP3S [Auer et al., 2002a] designed for adversarial bandits and our SW-LB-SDA algorithm for the comparison. The different algorithms make use of the knowledge of T and Γ_T.

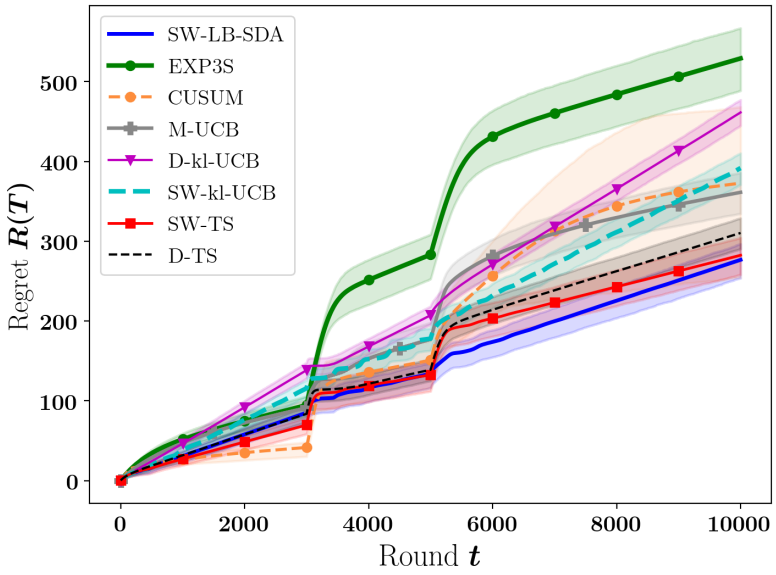


Figure 3.6: Performance on a Bernoulli instance averaged on 2000 independent replications.

To allow for fair comparison, we use for SW-LB-SDA, the same value of $\tau = 2\sqrt{T\log(T)/\Gamma_T}$ that is recommended for SW-UCB [Garivier and Moulines, 2011]. D-UCB uses the discount factor suggested by [Garivier and Moulines, 2011], $1/(1-\gamma) = 4\sqrt{T/\Gamma_T}$. The changepoint detection algorithms need extra information such has the minimal gap for a breakpoint and the minimum length of a stationary phase. For M-UCB, we set $w = 800$ and $b = \sqrt{w/2\log(2KT^2)}$ as recommended by [Cao et al., 2019] but set the amount of exploration to $\gamma = \sqrt{K\Gamma_T\log(T)/T}$ following [Besson et al., 2020]. In practice, using this value rather than the theoretical suggestion from [Cao et al., 2019] improved significantly the empirical performance of M-UCB for the horizon considered here. For CUSUM, α and h are tuned using suggestions from [Liu et al., 2018], namely $\alpha = \sqrt{\Gamma_T/T\log(T/\Gamma_T)}$ and $h = \log(T/\Gamma_T)$. On this specific instance, using $\epsilon = 0.05$ (to satisfy Assumption 2 of [Liu et al., 2018]) and $M = 50$ gives good performance. For the EXP3S algorithm, following [Auer et al., 2002a] the parameters α and γ are tuned as follows: $\alpha = 1/T$ and $\gamma = \min(1, \sqrt{K(e+\Gamma_T\log(KT)/((e-1)T)})$.

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

the other algorithms in this abruptly changing stochastic environment.

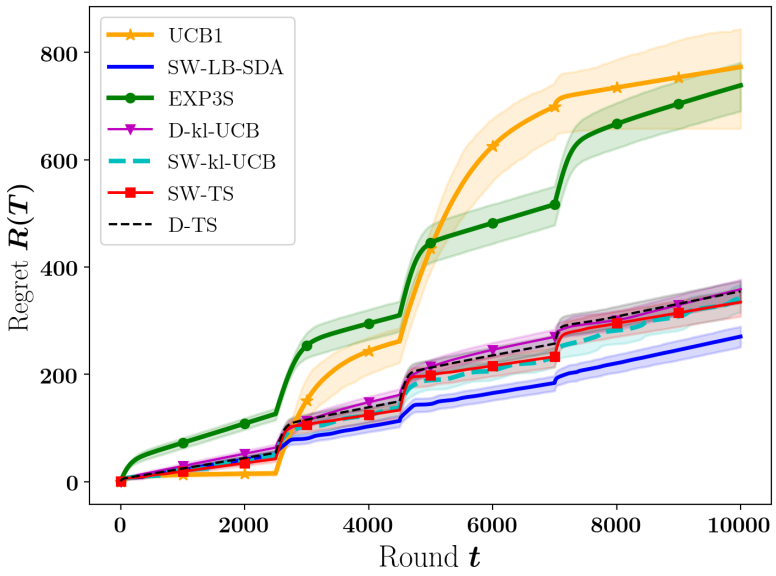


Figure 3.7: Performance on a Gaussian instance with a constant standard deviation of $\sigma = 0.5$ averaged on 2000 independent runs.

In the third experiment with $\Gamma_T = 3$ breakpoints, the $K = 3$ arms comes from Gaussian distributions with a fixed standard deviation of $\sigma = 0.5$ but time dependent means. The evolution of the arm's means is pictured on the right of Figure 3.5 and Figure 3.7 displays the performance of the algorithms.

CUSUM and M-UCB can not be applied in this setting because CUSUM is only analyzed for Bernoulli distributions and M-UCB assume that the distributions are bounded. Even if no theoretical guarantees exist for Thompson sampling with a sliding window or discount factors, when the distribution are Gaussian with known variance, we add them as competitors.

The analysis of SW-UCB and D-UCB was done under the bounded reward assumption but the algorithms can be adapted to the Gaussian case. Yet, the tuning of the discount factor and the sliding window had to be adapted to obtain reasonable performance, using $\tau = 2(1 + 2\sigma)\sqrt{T \log(T)/\Gamma_T}$ for D-UCB and $\gamma = 1 - 1/(4(1 + 2\sigma))\sqrt{\Gamma_T/T}$ for SW-UCB (considering that, practically, most of the rewards lie under $1 + 2\sigma$).

For reference, Figure 3.7 also displays the performance of the UCB1 algorithm that ignores the non-stationary structure. Clearly, SW-LB-SDA, in addition of being the only algorithm analyzed in this setting with unbounded rewards, also has the best empirical performance.

3.5.3 Non-stationarity Affecting the Variance

The last experiment features the same Gaussian means but with different standard errors. The standard error takes the values $0.5, 0.25, 1$ and 0.25, respectively, in the four stationary phases. The algorithms based on upper confidence bound are given the maximum standard error $\sigma = 1$, whereas SW-LB-SDA is not provided with any information of this sort. Figure 3.8 shows that the non-parametric nature of SW-LB-SDA is effective, with a significant improvement over state-of-the-art methods in such settings.

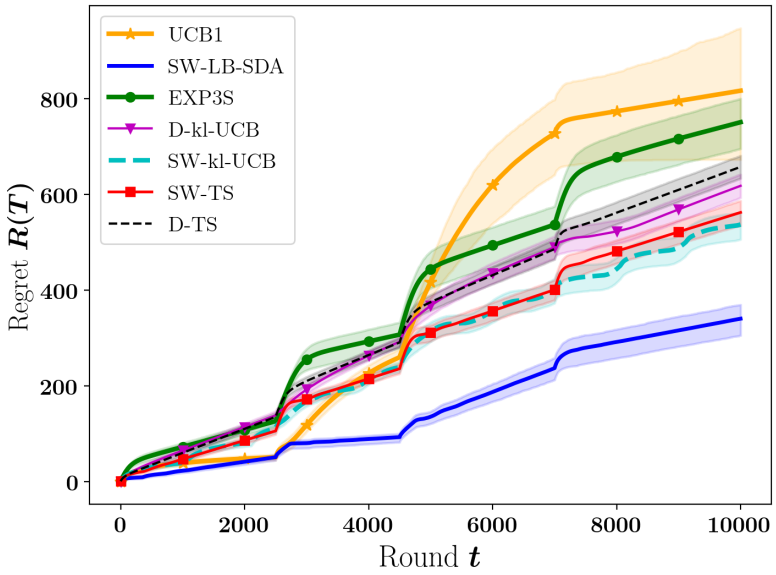


Figure 3.8: Performance on a Gaussian instance with time dependent standard deviations averaged on 2000 independent replications.

3.6 Conclusion

In this chapter, we considered the stochastic multi-armed bandit framework. We designed a non-parametric and asymptotically optimal policy for the exponential family bandit model. We also presented ideas for significantly reducing the storage required for performing the subsampling while preserving theoretical guarantees. We extended those ideas in abruptly changing environments where when given an upper-bound on the number of breakpoints, we propose an optimal algorithm. Finally, we assessed the performance of the approach empirically and obtained a strong competitor for this setting that can be analyzed with fewer restrictions on the distribution of the arms.

Appendix

Organization of the appendix

The appendix is organized as follows:

- In Section 3.A we provide some details on our analysis for the vanilla LB-SDA algorithm.
- In Section 3.B explain how to adapt LB-SDA when a limited memory is used and derive an upper-bound for the regret of this variant of LB-SDA.
- In Section 3.C a detailed analysis of LB-SDA with a sliding window in any abruptly changing environment is proposed.

Appendix 3.A Auxiliary Results for LB-SDA

Lemma 3.6. *With $W_r = 1 + \sum_{s=1}^{r-1} \mathbb{1}(\mathcal{A}_{s+1} = \{\ell(s)\})$, for any round r under LB-SDA it holds that*

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

Proof. We consider any trajectory of the bandit algorithm. For this trajectory we consider the sequence of the rounds where a change of leader occurred and write them as the (potentially infinite) set $\mathcal{Y} = [r_0, r_1, r_2, \dots]$. These are basically all the rounds r satisfying $\ell(r) \neq \ell(r-1)$. $r_0 = 1$ as it is the first round where we start defining the leader in the algorithm, and it holds that $N_{\ell(1)}(1) = 1$ as every arm is drawn once at the first round. As the leader was not defined before it holds that $W_1 = 1 = N_{\ell(1)}(1)$ so the property holds in r_0. As a first step, we show that the property is valid for all r_i when $i \in \mathbb{N}$. Let $i \in \mathbb{N}$, we assume that the property holds in r_i and we consider the round r_{i+1}. It holds that

$$W_{r_{i+1}} = W_{r_i} + \sum_{s=r_i}^{r_{i+1}-1} \mathbb{1}(\mathcal{A}_{s+1} = \ell(s)) \ .$$

The sum is exactly the number of duels won by the arm that is leader during the interval $[r_i, r_{i+1} - 1]$ and it holds that $\sum_{s=r_i}^{r_{i+1}-1} \mathbb{1}(\mathcal{A}_{s+1} = \ell(s)) = N_{\ell(r_i)}(r_{i+1}) - N_{\ell(r_i)}(r_i)$. Furthermore, when a change of leader happens the number of elements of the new and former leader are the same, i.e. $N_{\ell(r_{i+1})}(r_{i+1}) = N_{\ell(r_i)}(r_{i+1})$. This is due to the fact that when a challenger reaches the history size of the leader then the arm with the largest mean is chosen as the leader. In particular, if the challenger has a lower index than the leader at this round it cannot take the leadership at the next round as it will otherwise lose its duel against the leader. For this reason, the only possibility for a challenger to take the leadership is to reach to number of samples of the leader and to have a better index at this moment. We can write

$$\begin{aligned} W_{r_{i+1}} &= W_{r_i} + \sum_{s=r_i}^{r_{i+1}-1} \mathbb{1}(\mathcal{A}_{s+1} = \{\ell(s)\}) \\ &= W_{r_i} + N_{\ell(r_i)}(r_{i+1}) - N_{\ell(r_i)}(r_i) \\ &= W_{r_i} + N_{\ell(r_{i+1})}(r_{i+1}) - N_{\ell(r_i)}(r_i) \\ &= N_{\ell(r_i)}(r_i) + N_{\ell(r_{i+1})}(r_{i+1}) - N_{\ell(r_i)}(r_i) \quad \text{(Inductive step)} \\ &= N_{\ell(r_{i+1})}(r_{i+1}) \,. \end{aligned}$$

Therefore, if the property holds in r_i then it holds in r_{i+1} which gives the result. The extension to any round is obtained with similar arguments: $\forall r \notin \mathcal{Y}$, $\exists i : r_i < r < r_{i+1}$. Then we write

$$\begin{aligned} W_r &= W_{r_i} + \sum_{s=r_i}^{r-1} \mathbb{1}(\mathcal{A}_{s+1} = \ell(s)) \\ &= N_{\ell(r_i)}(r_i) + (N_{\ell(r_i)}(r) - N_{\ell(r_i)}(r_i)) \\ &= N_{\ell(r_i)}(r) = N_{\ell(r)}(r) \,, \end{aligned}$$

where the last inequality comes from the fact that the leader is unchanged between the rounds r_i and r. We conclude the proof by using the property that as the leader always has a number of samples larger than r/K, as it is the arm with the largest number of pulls at each round. □

Before proving Lemma 3.7 we prove an intermediary result that will also be useful to handle the balance function in the proof for switching bandits in Appendix 3.C. This result was already presented in [Chan, 2020], but we provide its proof for completeness.

Lemma 3.12. *Let F_1 and F_2 be the cumulative distribution function of two distributions with respective means μ_1 and μ_2, $\mu_1 > \mu_2$. For any integer $j \geq 1$ we denote $F_{1,j}$ and $F_{2,j}$ the cumulative distribution function of the sum of j independent random variables drawn respectively from F_1 and F_2, and*

$$\alpha(M, j) \coloneqq \mathbb{E}_{X \sim F_{1,j}} \left[(1 - F_{2,j}(X))^M \right]$$

the balance function of these two distributions. For any $u \in \mathbb{R}$ it holds that

$$\alpha(M, j) \leq F_{1,j}(u) + (1 - F_{2,j}(u))^M \,.$$

Furthermore, if we assume that F_1 and F_2 come from the same one-parameter exponential family of distributions, for any $u \in [0, 1]$ satisfying $F_2(u) \leq F_2(\mu_2)$ the following result holds

$$\alpha(M, j) \leq e^{-j\mathrm{kl}(\theta_2, \theta_1)} u + (1 - u)^M \,,$$

where $\mathrm{kl}(\theta_2, \theta_1)$ is the Kullback-Leibler divergence between F_2 and F_1, expressed with their canonical parameters θ_1 and θ_2.

Proof. We prove the first result, that is valid for any distribution F_1 and F_2 and is a direct

property of the definition of the balance function. For $u \in \mathbb{R}$, it holds that

$$\begin{aligned}\alpha(M,j) &= \int_{-\infty}^{+\infty} (1-F_{2,j}(x))^M dF_{1,j}(x) \\ &\leq \int_{-\infty}^{u} (1-F_{2,j}(x))^M dF_{1,j}(x) + \int_{u}^{+\infty} (1-F_{2,j}(x))^M dF_{1,j}(x) \\ &\leq F_{1,j}(u) + (1-F_{2,j}(u))^M \ .\end{aligned}$$

We now assume that F_1 and F_2 come from the same one-parameter exponential family of distributions. In this case they admit a density $f_\theta(y) = f(y,0)e^{\eta(\theta)y - \psi(\theta)}$ for some natural parameter $\theta \in \mathbb{R}$. We write θ_1 the parameter of F_1, and θ_2 the parameter of F_2. We then define some $y_1, ..., y_j \in \mathbb{R}^j$. If the sequence $y_1, \ldots, y_j$ satisfies $\sum_{u=1}^{j} y_u \leq j\mu_2$, it holds that

$$\prod_{u=1}^{j} f_{\theta_1}(y_u) = \prod_{u=1}^{j} e^{(\eta(\theta_1)-\eta(\theta_2))y_u - (\psi(\theta_1)-\psi(\theta_2))} f_{\theta_2}(y_u) \leq e^{-j\mathrm{kl}(\theta_2,\theta_1)} \prod_{u=1}^{j} f_{\theta_2}(y_u) \ .$$

where we write $\mathrm{kl}(\theta_2, \theta_1)$ for the Kullback-Leibler divergence between F_1 and F_2. This inequality first ensures that for all $x \leq \mu_2$

$$F_{1,j}(x) \leq e^{-j\mathrm{kl}(\theta_2,\theta_1)} F_{2,j}(x) \ .$$

If we insert this expression in the first result, we have that for any $u \in [0,1]$ satisfying $F_2(u) \leq F_2(\mu_2)$ the following result holds

$$\alpha(M,j) \leq e^{-j\mathrm{kl}(\theta_2,\theta_1)} u + (1-u)^M \ .$$

□

Remark 3.3. *The second result is particularly interesting because there is a trade-off in the choice of u. If we want to upper bound $\alpha(M,j)$ by a relatively small quantity we need to choose small values for u, however if u is too small then the second term may become too large. In particular, making the approximation $(1-u)^M \approx e^{-Mu}$ provides an optimal scaling of u of the form*

$$u^* = \frac{j\mathrm{kl}(\theta_2,\theta_1) + \log M}{M} \ ,$$

and as a consequence

$$\begin{aligned}\alpha(M,j) &\leq e^{-j\mathrm{kl}(\theta_2,\theta_1)} u^* + (1-u^*)^M \\ &\leq \frac{j\mathrm{kl}(\theta_2,\theta_1) + \log M}{M} e^{-j\mathrm{kl}(\theta_2,\theta_1)} + e^{M\log\left(1 - \frac{j\mathrm{kl}(\theta_2,\theta_1)+\log M}{M}\right)} \\ &\leq \frac{j\mathrm{kl}(\theta_2,\theta_1) + \log M}{M} e^{-j\mathrm{kl}(\theta_2,\theta_1)} + C_1 \frac{e^{-j\mathrm{kl}(\theta_2,\theta_1)}}{M} \\ &= \frac{j\mathrm{kl}(\theta_2,\theta_1) + \log M + C_1}{M} e^{-j\mathrm{kl}(\theta_2,\theta_1)} \ ,\end{aligned}$$

for some constant C_1.

With these technical results we can now prove Lemma 3.7 by simply replacing M by its value in the double sum.

Lemma 3.7. *If the arms k and $k^\star$ come from the same one-parameter exponential family of distributions it holds that*

$$\sum_{r=r(\beta,K)}^{T} \sum_{j=\lfloor \log(r)-1 \rfloor}^{\lfloor (\log r)^2 \rfloor} \alpha_k\left(\beta \frac{r}{K(K-1)(\log r)^2 j}, j\right) = O(1)\,.$$

Proof. We denote α_k the balance function between the arm $k^\star$ and an arm k and want to upper bound

$$\sum_{r=r(\beta,K)}^{T} \sum_{j=\lfloor \sqrt{\log r}-1 \rfloor}^{\lfloor (\log r)^2 \rfloor} \alpha_k\left(\beta \frac{r}{K(K-1)(\log r)^2 j}, j\right)\,.$$

We directly use the second result of Lemma 3.12, and choose the tuning of u from Remark 3.3. If we write $a_{r,j} = \alpha_k\left(\beta \frac{r}{K(K-1)(\log r)^2 j}, j\right)$ and try to extract the order of $a_{r,j}$ just in terms of r and j we obtain

$$a_{k,j} = O_{r,j}\left(\frac{j^2(\log r)^2}{r} e^{-j\mathrm{kl}(\theta_k,\theta_{k^\star})}\right)\,.$$

We then upper bound the term in j^2 by another $(\log r)^4$ using the upper limit on the sum on j, hence the only term left in j is $e^{-j\mathrm{kl}(\theta_2,\theta_1)}$, which sums in a term of order $\exp(-\sqrt{\log r})$. So we then obtain a term of the form

$$\sum_{r=r(\beta,K)}^{T} \sum_{j=\lfloor \sqrt{\log r}-1 \rfloor}^{\lfloor (\log r)^2 \rfloor} \alpha_k\left(\beta \frac{r}{K(K-1)(\log r)^2 j}, j\right) = O\left(\sum_{r=1}^{T} \frac{(\log r)^6 e^{-\sqrt{\log r}}}{r}\right)\,.$$

We conclude, using that for any integer $k > 1$, $(\log r)^k = o(e^{\sqrt{\log r}})$. Hence

$$\frac{(\log r)^6 e^{-\sqrt{\log r}}}{r} = o\left(\frac{1}{r(\log r)^2}\right)\,,$$

which is the general term of a convergent series. Hence we finally obtain

$$\sum_{r=r(\beta,K)}^{T} \sum_{j=\lfloor \sqrt{\log r}-1 \rfloor}^{\lfloor (\log r)^2 \rfloor} \alpha_k\left(\beta \frac{r}{K(K-1)(\log r)^2 j}, j\right) = O(1)\,.$$

□

Appendix 3.B LB-SDA with a Limited Memory

In this section the variant of LB-SDA using a limited storage memory introduced in Section 3.3.4 is analyzed. After introducing a few notations, we present a detailed version of the algorithm. We then provide a detailed proof of Theorem 3.8.

3.B.1 Notation for the Proof of Theorem 3.8

For simplifying the notation in this section, we assume without loss of generality that arm 1 is the optimal arm (with the highest mean). We now introduce the main notations.

- K number of arms
- ν_k distribution of the arm k, with mean μ_k. We assume that $\forall k$, $\nu_k \in \mathcal{P}$, a one-parameter exponential family.
- We assume that $\mu_1 = \max_{k\in[K]} \mu_k$ so we call the (unique) optimal arm "arm 1". In this part, the optimal arm is either denoted 1 or $k^\star$.
- $I_k(x)$ some large deviation rate function of the arm k, evaluated in x. For one-parameter exponential families this function will always be the KL-divergence between ν_k and the distribution from the same family with mean x.
- $N_k(r)$ number of pull of arm k up to (and including) round r.
- $Y_{k,i}$ reward obtained at the i-th pull of arm k.
- $\bar{Y}_{k,i}$ mean of the i-th first reward of arm k, $\bar{Y}_{k,n:m}$ mean of the rewards of k on a subset of indices $n < m$: $\bar{Y}_{k,n:m} = \frac{1}{m-n+1}\sum_{i=n}^{m} Y_{k,i}$. If $m - n = s$, then $\bar{Y}_{k,s}$ and $\bar{Y}_{k,n:m}$ have the same distribution.
- $\ell(r)$ leader at round r, $\ell(r) = \text{argmax}_{k\in\{1,\dots,K\}} N_k(r)$.
- $\mathcal{A}_r$ set of arms pulled at a round r.

Notations for the regret analysis, part relying on concentration:

- $\mathcal{Z}^r = \{\ell(r) \neq 1\}$, the leader used at round $r+1$ is suboptimal.
- $\mathcal{D}^r = \{\exists u \in \{\lfloor r/4 \rfloor, ..., r\}$ such that $\ell(u-1) = 1\}$, the optimal arm has been leader at least once between $\lfloor r/4 \rfloor$ and r.
- $\mathcal{B}^u = \{\ell(u) = 1, k \in \mathcal{A}_{u+1}, N_k(u) = N_1(u) - 1$ for some arm $k\}$, the optimal arm is leader in u but loses its duel against arm k, that have been pulled enough to possibly take over the leadership at next round.
- $\mathcal{C}^u = \{\exists k \neq 1, N_k(u) \geq N_1(u), \bar{Y}_{k,N_k(u)-N_1(u)+1:N_k(u)} \geq \bar{Y}_{1,N_1(u)}\}$, the optimal arm is not the leader and has lost its duel against the suboptimal leader.
- $\mathcal{L}^r = \sum_{u=\lfloor r/4 \rfloor}^{r} \mathbb{1}_{\mathcal{C}^u}$.

3.B.2 Proof of Theorem 3.8

The beginning of the proof of [Baudry et al., 2020] is valid for LB-SDA, however it has to be rewritten completely to introduce the storage limitation. We use the same notation as in Section 3.3.4 and introduce a sequence m_r of allowed memory for each arm at a round r. In the beginning of the proof we do not make any assumption on the sequence m_r except that $m_r/\log(r) \to +\infty$, which is required in the statement of Theorem 3.8. We further assume that m_r is an integer for any round r, which does not change anything for the algorithm but simplifies

the notations for the proof. In this section, without loss of generality, we assume that the arm 1 is the unique optimal arm $\mu_1 = \max_{k\in[K]} \mu_k$. We also recall that the arms are assumed to come from the same one-parameter exponential family of distributions.

In terms of notation, we remark that if $N_k(r) \geq m_r$ and $\ell(r) \neq k$ then the duel between k and $\ell(r)$ is the comparison between $\bar{Y}_{k,N_k(r)-m_r:N_k(r)}$ and $\bar{Y}_{\ell(r),N_{\ell(r)}(r)-m_r:N_{\ell(r)}(r)}$. Otherwise, if $N_k(r) \leq m_r$ and $\ell(r) \neq k$ then the duel is the comparison between $\bar{Y}_{k,N_k(r)}$ and $\bar{Y}_{\ell(r),N_{\ell(r)}(r)-N_k(r):N_{\ell(r)}(r)}$, which is the same as for the vanilla LB-SDA. We recall that the set of *saturated arms* at round r is defined as

$$\mathcal{S}_r = \{k \in \{1,\dots,K\} : N_k(r) \geq m_r\} \,. \tag{3.4}$$

However, we do not change the definition of the leader that is still defined as $\ell(r) = \text{argmax}_{k\leq K} N_k(r)$ nor the corresponding tie-breaking rules. All along the proof we will use the Chernoff inequality, that states that for any exponential family of distribution and any x, y satisfying $x < \mu_k < y$, then

$$\mathbb{P}\left(\bar{Y}_{k,n} \leq x\right) \leq e^{-n\text{kl}(x,\mu_k)} \text{ and } \mathbb{P}(\bar{Y}_{k,n} \geq y) \leq e^{-n\text{kl}(y,\mu_k)} \,.$$

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

We write $N_k(T)$ as $N_k(T) = 1 + \sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1})$. The first step of the proof is to decompose the number of pulls according to the events $\{\ell(r) = 1\}$ and $k \in \mathcal{S}_r$,

$$\begin{aligned}
\mathbb{E}[N_k(T)] &= 1 + \mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, \ell(r) \neq 1)\right] + \mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, k \in \mathcal{S}_r, \ell(r) = 1)\right] \\
&\quad + \mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, k \notin \mathcal{S}_r, \ell(r) = 1)\right] \\
&\leq 1 + \mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(\ell(r) \neq 1)\right] + \underbrace{\mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, k \in \mathcal{S}_r, \ell(r) = 1)\right]}_{E_1} \\
&\quad + \underbrace{\mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, k \notin \mathcal{S}_r, \ell(r) = 1)\right]}_{E_2} \,.
\end{aligned}$$

We first study the term $E_1 = \mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, k \in \mathcal{S}_r, \ell(r) = 1)\right]$ and use that under $k \in \mathcal{S}_r$ the index of both arms will be a subsample of size m_r of their history. We start the sum on the rounds at $2m_1$ because two arms cannot be saturated before this round is reached, so it

holds that

$$\begin{aligned}
E_1 &\leq \sum_{r=2m_1}^{T-1} \mathbb{P}\left(\ell(r)=1, k\in\mathcal{A}_{r+1}, N_k(r)\geq m_r, N_1(r)\geq m_r\right)\\
&\leq \sum_{r=2m_1}^{T-1} \mathbb{P}\left(N_k(r)\geq m_r, \bar{Y}_{k,N_k(r)-m_r+1:N_k(r)}\geq x_k\right)\\
&\quad + \sum_{r=2m_1}^{T-1} \mathbb{P}\left(N_1(r)\geq m_r, \bar{Y}_{1,N_1(r)-m_r+1:N_1(r)}\leq x_k\right)\\
&\leq \sum_{r=2m_1}^{T-1}\sum_{n_k=m_r}^{r} \mathbb{P}\left(\bar{Y}_{k,n_k-m_r+1:n_k}\geq x_k, N_k(r)=n_k\right)\\
&\quad + \sum_{r=2m_1}^{T-1}\sum_{n_1=m_r}^{r} \mathbb{P}\left(\bar{Y}_{1,n_1-m_r+1:n_1}\leq x_k, N_1(r)=n_1\right)\\
&\leq \sum_{r=2m_1}^{T-1}\sum_{n_k=m_r}^{r} \mathbb{P}\left(\bar{Y}_{k,n_k-m_r+1:n_k}\geq x_k\right) + \sum_{r=2m_1}^{T-1}\sum_{n_1=m_r}^{r} \mathbb{P}\left(\bar{Y}_{1,n_1-m_r+1:n_1}\leq x_k\right)\\
&\leq 2\sum_{r=2m_1}^{T-1} re^{-m_r\omega_k} \ ,
\end{aligned}$$

where we used two main elements: 1) if two random variables X and Y satisfy $X\geq Y$ then for any threshold η it holds that either $X\geq\eta$ or $Y\leq\eta$ (second line), and 2) the empirical averages of the fixed blocks of observations satisfy the Chernoff concentration inequality. Using the notation, we introduced

$$\mathbb{P}(\bar{Y}_{1,n_1-m_r+1:n_1}\leq x_k) = \mathbb{P}(\bar{Y}_{1,m_r}\leq x_k)\leq e^{-m_r\omega_k}$$

and

$$\mathbb{P}(\bar{Y}_{k,n_k-m_r+1:n_k}\geq x_k) = \mathbb{P}(\bar{Y}_{k,m_r}\geq x_k)\leq e^{-m_r\omega_k} \ .$$

Therefore, the following holds

$$E_1 = \sum_{r=1}^{T-1}\mathbb{P}(k\in\mathcal{A}_{r+1}, k\in\mathcal{S}_r, \ell(r)=1)\leq 2\sum_{r=2m_1}^{T-1} re^{-m_r\omega_k} \ . \tag{3.5}$$

We then study $E_2 = \mathbb{E}\left[\sum_{r=1}^{T-1}\mathbb{1}(k\in\mathcal{A}_{r+1}, k\notin\mathcal{S}_r, \ell(r)=1)\right]$. We further distinguish two cases, whenever $N_k(r)\leq n_0(T)$ holds or not at each round, for some $n_0(T)$ that will be specified later.

$$E_2 \leq n_0(T) + \mathbb{E}\left[\sum_{r=1}^{T-1}\mathbb{1}(k\in\mathcal{A}_{r+1}, k\notin\mathcal{S}_r, \ell(r)=1, N_k(r)\geq n_0(T))\right] \ .$$

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

$$
\begin{aligned}
E_2 &\leq n_0(T) + \mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, k \notin \mathcal{S}_r, \ell(r) = 1, N_k(r) \geq n_0(T))\right] \\
&\leq n_0(T) + \mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, \bar{Y}_{k,N_k(r)} \geq \bar{Y}_{1,N_1(r)-N_k(r)+1:N_1(r)}, N_k(r) \geq n_0(T))\right] \\
&\leq n_0(T) + \sum_{r=1}^{T-1} \mathbb{P}\left(k \in \mathcal{A}_{r+1}, \bar{Y}_{k,N_k(r)} \geq \eta_k, N_k(r) \geq n_0(T)\right) \\
&+ \sum_{r=1}^{T-1} \mathbb{P}\left(k \in \mathcal{A}_{r+1}, \bar{Y}_{1,N_1(r)-N_k(r)+1:N_1(r)} \leq \eta_k, N_k(r) \geq n_0(T), N_1(r) \geq n_0(T)\right) ,
\end{aligned}
$$

where we used the same trick as for E_1 to obtain the last result. We then use a union bound on the values of $N_k(r)$ for the first sum and on both $N_k(r)$ and $N_1(r)$ for the second sum, leading to

$$
\begin{aligned}
E_2 &\leq n_0(T) + \sum_{r=1}^{T-1} \sum_{n_k=n_0(T)}^{T-1} \mathbb{P}\left(k \in \mathcal{A}_{r+1}, \bar{Y}_{k,n_k} \geq \eta_k, N_k(r) = n_k\right) \\
&+ \sum_{r=1}^{T-1} \sum_{n_1=n_0(T)}^{T-1} \sum_{n_k=n_0(T)}^{n_1} \mathbb{P}\left(k \in \mathcal{A}_{r+1}, \bar{Y}_{1,n_1-n_k+1:n_1} \leq \eta_k, N_k(r) = n_k, N_1(r) = n_1\right) \\
&\leq n_0(T) + \sum_{n_k=n_0(T)}^{T-1} \mathbb{P}\left(\bar{Y}_{k,n_k} \geq \eta_k\right) + \sum_{n_k=n_0(T)}^{T-1} \sum_{n_1=n_0(T)}^{T-1} \mathbb{P}\left(\bar{Y}_{1,n_1-n_k+1:n_1} \leq \eta_k\right) ,
\end{aligned}
$$

where we used that $\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, N_k(r) = n_k) \leq 1$ to remove the sums in r (simply ignoring the event $N_1(r) = n_1$ in the second term). Using the Chernoff inequality, we write

$$
E_2 \leq n_0(T) + \frac{e^{-n_0(T)\mathrm{kl}(\eta_k,\mu_k)}}{1 - e^{-\mathrm{kl}(\eta_k,\mu_k)}} + T\frac{e^{-n_0(T)\mathrm{kl}(\eta_k,\mu_1)}}{1 - e^{-\mathrm{kl}(\eta_k,\mu_1)}} .
$$

We then calibrate $n_0(T)$ and η_k in order to make these terms converge properly. We define $\epsilon > 0$ and take $n_0(T) = \frac{1+\epsilon}{\mathrm{kl}(\mu_k,\mu_1)} \log T$. We then use the continuity of the kullback-leibler divergence on (μ_k, μ_1) to state that for any $\delta > 0$, there exists some $\epsilon > 0$ and $\eta_k \in (\mu_k, \mu_1)$ satisfying $\mathrm{kl}(\eta_k, \mu_1) \geq \mathrm{kl}(\mu_k, \mu_1) - \delta \geq \frac{\mathrm{kl}(\mu_k,\mu_1)}{1+\epsilon}$. This means that for any $\epsilon > 0$, there exists some $\eta_k > 0$ satisfying $Te^{-n_0(T)\mathrm{kl}(\eta_k,\mu_1)} \leq Te^{-n_0(T)\frac{\mathrm{kl}(\mu_k,\mu_1)}{1+\epsilon}} \leq 1$. Hence, for any $\epsilon > 0$ it holds that

$$
E_2 \leq \frac{1+\epsilon}{I_1(\mu_k)} \log T + C_{k,\epsilon} ,
$$

where $C_{k,\epsilon}$ is a constant.

Combining these results we can write a first decomposition of $\mathbb{E}[N_k(T)]$ as

$$
\mathbb{E}[N_k(T)] \leq 1 + \frac{1+\epsilon}{I_1(\mu_k)} \log T + 2 \sum_{r=2m_1}^{T-1} re^{-m_r\omega_k} + C_{k,\epsilon} + \sum_{r=2m_1}^{T-1} \mathbb{P}(\ell(r) \neq 1) . \tag{3.6}
$$

We remark that this expression provides an explicit dependence in m_r in the second term, that justifies the condition in Theorem 3.8 for m_r (namely, $m_r/(\log r) \to +\infty$). Indeed, this

condition is sufficient to ensure for instance that $m_r \geq \frac{3}{\omega_k} \log r$ for r large enough, making the term inside the sum a $o(r^{-2})$.

The next step is to prove that $\sum_{r=1}^{T-1} \mathbb{P}(\ell(r) \neq 1) = o(\log T)$. As in the proof of [Chan, 2020] this part causes a lot of technical challenges, and we need to define several new events to analyze the different scenarios that could lead a suboptimal arm to be the leader at a round r. In the next steps we will consider the same events as in the original proof, but the storage limitation will add some complexity to the task. We will use the following property, issued from the definition of the leader

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

Adding the storage constraint we have that for any r satisfying $r \geq K m_r$ the leader has necessarily more than m_r observations. For this reason, its history will be truncated to the m_r last observations. However, we leverage the property that when r is reasonably large, m_r is large enough to guarantee a good concentration of the empirical mean of the saturated arms around their true mean. We will explain how this can be done in this section. We define $a_r = \lceil \frac{r}{4} \rceil$, and write the following decomposition

$$\mathbb{P}\left(\ell(r) \neq 1\right) = \mathbb{P}\left(\{\ell(r) \neq 1\} \cap \mathcal{D}^r\right) + \mathbb{P}\left(\{\ell(r) \neq 1\} \cap \bar{\mathcal{D}}^r\right) , \tag{3.7}$$

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

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

We now explain how to upper bound the term in the left hand side of Equation (3.7). We look at the rounds larger than some round r_0 that will be specified later in the proof.

3.B.2.1 Arm 1 has been leader between a_r and r

We introduce a new event

$$\mathcal{B}^u = \{\ell(u) = 1, k \in \mathcal{A}_{u+1}, N_k(u) = N_1(u) - 1 \text{ for some arm } k\} .$$

Under the event $\mathcal{D}^r$, $\{\ell(r) \neq 1\}$ can only be true only if the leadership has been taken over by a suboptimal arm at some round between a_r and r, that is

$$\{\ell(r) \neq 1\} \cap \mathcal{D}^r \subset \cup_{u=a_r}^{r-1} \{\ell(u) = 1, \ell(u+1) \neq 1\} \subset \cup_{u=a_r}^{r-1} \mathcal{B}^u . \tag{3.8}$$

Indeed, a leadership takeover can only happen after a challenger has defeated the leader while having at least the same number of observations minus one (however this situation is necessary but not sufficient to cause a change of leader, hence the strict inclusion). We now upper bound $\sum_{r=r_0}^{T-1} \sum_{u=a_r}^{r} \mathbb{P}(\mathcal{B}^u)$. We use the notation $b_r = \lceil a_r / K \rceil$ representing the minimum of samples of the leader at the round a_r. Hence we are sure that under $\mathcal{B}^u$ arm 1 had at least b_u observations when it lost the duel that cost it the leadership. We then take an union bound on all the suboptimal arms $k \in \{2, ..., K\}$, with $\mathcal{B}^u = \cup_{k=2}^{K} \mathcal{B}_k^u$ where

$$\mathcal{B}_k^u \coloneqq \{\ell(u) = 1, k \in \mathcal{A}_{u+1}, N_k(u) = N_1(u) - 1\} .$$

This fixes the specific suboptimal arm that could have taken the leadership.

Choosing x_k, ω_k as in the previous section we can write

$$\begin{aligned}\sum_{r=r_0}^{T-1}\sum_{u=a_r}^{r}\mathbb{P}(\mathcal{B}_k^u) &= \mathbb{E}\left[\sum_{r=r_0}^{T-1}\sum_{u=a_r}^{r}\mathbb{1}(\ell(u)=1, k\in\mathcal{A}_{u+1}, N_1(u)=N_k(u)+1)\right]\\ &\le \underbrace{\mathbb{E}\left[\sum_{r=r_0}^{T-1}\sum_{u=a_r}^{r}\mathbb{1}(\ell(u)=1, k\in\mathcal{A}_{u+1}, N_1(u)=N_k(u)+1, k\notin\mathcal{S}_u)\right]}_{B_1}\\ &\quad+\underbrace{\mathbb{E}\left[\sum_{r=r_0}^{T-1}\sum_{u=a_r}^{r}\mathbb{1}(\ell(u)=1, k\in\mathcal{A}_{u+1}, N_1(u)=N_k(u)+1, k\in\mathcal{S}_u)\right]}_{B_2}.\end{aligned}$$

We proceed similarly as in the previous part, analyzing separately the case $k\in\mathcal{S}_u$ and the case $k\notin\mathcal{S}_u$ with $\mathcal{S}_u$ defined in Equation (3.4). We start with the term B_1,

$$B_1 \le \mathbb{E}\left[\sum_{r=r_0}^{T-1}\sum_{u=a_r}^{r}\mathbb{1}(N_1(u)\ge b_r, \bar{Y}_{k,N_k(u)}\ge x_k, N_1(u)=N_k(u)+1, k\in\mathcal{A}_{u+1}, k\notin\mathcal{S}_u)\right] \tag{3.9}$$

$$+\mathbb{E}\left[\sum_{r=r_0}^{T-1}\sum_{u=a_r}^{r}\mathbb{1}(N_1(u)\ge b_r, \bar{Y}_{1,N_1(u)-N_k(u)+1:N_1(u)}\le x_k, N_1(u)=N_k(u)+1, k\in\mathcal{A}_{u+1})\right]. \tag{3.10}$$

We now separately upper bound each of these two terms. First,

$$\begin{aligned}\text{Equation 3.9} &\le \mathbb{E}\left[\sum_{r=r_0}^{T-1}\sum_{u=a_r}^{r}\sum_{n_k=b_r-1}^{m_u-1}\mathbb{1}(N_k(u)=n_k, k\in\mathcal{A}_{u+1}, \bar{Y}_{k,n_k}\ge x_k)\right]\\ &\le \mathbb{E}\left[\sum_{r=r_0}^{T-1}\sum_{n_k=b_r-1}^{r}\mathbb{1}(\bar{Y}_{k,n_k}\ge x_k)\underbrace{\sum_{u=a_r}^{r}\mathbb{1}(N_k(u)=n_k)\mathbb{1}(k\in\mathcal{A}_{u+1})}_{\le 1}\right]\\ &\le \sum_{r=r_0}^{T-1}\sum_{n_k=b_r-1}^{r}\mathbb{P}(\bar{Y}_{k,n_k}\ge x_k)\\ &\le \sum_{r=r_0}^{T-1}\sum_{n_k=b_r-1}^{r}\exp\left(-n_k\omega_k\right)\\ &\le \sum_{r=r_0}^{T-1}\frac{e^{-(b_r-1)\omega_k}}{1-e^{-\omega_k}}.\end{aligned}$$

We remark that by definition $b_r \ge a_r/K \ge r/(4K)$ and using $r_0\ge 8$, we conclude that

$$\text{Equation 3.9} \le \frac{e^{(1-\frac{2}{K})\omega_k}}{(1-e^{-\omega_k})(1-e^{-\omega_k/(4K)})}.$$

As the subsampling in LB-SDA is deterministic, thanks to $N_1(r)=N_k(u)+1$ we obtain the

same result for Equation (3.10),

$$\begin{aligned}
\text{Equation 3.10} &\leq \mathbb{E}\left[\sum_{r=r_0}^{T-1}\sum_{u=a_r}^{r}\sum_{n_k=b_r-1}^{r}\mathbb{1}(\bar{Y}_{1,n_k}\leq x_k)\mathbb{1}(N_k(u)=n_k)\mathbb{1}(k\in\mathcal{A}_{u+1})\right] \\
&\leq \mathbb{E}\left[\sum_{r=r_0}^{T-1}\sum_{n_k=b_r-1}^{r}\mathbb{1}(\bar{Y}_{1,n_k}\leq x_k)\underbrace{\sum_{u=a_r}^{r}\mathbb{1}(N_k(u)=n_k)\mathbb{1}(k\in\mathcal{A}_{u+1})}_{\leq 1}\right] \\
&\leq \sum_{r=r_0}^{T-1}\sum_{n_k=b_r-1}^{r}\mathbb{P}(\bar{Y}_{1,n_k}\leq x_k) \\
&\leq \sum_{r=r_0}^{T-1}\sum_{n_k=b_r-1}^{r}\exp\left(-n_k\omega_k\right) \\
&\leq \frac{e^{(1-\frac{2}{K})\omega_k}}{(1-e^{-\omega_k})(1-e^{-\omega_k/(4K)})}\,.
\end{aligned}$$

We then control B_2. For B_2 the condition $N_1(u)=N_k(u)+1$ will not be used but instead we use Equation (3.5) already established in the previous section.

$$\sum_{u=1}^{r}\mathbb{P}(k\in\mathcal{A}_{u+1},k\in\mathcal{S}_u,\ell(u)=1)\leq 2\sum_{u=2m_1}^{r}ue^{-m_u\omega_k}\,,$$

which leads to

$$\begin{aligned}
B_2 &= \mathbb{E}\left[\sum_{r=r_0}^{T-1}\sum_{u=a_r}^{r}\mathbb{1}(\ell(u)=1,k\in\mathcal{A}_{u+1},N_1(u)=N_k(u)+1,k\in\mathcal{S}_u)\right] \\
&\leq \sum_{r=r_0}^{T-1}\sum_{u=\max(a_r,2m_1)}^{r}2ue^{-m_u\omega_k}\,.
\end{aligned}$$

Then, if we consider $r_0=\min\{r:a_r\geq 2m_1\}$ we can further upper bound B_2 by

$$B_2\leq\sum_{r=r_0}^{T-1}\sum_{u=a_r}^{r}2ue^{-m_u\omega_k}\leq 2\sum_{r=r_0}^{T-1}r\sum_{u=a_r}^{r}2e^{-m_u\omega_k}\leq 2\sum_{r=r_0}^{T-1}r^2e^{-m_{a_r}\omega_k}\,.$$

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

$$\sum_{r=r_0}^{T-1}\mathbb{P}\left(\{\ell(r)\neq 1\}\cap\mathcal{D}^r\right)\leq 2\sum_{k=2}^{K}\left[\frac{e^{(1-\frac{2}{K})\omega_k}}{(1-e^{-\omega_k})(1-e^{-\omega_k/(4K)})}+\sum_{r=r_0}^{T-1}r^2e^{-m_{a_r}\omega_k}\right]\,. \tag{3.11}$$

Hence, the sums of the probability that arm 1 is not the leader while it has already been before is upper bounded by two terms: a problem-dependent constant, and a term that depends of the sequence of memory limits $(m_r)_{r\geq 1}$. We can further analyze this second term. First, we remark that contrarily to the term in m_r in Equation (3.6) this time we have both r^2 and m_{a_r} instead of m_r, with $a_r=\lceil r/4\rceil$. Hence, for a fixed r the term of the sum is larger in this case. However, the constraint $m_r/\log(r)\to+\infty$ is again sufficient to ensure a proper convergence of this sum to a constant with the same arguments. This is mainly because the choice of a_r as a fraction of r ensures that m_{a_r} will be sufficiently large.

3.B.2.2 Arm 1 has never been leader between a_r and r

The idea in this part is to leverage the fact that if the optimal arm is not leader between $\lfloor r/4 \rfloor$ and r, then it has necessarily lost a lot of duels against the current leader at each round. We then use the fact that when the leader has been drawn "enough", concentration prevents this situation with large probability. We introduce

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

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

$$\mathbb{P}(\ell(r) \neq 1 \cap \bar{\mathcal{D}}^r) \leq \mathbb{P}(\mathcal{L}^r \geq r/4) \ . \tag{3.12}$$

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

$$\mathbb{P}(\mathcal{L}^r \geq r/4) \leq \frac{\mathbb{E}(\mathcal{L}^r)}{r/4} = \frac{4}{r} \sum_{u=a_r}^{r} \mathbb{P}(\mathcal{C}^u) \ . \tag{3.13}$$

We further decompose the probability of $\mathbb{P}(\mathcal{C}^u)$ in two parts depending on the value of the number of selections of arm 1. For the next steps we define the following events, $\{N_1(u) \leq C/4 \log(u)\}$ and $\{N_1(u) \geq C/4 \log(u)\}$, for some constant C that is not known by the algorithm and that we will define later. The idea is to handle the memory limit through this parameter C. Indeed, we only know that the sequence $(m_r)_{r\geq 1}$ satisfies $m_r/(\log(r)) \to +\infty$. For this reason, we know that for any $C > 0$ there exists a round r_C such that for any $r \geq r_C$ then $m_r \geq C \log(r)$. Using Equation (3.12) and Equation (3.13), we have

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

Again, D can be upper bounded by splitting the cases when the optimal arm is saturated or not. We define $D_{k,1}$ and $D_{k,2}$ as

$$\begin{aligned} D_{k,1} &\coloneqq \sum_{r=r_0}^{T-1} \frac{4}{r} \sum_{u=a_r}^{r} \mathbb{P}\left(\mathcal{C}_k^u, N_1(u) \geq \frac{C}{4} \log(u), 1 \in \mathcal{S}_u\right) \\ D_{k,2} &\coloneqq \sum_{r=r_0}^{T-1} \frac{4}{r} \sum_{u=a_r}^{r} \mathbb{P}\left(\mathcal{C}_k^u, N_1(u) \geq \frac{C}{4} \log(u), 1 \notin \mathcal{S}_u\right) \end{aligned}$$

We also introduce $\mathcal{C}_k^u \coloneqq \{\ell(u) = k, 1 \notin \mathcal{A}_{u+1}\}$ for any $k \in \{2, \dots, K\}$ and obtain

$$D \leq \sum_{k=2}^{K} (D_{k,1} + D_{k,2}) \ .$$

For the event featuring $\{1 \in \mathcal{S}_u\}$ we can use the result of the previous sections because in the event we consider there is no difference between $\ell(r) = 1$ and $\ell(r) = k$ when both arms are saturated. Following the proof for obtaining Equation (3.5), one has

$$\sum_{u=a_r}^{r} \mathbb{P}(1 \notin \mathcal{A}_{u+1}, 1 \in \mathcal{S}_u, \ell(u) = k) \leq 2 \sum_{u=a_r}^{r} u e^{-m_u \omega_k} \ . \tag{3.14}$$

With this result we then obtain

$$\begin{aligned}
D_{k,1} &= \sum_{r=r_0}^{T-1} \frac{4}{r} \sum_{u=a_r}^{r} \mathbb{P}\left(\mathcal{C}_k^u, 1 \in \mathcal{S}_u\right) \\
&\leq \sum_{r=r_0}^{T-1} \frac{4}{r} \sum_{u=a_r}^{r} \mathbb{P}\left(1 \notin \mathcal{A}_{u+1}, 1 \in \mathcal{S}_u, \ell(u) = k\right) \\
&\leq \sum_{r=r_0}^{T-1} \frac{4}{r} \sum_{u=a_r}^{r} 2u e^{-m_u \omega_k} \quad \text{(Equation 3.14)} \\
&\leq 8 \sum_{r=r_0}^{T-1} \sum_{u=a_r}^{r} e^{-m_u \omega_k} \\
&\leq 8 \sum_{r=r_0}^{T-1} r e^{-m_{a_r} \omega_k} \ .
\end{aligned}$$

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

So finally

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

At this step we remark that we need to choose the constant C large enough in order to make this sum converge to a constant. We remind here, that C is only an analysis parameter. We then consider the term B. As in [Baudry et al., 2020] we transform the double sum in a simple sum by simply counting the number of times each term is included. For any integer s and any round r, the term $\frac{4}{s}$ appears only if $a_s \leq r \leq s$. With the value $a_r = \lceil \frac{r}{4} \rceil$ we obtain

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

If we remark that $\sum_{t=1}^{r} \frac{4}{t}\mathbb{1}(t \in [s, 4s]) \leq (4s - s + 1) \times \frac{4}{s} \leq 16$, we finally get:

$$\sum_{r=r_0}^{T} \mathbb{P}(\{\ell(r) \neq 1\} \cap \overline{\mathcal{D}}^r) \leq r_0 + 16 \sum_{r=r_0}^{T} \mathbb{P}\left(N_1(r) \leq \frac{C}{4}\log(r)\right) + D(\boldsymbol{\nu}). \tag{3.15}$$

Combining Equation (3.11) and Equation (3.15) yields

$$\sum_{r=r_0}^{T} \mathbb{P}\left(\ell(r) \neq 1\right) \leq r_0 + 16 \sum_{r=r_0}^{T} \mathbb{P}\left(N_1(r) \leq \frac{C}{4}\log(r)\right) + D'_k(\boldsymbol{\nu})$$

for some constant $D'_k(\boldsymbol{\nu})$ that depends on k and $\boldsymbol{\nu}$. Hence, the storage limit may introduce larger constant terms in the proof, but asymptotically the dominant terms are the same as in the proof of the vanilla LB-SDA algorithm.

The last step is to show that we can upper the last term as we did in Appendix 3.A. To do so, we only need to prove that if r_0 is large enough and $\{N_1(r) \leq C/4\log(r)\}$, then the arm 1 has not been saturated for a long time. This way we would handle the saturation exactly as we handled the forced exploration (which is still present here) in the proof for the vanilla LB-SDA. To do so, we define the function $m^{-1}(x) = \inf\{r : m_r \geq x\}$. If we had exactly $m_r = C\log r$ then this function would be $m^{-1}(x) = \exp(x/C)$. Up to choosing a slightly larger r_0, we consider that for any $r > r_0$ we also have $m^{-1}(C/4\log r) \leq \exp(C/4\log(r)C^{-1}) = r^{1/4}$. Hence, after the round r_0 we are sure that arm 1 has never been saturated since the round $r^{1/4}$, hence we can apply the same sketch of proof as in Appendix 3.A to conclude that

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

Appendix 3.C Proof for Switching Bandits

Bounding $\mathbb{E}[N_k^\phi]$, the number of pulls of a suboptimal arm k during a *phase* ϕ is sufficient to control the dynamic regret. During the phase ϕ the best arm is denoted k^*_ϕ. We consider the SW-LB-SDA policy with a sliding window of size τ. We also define $\hat{\delta}_\phi = r_{\phi+1} - r_\phi$, the random number of rounds in the phase ϕ. Due to the sliding window, we use the definition of the leader introduced in Section 3.4 and recall that $N_k^\tau(r) = \sum_{s=r-\tau}^{r-1} \mathbb{1}\left(k \in \mathcal{A}_{s+1}\right)$, i.e. number of times arm k has been pulled during the τ last rounds.

3.C.1 Details for SW-LB-SDA Implementation

With our new definition of the leader, it could happen that for some rounds the leader is not the arm with the largest number of samples when $K \geq 3$. We give an example of such a behavior: assume that the first round is $r = 1$, there are $2n + m$ rounds and $K = 3$ arms drawn in the following order (1 arm per round): m pulls of arm 1, followed by $n > m$ pulls of arm 3 and then $n - m$ pulls of arm 1. If the length of the sliding window is $\tau = 2n$ and the leader at the round $(m + n + (n - m) = 2n)$ is 1, then we see that 1 will lose samples during the next m rounds. If for those m successive rounds only the arm 2 is pulled, then 1 will stay leader with $n - m$ samples while 3 still have n samples. At the end (round $2n + m$), the leader is arm 1, we have $N_1^\tau(2n + m) = n - m < N_3^\tau(2n + m) = n$. This example highlights that is it possible that the leader is not the arm that has been played the most with a sliding window.

For this reason, the duels are slightly different to the stationary case. The index of the leader for duels against an arm with a larger number of samples is simply the mean of its observations collected during the last τ rounds. Indeed, in this case both arms have a large number of samples hence subsampling is not necessary. This explain why the term $\hat{\mu}^{\tau}_{\ell,k}$ is used in Algorithm 9.

3.C.2 Analysis

We use the notation introduced in Section 3.4. The beginning of the proof takes elements from [Garivier and Moulines, 2008] and [Baudry et al., 2020]. For $k \neq k^*_\phi$ and an arbitrary function $A^{\phi,\tau}_k$, we write

$$\begin{aligned}
N^{\phi}_k &= \sum_{r=r_\phi-1}^{r_{\phi+1}-2} \mathbb{1}\left(k \in \mathcal{A}_{r+1}\right) \leq 2\tau + \sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(k \in \mathcal{A}_{r+1}\right) \\
&\leq 2\tau + \sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(k \in \mathcal{A}_{r+1}, \ell^\tau(r) = k^*_\phi, N^\tau_k(r) \geq A^{\phi,\tau}_k\right) \\
&\quad + \sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(k \in \mathcal{A}_{r+1}, N^\tau_k(r) < A^{\phi,\tau}_k\right) + \sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(k \in \mathcal{A}_{r+1}, \ell^\tau(r) \neq k^*_\phi\right) \\
&\leq 2\tau + \sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(k \in \mathcal{A}_{r+1}, \ell^\tau(r) = k^*_\phi, N^\tau_k(r) \geq A^{\phi,\tau}_k, D^\tau_k(r) = 0\right) \\
&\quad + \sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(\ell^\tau(r) = k^*_\phi, D^\tau_k(r) = 1\right) \\
&\quad + \sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(k \in \mathcal{A}_{r+1}, N^\tau_k(r) < A^{\phi,\tau}_k\right) + \sum_{r=r_\phi+2\tau-1}^{r_{\phi+1}-2} \mathbb{1}\left(k \in \mathcal{A}_{r+1}, \ell^\tau(r) \neq k^*_\phi\right) .
\end{aligned}$$

We then use the following lemma.

Lemma 3.13 (Adaptation of Lemma 25 from [Garivier and Moulines, 2008])**.**

$$\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(k \in \mathcal{A}_{r+1}, N^\tau_k(r) < A\right) \leq \frac{\widehat{\delta}_\phi A}{\tau} .$$

Therefore,

$$\begin{aligned}
N^{\phi}_k \leq 2\tau + \frac{\widehat{\delta}_\phi A^{\phi,\tau}_k}{\tau} &+ \underbrace{\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(k \in \mathcal{A}_{r+1}, \ell^\tau(r) = k^*_\phi, N^\tau_k(r) \geq A^{\phi,\tau}_k, D^\tau_k(r) = 0\right)}_{c^{\phi,\tau}_{k,1}} \\
&+ \underbrace{\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(\ell^\tau(r) = k^*_\phi, D^\tau_k(r) = 1\right)}_{c^{\phi,\tau}_{k,2}} + \underbrace{\sum_{r=r_\phi+2\tau-1}^{r_{\phi+1}-2} \mathbb{1}\left(\ell^\tau(r) \neq k^*_\phi\right)}_{c^{\phi,\tau}_{k,3}} .
\end{aligned}$$

We control the expectation of these terms separately.

3.C.2.1 Upper bounding $\mathbb{E}[c_{k,1}^{\phi,\tau}]$

We recall that

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

We start by stating a lemma on the concentration of subsample means in Last Block sampling that is crucial for the proof.

Lemma 3.14. *We consider a stationary phase ϕ and the multi-arm bandit model characterized by $(\nu_1^\phi, \dots, \nu_K^\phi)$. Let k_ϕ^* denote the arm with the largest mean.*

Then, for any constant $n \in \mathbb{N}$ satisfying $n \geq f(\tau) = \sqrt{\log \tau}$, by letting $\tilde{n} = \min(n, \lfloor \tau/(2K) \rfloor)$ it holds that

$$\mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(k \in \mathcal{A}_{r+1}, \ell^\tau(r) = k_\phi^*, N_k^\tau(r) \geq n, D_k^\tau(r) = 0\right)\right] \leq \delta_\phi(\tau+1)\frac{e^{-\tilde{n}\omega_k}}{1-e^{-\omega_k}} , \tag{3.16}$$

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

$$\mathbb{E}\left[\sum_{r=r_\phi+\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(k_\phi^* \notin \mathcal{A}_{r+1}, \ell^\tau(r) = k, N_{k_\phi^*}^\tau(r) \geq n\right)\right] \leq \delta_\phi(\tau+1)\frac{e^{-\tilde{n}\omega_k}}{1-e^{-\omega_k}} . \tag{3.17}$$

Proof. We start with the first claim. Under the considered event, an arm k can be drawn for three reason: 1) $D_k^\tau(r) = 1$, the diversity flag of this arm is raised 2) $N_k^\tau(r) \leq \sqrt{\log \tau}$, the forced exploration is used, or 3) k has won its duel against the leader k_ϕ^* here. In our case, as $D_k^\tau(r) = 0$ and $N_k^\tau(r) \geq n \geq \sqrt{\log \tau}$, if k is pulled while k_ϕ^* is leader then k has won its duel against k_ϕ^*.

Under this event, the duel between k and k_ϕ^* is a comparison between the mean of two blocks containing at least $\min(n, \tau/(2K))$ observations because of the definition of the leader. As in [Baudry et al., 2020] we use that for any threshold ξ_k, k wins the duel only if either $\widehat{\mu}_k^\tau(r) \geq \xi_k$ or $\widehat{\mu}_{\ell,k}^\tau(r) \leq \xi_k$. For the sake of simplicity in our results we choose ξ_k as the number satisfying $\xi_k = \frac{1}{2}(\mu_k^\phi + \mu_{k_\phi^*}^\phi)$, and this choice will remain the same for the rest of the appendix. We then write

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

First note that for a given arm k all possible blocks of observations are uniquely described by two quantities: $N_k^{\phi}(r)$ the number of observations of arm k from the beginning of the phase ϕ and $N_k^{\tau}(r)$ number of observations of arm k over the last τ rounds. We will use this property to bound the two previous sums.
Starting by the simpler term featuring the arm k, we use

$$\{k \in \mathcal{A}_{r+1}, \widehat{\mu}_k^{\tau}(r) \geq \xi_k, N_{k_\phi^*}^{\tau}(r) \geq \frac{\tau}{2K}, N_k^{\tau}(r) \geq n\} \subset \{k \in \mathcal{A}_{r+1}, \widehat{\mu}_k^{\tau}(r) \geq \xi_k, N_k^{\tau}(r) \geq n\} \ . \quad (3.18)$$

N_k^{ϕ} is defined by $N_k^{\phi}(r) = \sum_{s=r_\phi-1}^{r-1} \mathbb{1}(k \in \mathcal{A}_{s+1})$. For a given round r if the indicator from the RHS of Equation (3.18) is equal to 1, it implies that there is a block of length at least n with a mean at least ξ_k. More formally, when introducing

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

the following holds,

$$\{k \in \mathcal{A}_{r+1}, \widehat{\mu}_k^{\tau}(r) \geq \xi_k, N_k^{\tau}(r) \geq n\} \subset \bigcup_{n_k=n}^{\hat{\delta}_\phi} \bigcup_{m_k=1}^{\hat{\delta}_\phi} S_k^{n_k,m_k}(r) \ . \quad (3.19)$$

For the sake of clarity, we denote $Y_{k,1}, ..., Y_{k,\hat{\delta}_\phi}$ the set of possible rewards for the arm k for the phase ϕ. If the indicator function equals one for a given round r_0, then $\{k \in \mathcal{A}_{r_0+1}\}$ holds. The same block (same value for both n and m) can not be used for upcoming rounds because $N_k^{\phi}(r_0+1)$ will satisfy $N_k^{\phi}(r_0+1) = 1 + N_k^{\phi}(r_0)$. More specifically, for the arm k for any possible block there is at most one round for which the indicator function can be 1., i.e.

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

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

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

We also have

$$\{k \in \mathcal{A}_{r+1}, \widehat{\mu}_{k_\phi^\star,k}^{\tau}(r) \leq \xi_k, N_{k_\phi^\star}^{\tau}(r) \geq n'\} \subset \bigcup_{n^\star=n'}^{\hat{\delta}_\phi} \bigcup_{m^\star=1}^{\hat{\delta}_\phi} S_{k_\phi^\star}^{n^\star,m^\star}(r) \ . \quad (3.20)$$

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

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

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

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

We then have to handle carefully the fact that $\widehat{\delta}_\phi$ is actually a random variable depending on the bandit algorithm. Indeed, as several arms can be pulled at each round we don't know what will be the length of a phase in terms of rounds. However, this quantity is upper bounded by the actual length of the phase in terms of arms pulled δ_ϕ.
Thus, using the concentration inequality corresponding to the family of distributions for an appropriate rate function we can write

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

where in the last inequality we have introduced $\tilde{n} = \min(n, n') = \min(n, \lfloor \tau/(2K) \rfloor)$.
Finally, the proof of the second statement is a direct adaptation of this proof by inverting k and k_ϕ^*. We don't need the event $D_k^\phi(r) = 0$ because if k_ϕ^* is not drawn it has necessarily lost its duel against the leader k. □

We then remark that Equation (3.16) in Lemma 3.14 can be used to upper bound the term $c_{k,1}^{\phi,\tau}$, by replacing n by $A_k^{\phi,\tau}$. Assuming that $A_k^{\phi,\tau} \leq \tau/(2K)$ it holds that

$$\mathbb{E}[c_{k,1}^{\phi,\tau}] \leq \delta_\phi(\tau+1)\frac{e^{-A_k^{\phi,\tau}\omega_k}}{1-e^{-\omega_k}} . \tag{3.21}$$

3.C.2.2 Upper bounding $\mathbb{E}[c_{k,2}^{\phi,\tau}]$

We recall that,

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

To upper bound $\mathbb{E}[c_{k,2}^{\phi,\tau}]$ we have to study the probability that the optimal arm for the phase ϕ loses $\lceil (K-1)(\log\tau)^2 \rceil$ successive duels while being leader. We derive in Lemma 3.15 an intuitive consequence of this property: the optimal arm has necessarily lost at least one duel against a concentrated arm.

Lemma 3.15. *Consider K arms, and assume that some arm k has been leader for M consecutive rounds with $M \leq \tau$. For any m satisfying $(K-1)m \leq M$, if k has lost more than $(K-1)m$ duels then it has lost at least one duel against an arm with more than m samples.*

Proof. We assume that arm k has been leader for M consecutive rounds and that arm k lost strictly more than $(K-1)m$ duels. We also assume that all the challengers that have won against the arm k have less than m samples. We assume there exists an arm $k' \neq k$ such that k' won at least $m+1$ duels against arm k while having less than m samples. We denote the rounds corresponding to the first $m+1$ wins of k' against k by $r_1, \dots, r_{m+1}$. The following holds,

$$N_{k'}^{\tau}(r_{m+1}) = N_{k'}^{\tau}(r_1) + m - \sum_{s=r_1}^{r_{m+1}} \mathbb{1}(k' \in \mathcal{A}_{s-\tau+1}) \, .$$

As the number of rounds where k' wins against k is smaller than τ, we have $\sum_{s=r_1}^{r_{m+1}} \mathbb{1}(k' \in \mathcal{A}_{s-\tau+1}) \leq N_{k'}^{\tau}(r_1)$. Plugging this in the previous equation gives,

$$N_{k'}^{\tau}(r_{m+1}) \geq m \, .$$

We have the contradiction and it concludes the proof. □

Under the event $c_{k,2}^{\phi,\tau}$, the optimal arm k_ϕ^* is the leader and the diversity flag for the arm k is raised. If $D_k^\tau(r) = 1$, and k_ϕ^* is the leader, it means that the leader has not changed for $\lceil (K-1)(\log\tau)^2 \rceil$ successive rounds and has lost more than $(K-1)(\log\tau)^2$ duels. All the conditions for applying Lemma 3.15 are met. Using Lemma 3.15 and the fact that the diversity flag cannot be activated in r if it has already been activated in the last $\lceil (K-1)(\log\tau)^2 \rceil$ rounds it holds that

$$\begin{aligned}
&\mathbb{1}\left(\ell^\tau(r) = k_\phi^*, D_k^\tau(r) = 1\right) \\
&\leq \sum_{k' \neq k_\phi^*} \sum_{s=r-\lceil (K-1)(\log\tau)^2 \rceil}^{r-1} \mathbb{1}(\ell^\tau(s) = k_\phi^*, N_{k'}^\tau(s) \geq (\log\tau)^2, k' \in \mathcal{A}_{s+1}, D_{k'}^\tau(s) = 0) \, . \qquad (3.22)
\end{aligned}$$

Furthermore, we can add that an event $\{\ell^\tau(r) = k_\phi^*, N_k^\tau(s) \geq (\log\tau)^2, k \in \mathcal{A}_{s+1}, D_k^\tau(s) = 0\}$ can only be associated with at most one event $D_k^\tau(r) = 1$ for some r. Indeed, if the diversity flag is activated it cannot be anymore before at least $\lceil (K-1)(\log\tau)^2 \rceil$ rounds. Hence, combining these results we obtain

$$\begin{aligned}
&\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}(\ell^\tau(r) = k_\phi^*, D_k^\tau(r) = 1) \\
&\leq \sum_{k' \neq k_\phi^*} \sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}(k' \in \mathcal{A}_{r+1}, \ell^\tau(r) = k_\phi^*, N_{k'}^\tau(r) \geq (\log\tau)^2, D_{k'}^\tau(r) = 0) \, .
\end{aligned}$$

Applying the first equality from Lemma 3.14 with $n = (\log\tau)^2$ gives,

$$\mathbb{E}[c_{k,2}^{\phi,\tau}] \leq \sum_{k' \neq k_\phi^*} \delta_\phi(\tau+1) \frac{e^{-(\log\tau)^2\omega_{k'}}}{1 - e^{-\omega_{k'}}} \, . \qquad (3.23)$$

3.C.2.3 Upper bounding $\mathbb{E}[c_{k,3}^{\phi,\tau}]$

We recall that,

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

As for the stationary case the trickiest part is to prove that the leader is the best arm with high probability. We will first look at the terms involving the event that the best arm has already been leader after the first τ rounds of the phase, and then analyze the situation where it has never been leader. As the upper bound for $c_{k,3}^{\phi,\tau}$ is difficult to obtain, we break this section into different parts.

Part 1: *the optimal arm has been leader between $r-\tau$ and $r-1$.* If the best arm has already been leader between $r-\tau$ and $r-1$ then it has necessarily lost its leadership at some intermediate round. Loosing the leadership can be done in two different ways. The first one called the *active leadership takeover* corresponds to the case where an arm takes the leadership by winning against the leader. The second one, *passive leadership takeover* is simply the case where the leader loses so many duels that its number of samples falls below $\tau/(2K)$. We handle the first case similarly as in [Baudry et al., 2020], while for the second we use Lemma 3.15. We denote

$$\mathcal{D}(r) = \{\exists s \in [r-\tau, r-1] : \ell^\tau(s) = k_\phi^*\} .$$

We will upper bound $\mathbb{P}(\{\ell^\tau(r) \neq k_\phi^*\} \cap \mathcal{D}(r))$. We introduce,

$$\begin{aligned}\mathcal{B}(r) &:= \left\{\exists s \in [r-\tau, r-1] : \ell^\tau(s) = k_\phi^*, \ell^\tau(s+1) \neq k_\phi^*\right\} \\ &= \cup_{s=r-\tau}^{r-1} \left\{\ell^\tau(s) = k_\phi^*, \ell^\tau(s+1) \neq k_\phi^*\right\} .\end{aligned}$$

One has,

$$\mathbb{1}(\ell^\tau(r) \neq k_\phi^*, \mathcal{D}(r)) \leq \mathbb{1}(\mathcal{B}(r)) .$$

The change of leader can happen under three different scenarios: 1) some arm k takes the leadership after winning against k_ϕ^* (active takeover), 2) arm k_ϕ^* loses the leadership because its number of samples falls below the threshold $\tau/(2K)$ and 3) some arm takes the leadership after being pulled because of the diversity flag. We remark that the activation of the diversity flag for some arm k cannot lead to a leadership takeover by arm k if $(\log\tau)^2 \leq \tau/K$, so this scenario can only happen for relatively small values of τ. These properties can be formulated as

$$\begin{aligned}\left\{\ell^\tau(s) = k_\phi^*, \ell^\tau(s+1) \neq k_\phi^*\right\} \subset \cup_{k\neq k_\phi^*} &\left\{\ell^\tau(s) = k_\phi^*, \ell^\tau(s+1) = k, k \in \mathcal{A}_{s+1}, D_k^\tau(s) = 0\right\} \\ &\cup \left\{\ell^\tau(s) = k_\phi^*, N_{\ell^\tau(s)}^\tau(s+1) \leq \tau/(2K)\right\} \\ &\cup \left\{\ell^\tau(s) = k_\phi^*, \exists k \neq k_\phi^* : \ell^\tau(s+1) = k, D_k^\tau(s) = 1\right\} .\end{aligned}$$

Using this property it holds that

$$\begin{aligned}
&\sum_{r=r_{\phi}+2\tau-1}^{r_{\phi+1}-2} \mathbb{1}(\{\ell^\tau(r) \neq k^*_\phi\} \cap \mathcal{D}(r)) \\
&\leq \sum_{r=r_{\phi}+2\tau-1}^{r_{\phi+1}-2} \sum_{s=r-\tau}^{r-1} \sum_{k\neq k^*_\phi} \mathbb{1}\left(k \in \mathcal{A}_{s+1}, \ell^\tau(s) = k^*_\phi, \ell^\tau(s+1) = k, D^\tau_k(s) = 0\right) \\
&+ \sum_{r=r_{\phi}+2\tau-1}^{r_{\phi+1}-2} \sum_{s=r-\tau}^{r-1} \mathbb{1}\left(\ell^\tau(s) = k^*_\phi, N^\tau_{\ell^\tau(s)}(s+1) \leq \tau/(2K)\right) \\
&+ \sum_{r=r_{\phi}+2\tau-1}^{r_{\phi+1}-2} \sum_{s=r-\tau}^{r-1} \sum_{k\neq k^*_\phi} \mathbb{1}\left(\ell^\tau(s) = k^*_\phi, \ell^\tau(s+1) = k, D^\tau_k(s) = 1\right) .
\end{aligned}$$

We remark that if we reorganize the sums in s and r each element in the range $[r_\phi + 2\tau - 1, r_{\phi+1} - 2]$ will appear at most τ times, which leads to

$$\begin{aligned}
&\sum_{r=r_{\phi}+2\tau-1}^{r_{\phi+1}-2} \mathbb{1}(\{\ell^\tau(r) \neq k^*_\phi\} \cap \mathcal{D}(r)) \\
&\leq \underbrace{\sum_{r=r_{\phi}+2\tau-2}^{r_{\phi+1}-2} \tau \sum_{k\neq k^*_\phi} \mathbb{1}\left(\ell^\tau(r) = k^*_\phi, \ell^\tau(r+1) = k, k \in \mathcal{A}_{r+1}, D^\tau_k(r) = 0\right)}_{C_1} \\
&+ \underbrace{\sum_{r=r_{\phi}+2\tau-2}^{r_{\phi+1}-2} \tau \mathbb{1}\left(\ell^\tau(r) = k^*_\phi, N^\tau_{\ell^\tau(r)}(r+1) \leq \tau/(2K)\right)}_{C_2} \\
&+ \underbrace{\sum_{r=r_{\phi}+2\tau-1}^{r_{\phi+1}-2} \tau \sum_{k\neq k^*_\phi} \mathbb{1}\left(\ell^\tau(r) = k^*_\phi, \ell^\tau(r+1) = k, D^\tau_k(r) = 1\right)}_{C_3} .
\end{aligned}$$

We then bound the three terms separately. We can upper bound C_1 using Lemma 3.14 replacing n by the value $\tau/K - 2$,

$$\begin{aligned}
\mathbb{E}[C_1] &\leq \sum_{k\neq k^*_\phi} \tau \mathbb{E}\left[\sum_{r=r_{\phi}+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(k \in \mathcal{A}_{r+1}, \ell^\tau(r) = k^*_\phi, N^\tau_k(r) \geq \frac{\tau}{K} - 2, D^\tau_k(r) = 0\right)\right] \\
&\leq \sum_{k\neq k^*_\phi} \delta_\phi \tau(\tau+1) \frac{e^{-(\tau/K-2)\omega_k}}{1 - e^{-\omega_k}} .
\end{aligned}$$

To handle C_2 we will use Lemma 3.15. The definition of the leader ensures that when one arm takes the leadership is does it with at least τ/K observations. Hence, to make this number go below the threshold $\tau/(2K)$, k^*_ϕ has to lose at least $\tau/(2K)$ duels between the moment this arm took the leadership and the round r. There are two possibilities. The first one is that k^*_ϕ

was leader for at least τ rounds: as the index of each arms are computed from observations that have been all drawn under the leadership of k^*_ϕ then at least one arm has to beat k^*_ϕ while having more than $\tau/K - 1$ observations, which results in an active leadership takeover by this arm. Hence, a *passive* change of leader can only happen if k^*_ϕ was leader for less than τ rounds. In this case, we apply Lemma 3.15, it ensures that k^*_ϕ lost at least one duel with an arm with more than $\lfloor \frac{\tau}{2K(K-1)} \rfloor$ observations during the time it was leader. Formally,

$$\left\{ \ell^\tau(r) = k^*_\phi, N^\tau_{k^*_\phi}(r+1) \leq \frac{\tau}{2K} \right\} \subset \cup_{s=r-\tau}^{r-1} \left\{ \exists k, k \in \mathcal{A}_{s+1}, \ell^\tau(s) = k^*_\phi, N^\tau_k(s) \geq \left\lfloor \frac{\tau}{2K(K-1)} \right\rfloor \right\} .$$

We can write

$$\begin{aligned}
\mathbb{E}[C_2] &= \tau \mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}(\ell^\tau(r) = k^*_\phi, N^\tau_{k^*_\phi}(r+1) \leq \tau/(2K)) \right] \\
&\leq \tau \sum_{k \neq k^*_\phi} \mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \sum_{s=r-\tau}^{r-1} \mathbb{1}\left(k \in \mathcal{A}_{s+1}, \ell^\tau(s) = k^*_\phi, N^\tau_k(s) \geq \left\lfloor \frac{\tau}{2K(K-1)} \right\rfloor, D^\tau_k(s) = 0 \right) \right] \\
&\leq \tau^2 \sum_{k \neq k^*_\phi} \mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}\left(k \in \mathcal{A}_{r+1}, \ell^\tau(r) = k^*_\phi, N^\tau_k(r) \geq \left\lfloor \frac{\tau}{2K(K-1)} \right\rfloor, D^\tau_k(r) = 0 \right) \right] \\
&\leq \sum_{k \neq k^*_\phi} \delta_\phi \tau^2 (\tau+1) \frac{e^{-\left\lfloor \frac{\tau}{2K(K-1)} \right\rfloor \omega_k}}{1 - e^{-\omega_k}} .
\end{aligned}$$

In the second to last inequality, we have used that the terms can appear at most τ times and the last inequality result from the first inequality from Lemma 3.14.

We now focus on the term C_3. We use that $\{\ell^\tau(s+1) = k, D^\tau_k(s) = 1\}$ can happen only if $\tau/K \leq (\log \tau)^2$ because if $(\log \tau)^2 \leq \tau/K$, the activation of the diversity flag is not sufficient to take over the leadership. We recall that,

$$\mathbb{E}[C_3] = \mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \tau \sum_{k \neq k^*_\phi} \mathbb{1}\left(\ell^\tau(r) = k^*_\phi, \ell^\tau(r+1) = k, D^\tau_k(r) = 1 \right) \right] .$$

Using Equation (3.22), and letting $b = \lceil (K-1)(\log \tau)^2 \rceil$, one has

$$\begin{aligned}
\mathbb{E}[C_3] \leq \tau \sum_{k \neq k^*_\phi} \mathbb{E}\Bigg[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \sum_{k' \neq k^*_\phi} \sum_{s=r-b}^{r-1} \mathbb{1}(k' \in \mathcal{A}_{s+1}, \ell^\tau(s) = k^*_\phi, N^\tau_{k'}(s) \geq (\log \tau)^2, D^\tau_{k'}(s) = 0) \\
\times \mathbb{1}\left(\tau/K \leq (\log \tau)^2 \right) \Bigg] .
\end{aligned}$$

This can be further bounded using

$$\begin{aligned}
\mathbb{E}[C_3] \leq \tau(K-1) \sum_{k' \neq k^*_\phi} \mathbb{1}\left(\tau/K \leq (\log \tau)^2 \right) \\
\times \mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}(k' \in \mathcal{A}_{r+1}, \ell^\tau(r) = k^*_\phi, N^\tau_{k'}(r) \geq (\log \tau)^2, D^\tau_{k'}(r) = 0) \right] .
\end{aligned}$$

As $\mathbb{1}\left(\tau/K \leq (\log\tau)^2\right)$ is deterministic, we conclude by applying Lemma 3.14.

$$\mathbb{E}[C_3] \leq (K-1)\sum_{k\neq k^*_\phi} \delta_\phi \tau(\tau+1)\frac{e^{-(\log\tau)^2\omega_k}}{1-e^{-\omega_k}}\mathbb{1}\left(\tau/K \leq (\log\tau)^2\right) .$$

We then use the condition on τ to simply upper bound C_3 by

$$\mathbb{E}[C_3] \leq (K-1)\sum_{k\neq k^*_\phi} \delta_\phi \tau(\tau+1)\frac{e^{-(\tau/K)\omega_k}}{1-e^{-\omega_k}} .$$

We observe that the three terms $\mathbb{E}[C_1]$, $\mathbb{E}[C_2]$ and $\mathbb{E}[C_3]$ have very similar upper bounds, so we finally regroup them in a single term using $\left\lfloor \frac{\tau}{2K(K-1)} \right\rfloor \leq \tau/K - 2 \leq \tau/K$.

$$\mathbb{E}[C_1]+\mathbb{E}[C_2]+\mathbb{E}[C_3] \leq 3\delta_\phi\tau^2(\tau+1)(K-1)\sum_{k\neq k^*_\phi} \frac{e^{-\left\lfloor \frac{\tau}{2K(K-1)}\right\rfloor\omega_k}}{1-e^{-\omega_k}} . \tag{3.24}$$

Part 2: *the optimal arm has never been the leader after the 2τ first observations of the phase.* We now aim at upper bounding $\mathbb{E}\left[\sum_{r=r_\phi+2\tau-2}^{r_{\phi+1}-2} \mathbb{1}(\mathcal{D}(r)^c)\right]$, where $\mathcal{D}(r)^c$ is the event that k^*_ϕ has never been the leader between $r-\tau$ and $r-1$. To do so, we use that

$$\mathcal{D}(r)^c \subset \left\{ \sum_{s=r-\tau}^{r-1} \mathbb{1}\left(k^*_\phi \notin \mathcal{A}_{s+1}, \ell^\tau(s) \neq k^*_\phi\right) \geq \frac{\tau}{2} \right\} ,$$

and as in [Chan, 2020] we would like to handle this term using the Markov inequality. However, the problem in non-stationary environment is that the index of the sum is a random variable. Hence, to get back to a sum with a deterministic number of terms we introduce the set $\mathcal{R}_\phi = [r_\phi + 2\tau - 1, r_{\phi+1} - 2]$ and write

$$\begin{aligned}
\mathbb{E}\left[\sum_{r=r_\phi+2\tau-1}^{r_{\phi+1}-2} \mathbb{1}(\mathcal{D}(r)^c)\right] &= \mathbb{E}\left[\sum_{r=2\tau}^{T} \mathbb{1}(\mathcal{D}(r)^c, r\in\mathcal{R}_\phi)\right] \\
&\leq \sum_{r=2\tau}^{T} \mathbb{E}\left[\mathbb{1}(\mathcal{D}(r)^c, r\in\mathcal{R}_\phi)\right] \\
&\leq \sum_{r=2\tau}^{T} \mathbb{P}\left(\sum_{s=r-\tau}^{r-1} \mathbb{1}\left(k^*_\phi \notin \mathcal{A}_{s+1}, \ell^\tau(s) \neq k^*_\phi\right) \geq \frac{\tau}{2}, r\in\mathcal{R}_\phi\right) \\
&\leq \sum_{r=2\tau}^{T} \mathbb{P}\left(\sum_{s=r-\tau}^{r-1} \mathbb{1}(r\in\mathcal{R}_\phi)\mathbb{1}\left(k^*_\phi \notin \mathcal{A}_{s+1}, \ell^\tau(s) \neq k^*_\phi\right) \geq \frac{\tau}{2}\right) .
\end{aligned}$$

At this step we can use the Markov inequality, and obtain

$$\begin{aligned}\mathbb{E}\left[\sum_{r=r_{\phi}+2\tau-1}^{r_{\phi+1}-2}\mathbb{1}(\mathcal{D}(r)^c)\right] &\leq \sum_{r=2\tau}^{T}\frac{2}{\tau}\mathbb{E}\left[\sum_{s=r-\tau}^{r-1}\mathbb{1}(r\in\mathcal{R}_{\phi})\mathbb{1}\left(k_{\phi}^*\notin\mathcal{A}_{s+1},\ell^{\tau}(s)\neq k_{\phi}^*\right)\right]\\ &\leq \mathbb{E}\left[\sum_{r=2\tau}^{T}\mathbb{1}(r\in\mathcal{R}_{\phi})\frac{2}{\tau}\sum_{s=r-\tau}^{r-1}\mathbb{1}\left(k_{\phi}^*\notin\mathcal{A}_{s+1},\ell^{\tau}(s)\neq k_{\phi}^*\right)\right]\\ &\leq \mathbb{E}\left[\sum_{r\in\mathcal{R}_{\phi}}\frac{2}{\tau}\sum_{s=r-\tau}^{r-1}\mathbb{1}\left(k_{\phi}^*\notin\mathcal{A}_{s+1},\ell^{\tau}(s)\neq k_{\phi}^*\right)\right]\\ &= \mathbb{E}\left[\sum_{r=r_{\phi}+2\tau-1}^{r_{\phi+1}-2}\frac{2}{\tau}\sum_{s=r-\tau}^{r-1}\mathbb{1}\left(k_{\phi}^*\notin\mathcal{A}_{s+1},\ell^{\tau}(s)\neq k_{\phi}^*\right)\right].\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}\left[\sum_{r=r_{\phi}+2\tau-1}^{r_{\phi+1}-2}\mathbb{1}(\mathcal{D}(r)^c)\right] &\leq \mathbb{E}\left[\sum_{r=r_{\phi}+2\tau-1}^{r_{\phi+1}-2}\frac{2}{\tau}\sum_{s=r-\tau}^{r-1}\mathbb{1}\left(k_{\phi}^*\notin\mathcal{A}_{s+1},\ell^{\tau}(s)\neq k_{\phi}^*\right)\right]\\ &\leq D_1+D_2\,,\end{aligned}$$

where,

$$\begin{aligned}D_1 &= \mathbb{E}\left[\sum_{r=r_{\phi}+2\tau-1}^{r_{\phi+1}-2}\frac{2}{\tau}\sum_{s=r-\tau}^{r-1}\mathbb{1}\left(k_{\phi}^*\notin\mathcal{A}_{s+1},\ell^{\tau}(s)\neq k_{\phi}^*,N_{k_{\phi}^*}^{\tau}(s)\geq A_{k_{\phi}^*}^{\phi,\tau}\right)\right]\\ D_2 &= \mathbb{E}\left[\sum_{r=r_{\phi}+2\tau-1}^{r_{\phi+1}-2}\frac{2}{\tau}\sum_{s=r-\tau}^{r-1}\mathbb{1}\left(N_{k_{\phi}^*}^{\tau}(s)\leq A_{k_{\phi}^*}^{\phi,\tau}\right)\right].\end{aligned}$$

The different rounds can appear at most τ times in the double sum. Using this and the second equation of Lemma 3.14, D_1 can be upper bounded

$$D_1\leq 2\mathbb{E}\left[\sum_{r=r_{\phi}+2\tau-2}^{r_{\phi+1}-2}\mathbb{1}\left(k_{\phi}^*\notin\mathcal{A}_{r+1},\ell^{\tau}(r)\neq k_{\phi}^*,N_{k_{\phi}^*}^{\tau}(r)\geq A_{k_{\phi}^*}^{\phi,\tau}\right)\right]\leq 2\delta_{\phi}(\tau+1)\sum_{k\neq k_{\phi}^*}\frac{e^{-A_{k_{\phi}^*}^{\phi,\tau}\omega_k}}{1-e^{-\omega_k}}\,.$$

Contrarily to the stationary case, we cannot work directly with D_2 and have to further decompose $\mathbb{1}(N_{k_{\phi}^*}^{\tau}(r)\leq A_{k_{\phi}^*}^{\phi,\tau})$. Indeed, the proof in the stationary case use the sparsity of the observations of k_{ϕ}^* when it has not been pulled a lot, and the fact that in this case it has necessarily lost a lot of duel while having a fixed sample size. This is not the case in the non stationary environment, as for instance if k_{ϕ}^* has been pulled a lot in the previous window its index may change a lot. To avoid this we split the event according to the values of $N_k^{\tau}(r-\tau)$.

$$\begin{aligned}\mathbb{1}\left(N_{k_{\phi}^*}^{\tau}(r)\leq A_{k_{\phi}^*}^{\phi,\tau}\right) &\leq \mathbb{1}\left(N_{k_{\phi}^*}^{\tau}(r)\leq A_{k_{\phi}^*}^{\phi,\tau},N_{k_{\phi}^*}^{\tau}(r-\tau)>A_{k_{\phi}^*}^{\phi,\tau}\right)\\ &\quad+\mathbb{1}\left(N_{k_{\phi}^*}^{\tau}(r)\leq A_{k_{\phi}^*}^{\phi,\tau},N_{k_{\phi}^*}^{\tau}(r-\tau)\leq A_{k_{\phi}^*}^{\phi,\tau}\right).\end{aligned}$$

We then write $D_2 = 2(D_3 + D_4)$, with

$$D_3 = \mathbb{E}\left[\sum_{r=r_\phi+2\tau-1}^{r_{\phi+1}-1} \mathbb{1}\left(N^\tau_{k^*_\phi}(r) \leq A^{\phi,\tau}_{k^*_\phi}, N^\tau_{k^*_\phi}(r-\tau) > A^{\phi,\tau}_{k^*_\phi}\right)\right],$$
$$D_4 = \mathbb{E}\left[\sum_{r=r_\phi+2\tau-1}^{r_{\phi+1}-1} \mathbb{1}\left(N^\tau_{k^*_\phi}(r) \leq A^{\phi,\tau}_{k^*_\phi}, N^\tau_{k^*_\phi}(r-\tau) \leq A^{\phi,\tau}_{k^*_\phi}\right)\right].$$

D_3 can be upper bounded using Equation (3.17) in Lemma 3.14. Indeed, if $N^\tau_{k^*_\phi}(r) \leq A^{\phi,\tau}_{k^*_\phi}$ and $N^\tau_{k^*_\phi}(r-\tau,\tau) > A^{\phi,\tau}_{k^*_\phi}$, for large enough values of τ, k^*_ϕ can not be the leader and lost at least one duel against a suboptimal leader while having exactly $A^{\phi,\tau}_{k^*_\phi}$ samples between round $r-\tau$ and round $r-1$, thus

$$\left\{N^\tau_{k^*_\phi}(r) \leq A^{\phi,\tau}_{k^*_\phi}, N^\tau_{k^*_\phi}(r-\tau) > A^{\phi,\tau}_{k^*_\phi}\right\} \subset \cup_{s=r-\tau}^{r-1}\left\{k^*_\phi \notin \mathcal{A}_{s+1}, \ell^\tau(s) \neq k^*_\phi, N^\tau_{k^*_\phi}(s) = A^{\phi,\tau}_{k^*_\phi}\right\}.$$

We use the same trick as for D_1 and D_2 to handle the sums and write

$$\begin{aligned} D_3 &\leq \mathbb{E}\left[\sum_{r=r_\phi+2\tau-1}^{r_{\phi+1}-1}\sum_{s=r-\tau}^{r-1} \mathbb{1}\left(k^*_\phi \notin \mathcal{A}_{s+1}, \ell^\tau(s) \neq k^*_\phi, N^\tau_{k^*_\phi}(s) = A^{\phi,\tau}_{k^*_\phi}\right)\right] \\ &\leq \tau\mathbb{E}\left[\sum_{r=r_\phi+2\tau-1}^{r_{\phi+1}-1} \mathbb{1}\left(k^*_\phi \notin \mathcal{A}_{r+1}, \ell^\tau(r) \neq k^*_\phi, N^\tau_{k^*_\phi}(r) = A^{\phi,\tau}_{k^*_\phi}\right)\right]. \end{aligned}$$

We can directly use Lemma 3.14, however we remark that as we do not have to use an union bound on the values of $N^\tau_{k^*_\phi}$ we can remove the factor $1/(1-e^{-\omega_k})$. Hence, we finally get

$$D_3 \leq \delta_\phi\tau(\tau+1)\sum_{k\neq k^*_\phi} e^{-A^{\phi,\tau}_{k^*_\phi}\omega_k}.$$

We then handle D_4 by using the arguments introduced by [Baransi et al., 2014] with some novelty due to the sliding window. Indeed, we remark that if both $N^\tau_{k^*_\phi}(r-\tau) \leq A^{\phi,\tau}_{k^*_\phi}$ and $N^\tau_{k^*_\phi}(r) \leq A^{\phi,\tau}_{k^*_\phi}$, then k^*_ϕ competes with at most $2A^{\phi,\tau}_{k^*_\phi}$ *different index* in the entire window $[r-\tau, r-1]$. This is due to the fact that the index change only if k^*_ϕ is pulled (can happen at most $A^{\phi,\tau}_{k^*_\phi}$ times) or if k^*_ϕ loses one observation from the window $[r-2\tau, r-\tau-1]$ due to the sliding window (which can also happen at most $A^{\phi,\tau}_{k^*_\phi}$ times). Thanks to these properties we know that during the interval $[r-\tau, r-1]$ we are sure that k^*_ϕ lost at least $\tau - A^{\phi,\tau}_{k^*_\phi}$ duels, and that a fraction $1/(2A^{\phi,\tau}_{k^*_\phi})$ of them occurred while the index of k^*_ϕ remained the same.

Our objective is to highlight a property similar to the balance condition. To do so we need to identify the fraction of the duels played by k^*_ϕ with the *same index* and against *non-overlapping* blocks (i.e of mutually independent means) of any suboptimal arm $k \in \{1,\dots,K\}, k \neq k^*_\phi$. To avoid cumbersome notations we summarize the elements that allow this conclusion, first recalling the arguments of the previous paragraph:

- k^*_ϕ lost at least $\tau - A^{\phi,\tau}_{k^*_\phi}$ duels in the window $[r-\tau, r-1]$
- A fraction $1/(2A^{\phi,\tau}_{k^*_\phi})$ of them has been played with a fixed index for k^*_ϕ, i.e with the subsample mean of the same block. With a forced exploration $B(\tau) = \sqrt{\log \tau}$ this block can have any size between $\sqrt{\log \tau}$ and $A^{\phi,\tau}_{k^*_\phi}$.
- Among those duels, a fraction of at least $1/(K-1)$ of them has been played against the same suboptimal arm $k \neq k^*_\phi$.

The next step is to identify the proportion of these duels that have been played against non-overlapping blocks of k. As in the proof for the stationary case we proceed in 2 steps. First we identify the number of *different* duels (i.e the index of k is not based on the same block of observations of k) played by k^*_ϕ against k. However, thanks to the *diversity flag* we know a new duel happens after at most each $(K-1)(\log \tau)^2$ rounds. So we further process the set of duels previously identified stating that:

- A fraction of $\frac{1}{(K-1)(\log \tau)^2}$ has been played against different index of k based on different blocks of observations from the history of k, thanks to the diversity flag.
- As the blocks are of maximum size $A^{\phi,\tau}_{k^*_\phi}$ a fraction at least $1/A^{\phi,\tau}_{k^*_\phi}$ of them are *non-overlapping*.

We put all these elements together to state that there exist some $\beta \in (0,1)$ such that for any value of τ large enough, k^*_ϕ lost at least $C^\tau = \left\lfloor \frac{\beta\tau}{2(K-1)^2(\log\tau)^2 (A^{\phi,\tau}_{k^*_\phi})^2} \right\rfloor$ duels against non-overlapping blocks of some challenger k, with a fixed index. We write this event E^τ_j. Summing on all the arms, rounds, possible interval (index n) and size of the history of k^*_ϕ (index j), we obtain

$$D_4 \leq \mathbb{E}\left[\sum_{k \neq k^*_\phi} \sum_{r=r_\phi+2\tau-1}^{r_{\phi+1}-1} \sum_{n=1}^{2\left\lfloor A^{\phi,\tau}_{k^*_\phi} \right\rfloor} \sum_{j=\sqrt{\log\tau}}^{\left\lfloor A^{\phi,\tau}_{k^*_\phi} \right\rfloor} \mathbb{1}(E^\tau_j) \right] .$$

As these events do not depend on r and on n we have

$$\begin{aligned} D_4 \leq & 2\delta_\phi A^{\phi,\tau}_{k^*_\phi} \sum_{k \neq k^*_\phi} \sum_{j=\sqrt{\log\tau}}^{\left\lfloor A^{\phi,\tau}_{k^*_\phi} \right\rfloor} \mathbb{E}\left[\mathbb{1}(E^\tau_j)\right] \\ \leq & 2\delta_\phi A^{\phi,\tau}_{k^*_\phi} \sum_{k \neq k^*_\phi} \sum_{j=\sqrt{\log\tau}}^{\left\lfloor A^{\phi,\tau}_{k^*_\phi} \right\rfloor} \alpha^\phi_k(C^\tau, j) . \end{aligned}$$

Here α_k is the balance function. We index these functions by ϕ and k in order to denote the balance function between k^*_ϕ and k in the phase ϕ. We recall the definition of α_k, for any integer M

$$\alpha^\phi_k(M,j) = \mathbb{E}_{X \sim \nu^\phi_{k^*_\phi}}\left[(1 - F^\phi_{k,j}(X))^M\right] ,$$

where $\nu^{\phi}_{k',j}$ is the distribution of the sum of j random variables drawn from the distribution of an arm k' in the phase ϕ, and $F^{\phi}_{k',j}$ its cdf. We then use Lemma 3.12. We recall that this result states that for any $u \leq \mu^{\phi}_k$ it holds that

$$\alpha^{\phi}_k(C^\tau, j) \leq e^{-j\mathrm{kl}(\mu^{\phi}_k, \mu_{k^*_\phi})} u + (1-u)^{C^\tau} \ .$$

We write $\mathrm{kl}(\mu^{\phi}_k, \mu_{k^*_\phi}) = \omega^{\phi}_k$, and choose the value $u = \frac{3\log\tau}{C^\tau}$. Thanks to this choice, there exist a constant $\gamma > 1$ such that

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

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

$$\begin{aligned}
D_4 &\leq 2\delta_\phi A^{\phi,\tau}_{k^*_\phi} \sum_{k \neq k^*_\phi} \sum_{j=\sqrt{\log\tau}}^{\left\lfloor A^{\phi,\tau}_{k^*_\phi} \right\rfloor} \left[e^{-j\omega^{\phi}_k} \frac{3\log\tau}{C^\tau} + \frac{\gamma}{\tau^3} \right] \\
&\leq 2\delta_\phi A^{\phi,\tau}_{k^*_\phi} \sum_{k \neq k^*_\phi} \left[\frac{e^{-\sqrt{\log\tau}\omega^{\phi}_k}}{1 - e^{-\omega^{\phi}_k}} \frac{3\log\tau}{C^\tau} + \frac{\gamma A^{\phi,\tau}_{k^*_\phi}}{\tau^3} \right] \\
&\leq 2\delta_\phi A^{\phi,\tau}_{k^*_\phi} (K-1) \left[\frac{e^{-\sqrt{\log\tau}\omega^{\phi}}}{1 - e^{-\omega^{\phi}}} \frac{3\log\tau}{C^\tau} + \frac{\gamma A^{\phi,\tau}_{k^*_\phi}}{\tau^3} \right] \ ,
\end{aligned}$$

where $\omega^{\phi} = \min_{k \neq k^*_\phi} \omega^{\phi}_k$.

Even if these terms look impressive we explain in the next section that they are not first order terms in the regret analysis. Indeed, if we only look at the order of $A^{\phi,\tau}_{k^*_\phi}$, C^τ, we can use the same argument as in the proof of Lemma 3.7. Considering that for any integer $k > 1$, $(\log\tau)^k = o\left(e^{\sqrt{\log\tau}\omega}\right)$ we obtain that asymptotically D_4 is a $o\left(\frac{\delta_\phi}{\tau \log\tau^{k'}}\right)$ for any integer $k' \geq 1$ when $A^{\phi,\tau}_{k^*_\phi}$ is of order $(\log\tau)^u$ for some $u > 0$.

3.C.3 Summary: Upper Bound on the Dynamic Regret

Objective. Due to the many terms introduced in the analysis we provide in this section a clarification of the final terms in the regret. First of all we recall the decomposition introduced in the Section 3.4 to control the number of pulls of a suboptimal arm during a phase $\phi \in [1, \Gamma_T]$,

$$\mathbb{E}[N^{\phi}_k] \leq 2\tau + \frac{\delta_\phi A^{\phi,\tau}_k}{\tau} + \mathbb{E}[c^{\phi,\tau}_{k,1}] + \mathbb{E}[c^{\phi,\tau}_{k,2}] + \mathbb{E}[c^{\phi,\tau}_{k,3}] \ .$$

Results of Section 3.C We first provide the results we obtained in Appendix 3.C, that are true for any value of the sliding window τ and the function $A_k^{\phi,\tau}$, that we will properly calibrate later. We also recall that for any sub-optimal arm k in a phase ϕ we defined a constant ω_k^ϕ, satisfying $\omega_k^\phi = \min\left(\mathrm{kl}\left(\mu_k^\phi, \frac{1}{2}(\mu_k^\phi + \mu_{k_\phi^*}^\phi)\right), \mathrm{kl}\left(\mu_{k_\phi^*}^\phi, \frac{1}{2}(\mu_k^\phi + \mu_{k_\phi^*}^\phi)\right)\right)$.

We first obtained an upper bound on $\mathbb{E}[c_{k,1}^{\phi,\tau}]$, which controls the probability that a "concentrated" suboptimal arm k is pulled when the best arm is leader. Then we have proposed an upper bound for $\mathbb{E}[c_{k,2}^{\phi,\tau}]$ that represents the expectation of the number of pulls of the arm k because of the diversity flag when the best arm is leader. These upper bounds are

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

We then provided an upper bound of $\mathbb{E}[c_{k,3}^{\phi,\tau}]$ composed of multiple terms. This is because this term represents the expectation of the number of rounds when the best arm is not leader. To provide a general overview, this term is composed of two parts: the first one for the cases when the best arm *has already* been leader in the last τ rounds, and the case when the best arm *has never* been leader in the last τ round. The first general scenario was handled by the constants C_1, C_2 and C_3, that we have upper bounded in expectation by,

$$\mathbb{E}\left[C_1+C_2+C_3\right] \leq 3\delta_\phi\tau^2(\tau+1)(K-1)\sum_{k'\neq k_\phi^*}\frac{e^{-\left\lfloor\frac{\tau}{2K(K-1)}\omega_{k'}^\phi\right\rfloor}}{1-e^{-\omega_{k'}^\phi}}.$$

We observe that this term has a larger order in τ than the previous one before the exponential, but as a larger term in the exponential that compensates. After that, we handled the cases when the best arm has never been leader. We distinguish again different cases. The terms D_1 and D_3 provide terms that share similar order with the ones we obtained before, namely:

$$D_1 \leq 2\delta_\phi(\tau+1)\sum_{k'\neq k_\phi^*}\frac{e^{-A_{k'}^{\phi,\tau}\omega_{k'}^\phi}}{1-e^{-\omega_{k'}^\phi}} \quad \text{and} \quad D_3 \leq \delta_\phi\tau(\tau+1)e^{-A_{k_\phi^*}^{\phi,\tau}\omega_k}.$$

The last term is the one that corresponds to the balance condition in the stationary case. Yet, its adaptation to the non-stationary case was not trivial. We obtained

$$D_4 \leq 2\delta_\phi A_{k_\phi^*}^{\phi,\tau}(K-1)\left[\frac{e^{-\sqrt{\log\tau}\omega^\phi}}{1-e^{-\omega^\phi}}\frac{3\log\tau}{C^\tau} + \frac{\gamma A_{k_\phi^*}^{\phi,\tau}}{\tau^3}\right],$$

where $C^\tau = \left\lfloor\frac{\beta\tau}{2(K-1)^2(\log\tau)^2(A_{k_\phi^*}^{\phi,\tau})^2}\right\rfloor$ and $\omega^\phi = \min_{k\neq k_\phi^*}\omega_k^\phi$.

Tuning of the parameters The previous results allow to control precisely the dynamic regret of SW-LB-SDA for general values of τ and the constants of the problem. We first remark that one could tune each of the constants $A_{k_\phi^*}^{\phi,\tau}$ to optimize the term in each phase. However, in this paragraph we propose a more general asymptotic analysis that proves that an optimal tuning of τ allows the algorithm to reach optimal guarantees. To catch this generality we will simply

define $A^{\phi,\tau}_{k^*_\phi} = A(\tau) = B\log\tau$ for some constant B, and define $\omega = \min_{\phi\in[1,\Gamma_T]}\{\min_{k\neq k^*_\phi}\omega^\phi_k\}$. With these new definitions, we can group several terms together, and obtain for $\tau > K$

$$\mathbb{E}[N^\phi_k] \leq 2\tau + \frac{\delta_\phi A(\tau)}{\tau} + \frac{2\delta_\phi(\tau+1)K}{1-e^{-\omega}}e^{-A(\tau)\omega} + \frac{K\delta_\phi\tau(\tau+1)}{1-e^{-\omega}}e^{-(\log\tau)^2\omega} + 3\delta_\phi\tau^2(\tau+1)(K-1)^2\frac{e^{-\left\lfloor\frac{\tau}{2K(K-1)}\omega\right\rfloor}}{1-e^{-\omega}} + 2\delta_\phi A(\tau)(K-1)\left[\frac{e^{-\sqrt{\log\tau}\omega}}{1-e^{-\omega}}\frac{3\log\tau}{C^\tau} + \frac{\gamma A(\tau)}{\tau^3}\right].$$

As the only term that depends on the phase is δ_ϕ it is now straightforward to sum on the phases and the arms to obtain the dynamic regret, recalling that $\sum_{\phi=1}^{\Gamma_T}\delta_\phi = T$. Without loss of generality, we also assume that for all ϕ and for all $k \neq k^*_\phi$, $\Delta^\phi_k \leq 1$.

$$\begin{aligned}
\mathcal{R}(T) &= \sum_{\phi=1}^{\Gamma_T}\sum_{k\neq k^*_\phi}\mathbb{E}[N^\phi_k]\Delta^\phi_k \\
&\leq \underbrace{2(K-1)\tau\Gamma_T + \frac{(K-1)TA(\tau)}{\tau}}_{E_1} + \underbrace{\frac{2T(\tau+1)K(K-1)}{1-e^{-\omega}}e^{-A(\tau)\omega}}_{E_2} \\
&\quad + \underbrace{\frac{TK(K-1)\tau(\tau+1)}{1-e^{-\omega}}e^{-(\log\tau)^2\omega}}_{E_3} + \underbrace{\frac{3T(K-1)\tau^2(\tau+1)(K-1)^2}{1-e^{-\omega}}e^{-\left\lfloor\frac{\tau}{2K(K-1)}\omega\right\rfloor}}_{E_4} \\
&\quad + \underbrace{2TA(\tau)(K-1)^2\left[\frac{e^{-\sqrt{\log\tau}\omega}}{1-e^{-\omega}}\frac{3\log\tau}{C^\tau} + (K-1)\frac{\gamma A(\tau)}{\tau^3}\right]}_{E_5}
\end{aligned}$$

Knowing the horizon T and an order of the number of breakpoints Γ_T we propose a tuning for τ in $\sqrt{\frac{T\log T}{\Gamma_T}}$. We then prove that the only first order terms in the decomposition are the terms in E_1.

First, as $\log\tau$ is of order $\log T$, choosing $A(\tau) = \frac{6}{\omega}\log\tau$ ensures that E_2 is upper bounded by a constant. Then, the terms E_3 and E_4 are also both upper bounded by constants as the term in the exponent dominates the polynomial in τ preceding it. The term E_5 needs more work. Indeed, its second component causes no difficulty and is upper bounded by a constant. However, for the first term we need to use the fact C^τ is of order $\tau/\log(\tau)^j$, hence there exists some integer j' such that the dominant term in E_5 is of order $\frac{T}{\tau}\times(\log\tau)^{j'}e^{-\sqrt{\log\tau}\omega}$. As in Appendix 3.A we use that $(\log\tau)^{j'}e^{-\sqrt{\log\tau}\omega} = o(\log(\tau)^{-1})$ (for instance). Hence, thanks to the log terms E_5 is of lower order than E_1. Finally, we obtain

$$\mathcal{R}(T) = O(\sqrt{T\Gamma_T\log T}).$$

This concludes the proof of Theorem 3.11.

4 | Weighted Linear Bandits for Non-Stationary Environments

In the previous chapter, we discussed an alternative to upper-confidence bound based or Thompson Sampling algorithms for the stochastic multi-armed bandit problem. From now on, we consider richer structured models. This chapter is dedicated to the study of the linear contextual bandit problem in which the available actions correspond to arbitrary context vectors whose associated rewards follow a *non-stationary* linear regression model. In this setting, the unknown regression parameter is allowed to vary in time. To address this problem, we propose a novel optimistic algorithm based on discounted linear regression D-LinUCB, where exponential weights are used to smoothly forget the past. This involves studying the deviations of the sequential weighted least-squares estimator under generic assumptions. As a by-product, we obtain novel deviation results that can be used beyond non-stationary environments. We provide theoretical guarantees on the behavior of D-LinUCB in both slowly-varying and abruptly-changing environments. The results from this chapter are based on [Russac et al., 2019].

Outline

4.1 Introduction

In this chapter we are interested in structured bandit models, known as stochastic linear bandits, in which linear regression is used to predict rewards [Abbasi-Yadkori et al., 2011, Auer, 2002, Li et al., 2010].

Our first contribution consists in extending existing deviation inequalities to sequential weighted least-squares. Our result applies to a large variety of bandit problems and is of independent interest. In particular, it extends the recent analysis of heteroscedastic environments by [Kirschner and Krause, 2018]. It can also be useful to deal with class imbalance situations, or, as we focus on here, in non-stationary environments.

As a second major contribution, we apply our results to propose D-LinUCB, an adaptive linear bandit algorithm based on carefully designed exponential weights. D-LinUCB can be implemented fully recursively (without requiring the storage of past actions) with a numerical complexity that is comparable to that of LinUCB. To characterize the performance of the algorithm, we provide a unified regret analysis for abruptly-changing or slowly-varying environments.

The setting and notations are presented below and we state our main deviation result in Section 4.2. Section 4.3 is dedicated to non-stationary linear bandits: we describe our algorithms and provide regret upper bounds in abruptly-changing and slowly-varying environments. We complete this theoretical study with a set of experiments in Section 4.4.

4.1.1 Model and Notations

The setting we consider in this chapter is a non-stationary variant of the stochastic linear bandit problem studied in [Abbasi-Yadkori et al., 2011, Li et al., 2010] and presented in Section 1.4.2. We recall the main ingredients here. At each round $t \geq 1$, the learner receives a set of feasible actions $\mathcal{A}_t \subset \mathbb{R}^d$ and chooses an action $A_t \in \mathcal{A}_t$. Based on this choice, the learner receives a reward X_t satisfying Equation (1.20), i.e

$$X_t = \langle A_t, \theta_t^\star \rangle + \eta_t \ ,$$

where $\theta_t^\star \in \mathbb{R}^d$ is an unknown parameter and η_t is, conditionally on the past, a $\sigma-$subgaussian random noise.

The action set $\mathcal{A}_t$ may be arbitrary but its components are assumed to be bounded, in the sense that $\|a\|_2 \leq L$, $\forall a \in \mathcal{A}_t$. The time-varying parameter is also assumed to be bounded: $\forall t, \|\theta_t^\star\|_2 \leq S$. We further assume that $|\langle a, \theta_t^\star \rangle| \leq 1$, $\forall t, \forall a \in \mathcal{A}_t$, (obviously, this could be guaranteed by assuming that $L = S = 1$, but we indicate the dependence in L and S in order to facilitate the interpretation of some results). For a positive semi definite matrix M and a vector x, we denote by $\|x\|_M$ the norm $\sqrt{x^\top M x}$. The goal of the learner is to build a policy π that minimizes the *dynamic regret* defined as

$$R(T, \pi) = \sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} \langle a - A_t, \theta_t^\star \rangle \ . \tag{4.1}$$

Even in the stationary case (i.e. when $\theta_t^\star = \theta^\star$), there is, in general, no single fixed best action in this model.

When making stronger structural assumption on $\mathcal{A}_t$, one recovers specific instances that have also been studied in the literature. In particular, the canonical basis of $\mathbb{R}^d$, $\mathcal{A}_t = \{e_1, \ldots, e_d\}$,

yields the familiar non contextual multi-armed bandit model [Lattimore and Szepesvári, 2020]. Another variant, studied by [Goldenshluger and Zeevi, 2013] and others, is obtained when $\mathcal{A}_t = \{e_1 \otimes a_t, \dots, e_k \otimes a_t\}$, where $\otimes$ denotes the Kronecker product and a_t is a time-varying context vector shared by the k actions.

4.1.2 Related Work

In recent years, linear bandits have become the go-to paradigm to balance exploration and exploitation in contextual sequential decision making problems. Linear bandits have typically found applications for content-based recommendations [Li et al., 2010, Valko et al., 2014], real-time bidding [Flajolet and Jaillet, 2017] and even mobile-health interventions [Tewari and Murphy, 2017]. In these application, non-stationary often plays a crucial role. In this chapter, we focus on non-stationary environments. For the sake of conciseness, we restrict the discussion to works that consider specifically the stochastic linear bandit model from Equation (1.20), including its restriction to the simpler (non-stationnary) multi-armed bandit model. Note that there is also a rich line of works that consider possibly non-linear contextual models in the case where one can make probabilistic assumptions on the contexts [Chen et al., 2019, Luo et al., 2018].

Controlling the regret with respect to the non-stationary optimal action defined in Equation (4.1) depends on the assumptions that are made on the time-variations of $\theta_t^\star$. A generic way of quantifying them is through a *variation bound* $\mathcal{B}_T = \sum_{s=1}^{T-1} \|\theta_s^\star - \theta_{s+1}^\star\|_2$ [Besbes et al., 2014, Besbes et al., 2018, Cheung et al., 2019], similar to the penalty used in the group fused Lasso [Bleakley and Vert, 2011]. The main advantage of using the variation budget is that it includes both *slowly-varying* and *abruptly-changing* environments. Assuming that an upper-bound B_T for the variation bound is known, [Besbes et al., 2014, Besbes et al., 2015, Besbes et al., 2018] achieve the tight dynamic regret bound of $\widetilde{\mathcal{O}}(K^{1/3}B_T^{1/3}T^{2/3})$ for the $K-$armed bandits. For linear bandits, [Cheung et al., 2019, Cheung et al., 2021] propose an algorithm based on the use of a sliding-window and provide a $\widetilde{\mathcal{O}}(d^{2/3}B_T^{1/3}T^{2/3})$ dynamic regret bound; since this contribution is close to ours, we discuss it further in Section 4.3.2. [Zhao et al., 2020] suggest to restart the algorithm every H steps. H is tuned based on the knowledge of B_T.

A more specific non-stationary setting arises when the number of changes in the parameter is bounded by Γ_T, as in traditional change-point models. The problem is usually referred to as *switching bandits* or *abruptly-changing* environments. It is, for instance, the setting considered in the work by [Garivier and Moulines, 2011], who analyzed the dynamic regret of UCB strategies based on either a sliding-window or exponential discounting. For both policies, they prove upper bounds on the regret in $\widetilde{\mathcal{O}}(\sqrt{\Gamma_T T})$ when Γ_T is known. They also provide a lower bound in a specific non-stationary setting, showing that $R(T) = \Omega(\sqrt{T})$. We refer the reader to Section 1.3.1 for a better intuition on this lower bound. The algorithm ideas from [Garivier and Moulines, 2011] can be traced back to [Kocsis and Szepesvári, 2006b]. [Wei and Srivatsva, 2018] shows that an horizon-independent version of the sliding window algorithm can also be analyzed in a slowly-varying setting. [Keskin and Zeevi, 2017] analyze windowing and discounting approaches to address dynamic pricing guided by a (time-varying) linear regression model. Discount factors have also been used with Thomson sampling in dynamic environments as in [Gupta et al., 2011, Raj and Kalyani, 2017].

In abruptly-changing environments, the alternative approach relies on change-point detection [Auer et al., 2019, Besson et al., 2020, Cao et al., 2019, Wu et al., 2018, Yu and Mannor, 2009]. A bound on the regret in $\mathcal{O}((\frac{1}{\epsilon^2} + \frac{1}{\Delta})\log(T))$ is proven by [Yu and Mannor, 2009], where ϵ is the

smallest gap that can be detected by the algorithm, which had to be given as prior knowledge. [Cao et al., 2019] proves a minimax bound in $\mathcal{O}(\sqrt{\Gamma_T KT})$ if Γ_T is known. [Besson et al., 2020] achieves a rate of $\mathcal{O}(\sqrt{\Gamma_T KT})$ without any prior knowledge of the gaps or Γ_T. In the contextual case, [Wu et al., 2018] builds on the same idea: they use a pool of LinUCB learners called *slave models* as experts and they add a new model when no existing slave is able to give good prediction, that is, when a change is detected. A limitation however of such an approach is that it can not adapt to some slowly-varying environments, as will be illustrated in Section 4.4. From a practical viewpoint, the methods based either on sliding window or change-point detection require the storage of past actions whereas those based on discount factors can be implemented fully recursively.

Finally, non-stationarity may also arise in more specific scenarios connected, for instance, to the decaying attention of the users, as investigated in [Levine et al., 2017, Mintz et al., 2020, Seznec et al., 2019]. In the following, we consider the general case where the parameters satisfy the variation bound, i.e., $\mathcal{B}_T = \sum_{t=1}^{T-1} \|\theta_t^\star - \theta_{t+1}^\star\|_2 \leq B_T$ and we propose an algorithm based on discounted linear regression.

4.2 Confidence Bounds for Weighted Linear Bandits

In this section, we consider the concentration of the weighted regularized least-squares estimator, when used with general weights and regularization parameters. To the best of our knowledge there is no such results in the literature for sequential learning —i.e., when the current regressor may depend on the random outcomes observed in the past. The particular case considered in Lemma 5 of [Kirschner and Krause, 2018] (heteroscedastic noise with optimal weights) stays very close to the unweighted case and we show below how to extend this result. We believe that this new bound is of interest beyond the specific model considered in this chapter. For the sake of clarity, we first focus on the case of regression models with fixed parameter, where $\theta_t^\star = \theta^\star$, for all t.

First consider a deterministic sequence of regularization parameters $(\lambda_t)_{t\geq 1}$. The reason why these should be non-constant for weighted least-squares will appear clearly in Section 4.3. Next, define by $\mathcal{F}_t = \sigma(\mathcal{A}_1, A_1, X_1, \ldots, A_t, X_t, \mathcal{A}_{t+1}, A_{t+1})$ the filtration from Section 1.4.1.2. Using this filtration the actions A_t are predictable, that is, they are $\mathcal{F}_{t-1}$ measurable. We also assume that the positive weights $(w_t)_t$ are predictable. Defining by

$$\hat{\theta}_t = \mathrm{argmin}_{\theta\in\mathbb{R}^d} \left(\sum_{s=1}^{t} w_s (X_s - \langle A_s, \theta\rangle)^2 + \lambda_t \|\theta\|_2^2 \right) ,$$

the regularized weighted least-squares estimator of $\theta^\star$ at time t, one has

$$\hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} w_s A_s X_s \quad \text{where} \quad V_t = \sum_{s=1}^{t} w_s A_s A_s^\top + \lambda_t I_d \, , \tag{4.2}$$

and I_d denotes the d-dimensional identity matrix. We further consider an arbitrary sequence of positive parameters $(\mu_t)_{t\geq 1}$ and define the matrix

$$\widetilde{V}_t = \sum_{s=1}^{t} w_s^2 A_s A_s^\top + \mu_t I_d \, . \tag{4.3}$$

$\widetilde{V}_t$ is strongly connected to the variance of the estimator $\hat{\theta}_t$, which involves the squares of the weights $(w_s^2)_{s\geq 1}$. For the time being, μ_t is arbitrary and will be set as a function of λ_t in order to optimize the deviation inequality.

We now establish the following maximal deviation inequality.

Theorem 4.1. *For any $\mathcal{F}_t$-predictable sequences of actions $(A_t)_{t\geq 1}$ and positive weights $(w_t)_{t\geq 1}$ and for all $\delta > 0$,*

$$\mathbb{P}\left(\forall t, \|\hat{\theta}_t - \theta^\star\|_{V_t\widetilde{V}_t^{-1}V_t} \leq \frac{\lambda_t}{\sqrt{\mu_t}}S + \sigma\sqrt{2\log(1/\delta) + d\log\left(1 + \frac{L^2\sum_{s=1}^t w_s^2}{d\mu_t}\right)}\right) \geq 1-\delta\,.$$

Proof. We define the quantity $S_t = \sum_{s=1}^t w_s A_s \eta_s$. First note that,

$$\begin{aligned}\hat{\theta}_t &= V_t^{-1}\sum_{s=1}^t w_s A_s X_s = V_t^{-1}\sum_{s=1}^t w_s A_s(A_s^\top\theta^\star + \eta_s)\\ &= V_t^{-1}\left(\sum_{s=1}^t w_s A_s A_s^\top\theta^\star + \lambda_t\theta^\star - \lambda_t\theta^\star\right) + V_t^{-1}S_t = \theta^\star - \lambda_t V_t^{-1}\theta^\star + V_t^{-1}S_t\,.\end{aligned}$$

Thus,

$$\hat{\theta}_t - \theta^\star = V_t^{-1}S_t - \lambda_t V_t^{-1}\theta^\star\,. \tag{4.4}$$

$\forall x\in\mathbb{R}^d, \forall t>0$, we have

$$\begin{aligned}|x^\top(\hat{\theta}_t - \theta^\star)| &\leq \|x\|_{V_t^{-1}\widetilde{V}_t V_t^{-1}}\left(\|V_t^{-1}S_t\|_{V_t\widetilde{V}_t^{-1}V_t} + \|\lambda_t V_t^{-1}\theta^\star\|_{V_t\widetilde{V}_t^{-1}V_t}\right)\\ &\leq \|x\|_{V_t^{-1}\widetilde{V}_t V_t^{-1}}\left(\|S_t\|_{\widetilde{V}_t^{-1}} + \lambda_t\|\theta^\star\|_{\widetilde{V}_t^{-1}}\right)\,.\end{aligned}$$

By applying the previous inequality with $x = V_t\widetilde{V}_t^{-1}V_t(\hat{\theta}_t - \theta^\star)$, we have

$$\forall t,\ \|\hat{\theta}_t - \theta^\star\|_{V_t\widetilde{V}_t^{-1}V_t} \leq \|S_t\|_{\widetilde{V}_t^{-1}} + \lambda_t\|\theta^\star\|_{\widetilde{V}_t^{-1}}\,.$$

Knowing that $\widetilde{V}_t \geq \mu_t I_d$ and that $\widetilde{V}_t$ is positive definite, we have $\|\theta^\star\|_{\widetilde{V}_t^{-1}} \leq \frac{1}{\sqrt{\mu_t}}\|\theta^\star\|_2$. Finally,

$$\forall t, \|\hat{\theta}_t - \theta^\star\|_{V_t\widetilde{V}_t^{-1}V_t} \leq \|S_t\|_{\widetilde{V}_t^{-1}} + \frac{\lambda_t}{\sqrt{\mu_t}}\|\theta^\star\|_2\,. \tag{4.5}$$

The result is then obtained using Proposition 4.8, where the following any time high probability upper bound for $\|S_t\|_{\widetilde{V}_t^{-1}}$ is established,

$$\mathbb{P}\left(\forall t\geq 0, \|S_t\|_{\widetilde{V}_t^{-1}} \leq \sigma\sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\widetilde{V}_t)}{\mu_t^d}\right)}\right) \geq 1-\delta\,.$$

Therefore by using inequality 4.5,

$$\mathbb{P}\left(\forall t\geq 0, \|\hat{\theta}_t - \theta^\star\|_{\widetilde{V}_t^{-1}} \leq \frac{\lambda_t}{\sqrt{\mu_t}}S + \sigma\sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\widetilde{V}_t)}{\mu_t^d}\right)}\right) \geq 1-\delta\,.$$

We obtain the exact formula of Theorem 4.1 by upper bounding $\det(\widetilde{V}_t)$ as proposed in Proposition 4.9 □

The standard result used for least-squares [Lattimore and Szepesvári, 2020, Chapter 20] is recovered by taking $\mu_t = \lambda_t$ and $w_t = 1$ (note that $\widetilde{V}_t$ is then equal to V_t). When the weights are not equal to 1, the appearance of the matrix $\widetilde{V}_t$ is a consequence of the fact that the variance terms are proportional to the squared weights w_t^2, while the least-squares estimator itself is defined with the weights w_t. In the weighted case, the matrix $V_t\widetilde{V}_t^{-1}V_t$ must be used to define the confidence ellipsoid.

An important property of the least-squares estimator is to be scale-invariant, in the sense that multiplying all weights $(w_s)_{1\leq s\leq t-1}$ and the regularization parameter λ_t by a constant leaves the estimator $\hat{\theta}_t$ unchanged. In Theorem 4.1, the only choice of sequence $(\mu_t)_{t\geq 1}$ that is compatible with this scale-invariance property is to take μ_t proportional to λ_t^2: then the matrix $V_t\widetilde{V}_t^{-1}V_t$ becomes scale-invariant (i.e. unchanged by the transformation $w_s \mapsto \alpha w_s$) and so does the upper bound of $\|\hat{\theta}_t - \theta^\star\|_{V_t\widetilde{V}_t^{-1}V_t}$ in Theorem 4.1. In the following, we will stick to this choice, while particularizing the choice of the weights w_t to allow for non-stationary models.

It is possible to extend this result to heteroscedastic noise, when η_t is σ_t sub-Gaussian and σ_t is $\mathcal{F}_{t-1}$ measurable, by defining $\widetilde{V}_t$ as $\sum_{s=1}^t w_s^2\sigma_s^2 A_s A_s^\top + \mu_t I_d$. In the next section, we will also use an extension of Theorem 4.1 to the non-stationary linear model. In this case, Theorem 4.1 holds with $\theta^\star$ replaced by $V_t^{-1}(\sum_{s=1}^t w_s A_s A_s^\top \theta_s^\star + \lambda_t \theta_r^\star)$, where r is an arbitrary time index (proposition 4.12 in Appendix). The fact that r can be chosen freely is a consequence of the assumption that the sequence of ℓ_2-norms of the parameters $(\theta_t^\star)_{t\geq 1}$ is bounded by S.

4.3 Application to Non-stationary Linear Bandits

In this section, we consider the non-stationary linear bandit model from Section 4.1.1 and propose a bandit algorithm in Section 4.3.1, called Discounted Linear Upper Confidence Bound (D-LinUCB), that relies on weighted least-squares to adapt to changes in the parameters $\theta_t^\star$. Analyzing the performance of D-LinUCB in Section 4.3.2, we show that it achieves reliable performance both for abruptly changing or slowly drifting parameters.

4.3.1 The D-LinUCB Algorithm

Being adaptive to parameter changes indeed implies to reduce the influence of observations that are far back in the past, which suggests using weights w_t that increase with time. In doing so, there are two important caveats to consider. First, this can only be effective if the sequence of weights is growing sufficiently fast (see the analysis in the next section). We thus consider exponentially increasing weights of the form $w_t = \gamma^{-t}$, where $0 < \gamma < 1$ is the discount factor.

Next, due to the absence of assumptions on the action sets $\mathcal{A}_t$, the regularization is instrumental in obtaining guarantees of the form given in Theorem 4.1. In fact, if $w_t = \gamma^{-t}$ while λ_t does not increase sufficiently fast (and hence μ_t), then the term $\log(1 + (L^2\sum_{s=1}^t w_s^2)/(d\mu_t))$ will eventually dominate the radius of the confidence region since we choose μ_t proportional to λ_t^2. This occurs because there is no guarantee that the algorithm will persistently select actions A_t that span the entire space. With this in mind, we consider an increasing regularization factor of the form $\lambda_t = \gamma^{-t}\lambda$, where $\lambda > 0$ is a hyperparameter.

Note that due to the scale-invariance property of the weighted least-square estimator, we can equivalently consider that at time t, we are given *time-dependent* weights $w_{t,s} = \gamma^{t-s}$, for $1 \leq s \leq t$ and that $\hat{\theta}_t$ is defined as

$$\text{argmin}_{\theta\in\mathbb{R}^d}\left(\sum_{s=1}^{t}\gamma^{t-s}(X_s-\langle A_s,\theta\rangle)^2+\lambda\|\theta\|_2^2\right).$$

For numerical stability reasons, this form is preferable and is used in the statement of Algorithm 10. In the analysis of Section 4.3.2 however we revert to the standard form of the weights, which is required to apply the concentration result of Section 4.1. We are now ready to describe D-LinUCB in Algorithm 10.

Input: Failure probability δ, subgaussianity constant σ, dimension d, regularization λ, upper bound for actions L, upper bound for parameters S, discount factor γ.
Initialization: $b=0_{\mathbb{R}^d}$, $V=\lambda I_d$, $\widetilde{V}=\lambda I_d$, $\hat{\theta}=0_{\mathbb{R}^d}$
for $t\geq 1$ **do**
  Receive $\mathcal{A}_t$, compute $\beta_{t-1}=\sqrt{\lambda}S+\sigma\sqrt{2\log\left(\frac{1}{\delta}\right)+d\log\left(1+\frac{L^2(1-\gamma^{2(t-1)})}{\lambda d(1-\gamma^2)}\right)}$
  for $a\in\mathcal{A}_t$ **do**
    Compute $\text{UCB}(a)=a^\top\hat{\theta}+\beta_{t-1}\sqrt{a^\top V^{-1}\widetilde{V}V^{-1}a}$
  $A_t=\text{argmax}_{a\in\mathcal{A}_t}\text{UCB}(a)$
  Play action A_t
  Receive reward X_t
  Updating phase: $V=\gamma V+A_tA_t^\top+(1-\gamma)\lambda I_d$, $\widetilde{V}=\gamma^2\widetilde{V}+A_tA_t^\top+(1-\gamma^2)\lambda I_d$, $b=\gamma b+X_tA_t$, $\hat{\theta}=V^{-1}b$

Algorithm 10: D-LinUCB

4.3.2 Analysis

As discussed previously, we consider weights of the form $w_t=\gamma^{-t}$ (where $0<\gamma<1$) in the D-LinUCB algorithm. In accordance with the discussion at the end of Section 4.1, Algorithm 10 uses $\mu_t=\gamma^{-2t}\lambda$ as the parameter to define the confidence ellipsoid around $\hat{\theta}_{t-1}$. The confidence ellipsoid $\mathcal{C}_t$ is defined as $\left\{\theta:\|\theta-\hat{\theta}_{t-1}\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}}\leq\beta_{t-1}\right\}$ where

$$\beta_t=\sqrt{\lambda}S+\sigma\sqrt{2\log(1/\delta)+d\log\left(1+\frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)}\right)}. \tag{4.6}$$

Using standard algebraic calculations together with the remark above about scale-invariance it is easily checked that at time t, Algorithm 10 selects the action A_t that maximizes $\langle a,\theta\rangle$ for $a\in\mathcal{A}_t$ and $\theta\in\mathcal{C}_t$.

A strong conceptual advantage (at least from an analysis point of view) of forgetting strategies is that it allows for a natural decoupling of the *learning* and *tracking* aspects of non-stationary bandit problems. At each round t, the learning aspect is rooted in the noisy nature of the environment, which blurs the sequence of $\{\theta_s^\star\}_{s=1}^t$ that generated observed rewards. The learning guarantees of forgetting policies can be extended from existing stationary analyses, this is what we obtain with the expression of $\mathcal{C}_t$ and Equation 4.6.

On the other hand, the tracking aspect is inherited from the drift of $\theta_{t-1}^\star$ to $\theta_t^\star$ which induces an incompressible estimation error. It is therefore fundamentally tied to the variation-budget defined by $\mathcal{B}_T=\sum_{t=1}^{T-1}\|\theta_t^\star-\theta_{t+1}^\star\|_2$, which is an off-policy metric (*i.e* independent of the trajectory

that was played) characterized by the ℓ_2 norm. Both aspects are conflicting sources of regret; reaching optimality requires finding the correct balance between the two of them.

Using forgetting mechanisms is helpful for controlling the bias. When considering the ordinary least squares estimator (hence without forgetting) [Luo et al., 2021, Figure 1] build a simple example in 2 dimensional space without noise where the bias is large even when the true parameters at two consecutive times $\theta_1^\star$ and $\theta_2^\star$ are ϵ close to each other.

We now show how to isolate bias (related to the tracking aspect of the problem) and variance (related to the learning) terms. In contrast with the stationary case, the confidence ellipsoid $\mathcal{C}_t$ does not necessarily contain (with high probability) the actual parameter value $\theta_t^\star$ due to the (unknown) bias arising from the time variations of the parameter. We thus define

$$\bar{\theta}_t = V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{-s}A_sA_s^\top\theta_s^\star + \lambda\gamma^{-(t-1)}\theta_t^\star\right) ,$$

which is an action-dependent analogue of the parameter value $\theta^\star$ in the stationary setting (although this is a random value). As mentioned in section 4.2, $\bar{\theta}_t$ does belong to $\mathcal{C}_t$ with probability at least $1-\delta$ (see Proposition 4.12 in Appendix).

Let $A_t^\star = \operatorname{argmax}_{a\in\mathcal{A}_t}\langle a, \theta_t^\star\rangle$ and $\theta_t = \operatorname{argmax}_{\theta\in\mathcal{C}_t}\langle A_t, \theta\rangle$. The instantaneous regret r_t satisfies,

$$\begin{aligned} r_t &\coloneqq \max_{a\in\mathcal{A}_t}\langle a, \theta_t^\star\rangle - \langle A_t, \theta_t^\star\rangle = \langle A_t^\star - A_t, \theta_t^\star\rangle \\ &= \langle A_t^\star - A_t, \bar{\theta}_t\rangle + \langle A_t^\star - A_t, \theta_t^\star - \bar{\theta}_t\rangle . \end{aligned} \tag{4.7}$$

Under the event $\{\forall t > 0,\, \bar{\theta}_t \in \mathcal{C}_t\}$, that occurs with probability at least $1-\delta$ thanks to Proposition 4.12, we have,

$$\langle A_t^\star, \bar{\theta}_t\rangle \le \operatorname{argmax}_{\theta\in\mathcal{C}_t}\langle A_t^\star, \theta\rangle = \mathrm{UCB}_t(A_t^\star) \le \mathrm{UCB}_t(A_t) = \operatorname{argmax}_{\theta\in\mathcal{C}_t}\langle A_t, \theta\rangle = \langle A_t, \theta_t\rangle . \tag{4.8}$$

Then, with probability at least $1-\delta$, $\forall t > 0$,

$$\begin{aligned} r_t &\le \langle A_t, \theta_t - \bar{\theta}_t\rangle + \langle A_t^\star - A_t, \theta_t^\star - \bar{\theta}_t\rangle \\ &\le \|A_t\|_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\|\theta_t - \bar{\theta}_t\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} + \|A_t^\star - A_t\|_2\|\theta_t^\star - \bar{\theta}_t\|_2 \quad \text{(Cauchy-Schwarz)} \\ &\le \|A_t\|_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\|\theta_t - \bar{\theta}_t\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} + 2L\|\theta_t^\star - \bar{\theta}_t\|_2 \quad (\forall a\in\mathcal{A}_t,\ \|a\|_2 \le L) . \end{aligned}$$

The two terms are upper bounded using different techniques. The first term is handled with the equivalent in a non-stationary environment of the deviation inequality of Theorem 4.1 and the second term is the equivalent of the bias. We now explain how to control $\|\theta_t - \bar{\theta}_t\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}}$ which is related to the learning aspect of the problem. We have,

$$\|\theta_t - \bar{\theta}_t\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} \le \|\theta_t - \hat{\theta}_{t-1}\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} + \|\bar{\theta}_t - \hat{\theta}_{t-1}\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} \le 2\beta_{t-1} ,$$

where the last inequality holds because under our assumption $\bar{\theta}_t \in \mathcal{C}_t$ with high probability and by definition $\theta_t \in \mathcal{C}_t$.

4.3.2.1 An Error in the Control of the Bias Term

The tracking error can only be observed (at least at analysis time) in the directions that were actually played by the algorithm and for which rewards were collected. Henceforth, the main challenge when controlling the tracking error lies in converting its on-policy version to its off-policy counterpart (which is B_T). This is where current approaches make a mistake by claiming that this can be done at no cost on the regret.

The bias term can be bounded deterministically, from the assumption made on $\sum_{s=1}^{T-1}\|\theta_s^\star - \theta_{s+1}^\star\|_2$. In doing so, we introduce the analysis parameter D that, roughly speaking, corresponds to the window length equivalent to a particular choice of discount factor γ: the bias resulting from observations that are less than D time steps apart may be bounded in term of D. Let $D \in \mathbb{N}^\star$, one has

$$
\begin{aligned}
\|\theta_t^\star - \bar{\theta}_t\|_2 &= \left\| V_{t-1}^{-1} \sum_{s=1}^{t-1} \gamma^{-s} A_s A_s^\top (\theta_s^\star - \theta_t^\star) \right\|_2 \\
&\leq \left\| \sum_{s=t-D}^{t-1} V_{t-1}^{-1} \gamma^{-s} A_s A_s^\top (\theta_s^\star - \theta_t^\star) \right\|_2 + \left\| V_{t-1}^{-1} \sum_{s=1}^{t-D-1} \gamma^{-s} A_s A_s^\top (\theta_s^\star - \theta_t^\star) \right\|_2 \\
&\leq \left\| \sum_{s=t-D}^{t-1} V_{t-1}^{-1} \gamma^{-s} A_s A_s^\top \sum_{p=s}^{t-1} (\theta_p^\star - \theta_{p+1}^\star) \right\|_2 + \left\| \sum_{s=1}^{t-D-1} \gamma^{-s} A_s A_s^\top (\theta_s^\star - \theta_t^\star) \right\|_{V_{t-1}^{-2}} \\
&\leq \left\| \sum_{p=t-D}^{t-1} V_{t-1}^{-1} \gamma^{-s} A_s A_s^\top \sum_{s=t-D}^{p} (\theta_p^\star - \theta_{p+1}^\star) \right\|_2 + \frac{1}{\lambda} \sum_{s=1}^{t-D-1} \gamma^{t-1-s} \|A_s A_s^\top (\theta_s^\star - \theta_t^\star)\|_2 \\
&\leq \sum_{p=t-D}^{t-1} \left\| V_{t-1}^{-1} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top (\theta_p^\star - \theta_{p+1}^\star) \right\|_2 + \frac{2L^2 S}{\lambda} \sum_{s=1}^{t-D-1} \gamma^{t-1-s} \,.
\end{aligned}
$$

The first inequality is a consequence of the triangular inequality. The third inequality uses $V_{t-1}^{-2} \leq (\frac{\gamma^{t-1}}{\lambda})^2 I_d$.

In [Cheung et al., 2019, Russac et al., 2019, Zhao et al., 2020], the authors control the first term using the following argument

$$
\begin{aligned}
\left\| V_{t-1}^{-1} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top (\theta_p^\star - \theta_{p+1}^\star) \right\|_2 &\leq \lambda_{\max} \left(V_{t-1}^{-1} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top \right) \|\theta_p^\star - \theta_{p+1}^\star\|_2 \\
&\leq \|\theta_p^\star - \theta_{p+1}^\star\|_2 \,. \qquad (4.9)
\end{aligned}
$$

Unfortunately, this bound is in general false. $V_{t-1}^{-1} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top$ being not a symmetric matrix its operator norm cannot be bounded by its largest eigenvalue and the maximum eigenvalue itself is not necessarily smaller than 1. In Section 4.3.2.4, we give a correct bound for this term.

4.3.2.2 A Correct Bound with an Additional Assumption

We can however look at *sufficient* conditions for the current analysis to hold. In particular, it is sufficient that $V_{t-1}^{-1} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top$ is a *symmetric* matrix for the different values p. Equivalently, we can require for the two positive semidefinite matrices V_{t-1}^{-1} and $\sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top$

to share the same basis of eigenvectors. This is a strong requirement; not only should it hold for all $t \leq T$, but furthermore such matrices are generated by the algorithm itself. This co-diagonalizability requirement must therefore hold for virtually *any* sequence of arms $\{A_s\}$. The only reasonable situation where this can be verified arises when it is *de-facto* imposed by the geometry of the action set $\mathcal{A} \coloneqq \cup_{t \leq T} \mathcal{A}_t$; for instance, when $\mathcal{A}$ lies along an orthogonal basis.

Assumption 4.1 (Orthogonal arm-set)**.** *Let $\{e_i\}_{i=1}^d$ an orthonormal basis of $\mathbb{R}^d$. We call a collection of arm-sets $\{\mathcal{A}_t\}_t$* orthogonal *if for all $t \geq 1$ and any $a \in \mathcal{A}_t$, there exists α and i such that $a = \alpha e_i$.*

This assumption allows for more general models than the multi-armed bandit setting and in particular it allows each of the action set $\mathcal{A}_t$ to have an arbitrarily large number of actions. Equipped with this additional assumption, we can proceed in upper-bounding the bias using the following property.

$$V_{t-1}^{-1} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top = V_{t-1}^{-1/2} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top V_{t-1}^{-1/2} \coloneqq J_{t-1} \tag{4.10}$$

The advantage, now is that the matrix on the right-hand side of Equation (4.10) is symmetric and we can use the relation $\|Ma\| \leq \|M\| \, \|a\|_2$ that holds for every symmetric matrix M and where $\|M\|$ denotes the operator norm of M. The final step consists in upper-bounding the operator norm of J_{t-1}. Let $x \in \mathbb{R}^d$ such that $\|x\|_2 \leq 1$, we have:

$$x^\top J_{t-1} x = x^\top V_{t-1}^{-1/2} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top V_{t-1}^{-1/2} x \leq x^\top V_{t-1}^{-1/2} V_{t-1} V_{t-1}^{-1/2} x \leq x^\top x \leq 1 \, .$$

From which we deduce,

$$\|J_t\| \leq 1 \, . \tag{4.11}$$

Using this, under Assumption 6.4 the bias term can be bounded by:

$$\sum_{p=t-D}^{t-1} \left\| V_{t-1}^{-1} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top (\theta_p^\star - \theta_{p+1}^\star) \right\|_2 \leq \sum_{p=t-D}^{t-1} \|\theta_p^\star - \theta_{p+1}^\star\|_2 \, .$$

Combining the previous results and using the assumption on the action sets, we have the following regret guarantee:

Theorem 4.2. *With orthogonal arm-sets, assuming that $\sum_{s=1}^{T-1} \|\theta_s^\star - \theta_{s+1}^\star\|_2 \leq B_T$, the regret of the* D-LinUCB *algorithm is bounded for all $\gamma \in (0,1)$ and integer $D \geq 1$, with probability at least $1-\delta$, by*

$$R(T) \leq 2LDB_T + \frac{4L^3 S}{\lambda} \frac{\gamma^D}{1-\gamma} T + 2\sqrt{2}\beta_T \sqrt{dT} \sqrt{T \log(1/\gamma) + \log\left(1 + \frac{L^2}{d\lambda(1-\gamma)}\right)} \, . \tag{4.12}$$

Proof. With this additional assumption and combining previous arguments, the instantaneous regret is bounded with high probability by

$$r_t \leq 2L \sum_{p=t-D}^{t-1} \|\theta_p^\star - \theta_{p+1}^\star\|_2 + \frac{4L^3 S}{\lambda} \frac{\gamma^D}{1-\gamma} + 2\beta_{t-1} \|A_t\|_{V_{t-1}^{-1} \widetilde{V}_{t-1} V_{t-1}^{-1}} .$$

The assumption $\forall a \in \mathcal{A}_t, \ |\langle A_t, \theta_t^\star \rangle| \leq 1$ also implies $r_t \leq 2$. Hence, with probability at least

$1-\delta$:

$$r_t \leq 2L \sum_{p=t-D}^{t-1} \|\theta_p^\star - \theta_{p+1}^\star\|_2 + 4L^3 S \frac{\gamma^D}{1-\gamma} + 2\beta_{t-1} \min\left(1, \|A_t\|_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\right) . \tag{4.13}$$

To conclude the proof we use the results of Subsection 4.B.2 established in Appendix.

$$\begin{aligned}
R(T) &= \sum_{t=1}^{T} r_t \\
&\leq 2L \sum_{t=1}^{T} \sum_{p=t-D}^{t-1} \|\theta_p^\star - \theta_{p+1}^\star\|_2 + \frac{4L^3 S}{\lambda} \frac{\gamma^D}{1-\gamma} T + 2\beta_T \sum_{t=1}^{T} \min\left(1, \|A_t\|_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\right) \\
&\leq 2L \sum_{t=1}^{T} \sum_{p=t-D}^{t-1} \|\theta_p^\star - \theta_{p+1}^\star\|_2 + \frac{4L^3 S}{\lambda} \frac{\gamma^D}{1-\gamma} T + 2\beta_T \sqrt{T} \sqrt{\sum_{t=1}^{T} \min\left(1, \|A_t\|^2_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\right)} \\
&\leq 2LB_T D + \frac{4L^3 S}{\lambda} \frac{\gamma^D}{1-\gamma} T + 2\sqrt{2}\beta_T \sqrt{dT} \sqrt{T \log(1/\gamma) + \log\left(1 + \frac{L^2}{d\lambda(1-\gamma)}\right)}.
\end{aligned}$$

In the first inequality, we use that $t \mapsto \beta_t$ is increasing. The second inequality is an application of the Cauchy-Schwarz inequality to the third term and the last inequality is an application of Corollary 4.15 reported in Appendix. □

In Theorem 4.2 the first two terms of the r.h.s. of Equation (4.16) are the result of the bias due to the non-stationary environment. The last term is the consequence of the high probability bound established in the previous section.

4.3.2.3 Analysis in the General Case

A first fix to the flaw presented in Section 4.3.2.1 was proposed in [Touati and Vincent, 2020]. Here, we follow their line of proof for proposing a valid bound for the bias term without assumption on the geometry of the different action sets. We use:

$$\left\| V_{t-1}^{-1} \sum_{s=1}^{t-1} \gamma^{-s} A_s A_s^\top (\theta_s^\star - \theta_t^\star) \right\|_2 = \max_{x:\|x\|_2=1} \left| x^\top V_{t-1}^{-1} \sum_{s=1}^{t-1} \gamma^{-s} A_s A_s^\top (\theta_s^\star - \theta_t^\star) \right| . \tag{4.14}$$

Let $x \in \mathbb{R}^d$ such that $\|x\|_2 = 1$, we have

$$\begin{aligned}
\left| x^\top V_{t-1}^{-1} \sum_{s=1}^{t-1} \gamma^{-s} A_s A_s^\top (\theta_s^\star - \theta_t^\star) \right| \leq & \left| x^\top V_{t-1}^{-1} \sum_{s=1}^{t-D-1} \gamma^{-s} A_s A_s^\top (\theta_s^\star - \theta_t^\star) \right| \\
& + \left| x^\top V_{t-1}^{-1} \sum_{s=t-D}^{t-1} \gamma^{-s} A_s A_s^\top (\theta_s^\star - \theta_t^\star) \right| .
\end{aligned}$$

For the first term, we use Cauchy-Schwarz as before and obtain:

$$\left| x^\top V_{t-1}^{-1} \sum_{s=1}^{t-D-1} \gamma^{-s} A_s A_s^\top (\theta_s^\star - \theta_t^\star) \right| \leq \|x\|_2 \frac{2SL^2}{\lambda} \frac{\gamma^D}{1-\gamma} .$$

For the second term $b \coloneqq \left|x^\top V_{t-1}^{-1} \sum_{s=t-D}^{t-1} \gamma^{-s} A_s A_s^\top (\theta_s^\star - \theta_t^\star)\right|$ extra work is required.

$$
\begin{aligned}
b &\leq \sum_{s=t-D}^{t-1} \gamma^{-s} \left|x^\top V_{t-1}^{-1} A_s\right| \left|A_s^\top (\theta_s^\star - \theta_t^\star)\right| \leq \sum_{s=t-D}^{t-1} \gamma^{-s} \left|x^\top V_{t-1}^{-1} A_s\right| \left|A_s^\top \sum_{p=s}^{t-1} (\theta_p^\star - \theta_{p+1}^\star)\right| \\
&\leq L \sum_{s=t-D}^{t-1} \gamma^{-s} \left|x^\top V_{t-1}^{-1} A_s\right| \left\|\sum_{p=s}^{t-1} (\theta_p^\star - \theta_{p+1}^\star)\right\|_2 \quad \text{(Cauchy-Schwarz, } \|A_s\|_2 \leq L) \\
&\leq L \sum_{p=t-D}^{t-1} \sum_{s=t-D}^{p} \gamma^{-s} \left|x^\top V_{t-1}^{-1} A_s\right| \|\theta_p^\star - \theta_{p+1}^\star\|_2 \\
&\leq L \sum_{p=t-D}^{t-1} \sum_{s=t-D}^{p} \gamma^{-s} \sqrt{x^\top V_{t-1}^{-1} x} \sqrt{A_s^\top V_{t-1}^{-1} A_s} \|\theta_p^\star - \theta_{p+1}^\star\|_2 \quad \text{(Cauchy-Schwarz)} \\
&\leq L \sum_{p=t-D}^{t-1} \sqrt{\sum_{s=t-D}^{t-1} \gamma^{-s} x^\top V_{t-1}^{-1} x} \sqrt{\sum_{s=t-D}^{t-1} \gamma^{-s} A_s^\top V_{t-1}^{-1} A_s} \, \|\theta_p^\star - \theta_{p+1}^\star\|_2 \quad \text{(Cauchy-Schwarz)} \, .
\end{aligned}
$$

Now,

$$
\begin{aligned}
\sqrt{\sum_{s=t-D}^{t-1} \gamma^{-s} A_s^\top V_{t-1}^{-1} A_s} &= \sqrt{\operatorname{tr}\left(\sum_{s=t-D}^{t-1} \gamma^{-s} A_s^\top V_{t-1}^{-1} A_s\right)} = \sqrt{\operatorname{tr}\left(V_{t-1}^{-1} \sum_{s=t-D}^{t-1} \gamma^{-s} A_s A_s^\top\right)} \\
&\leq \sqrt{\operatorname{tr}(I_d)} = \sqrt{d} \, .
\end{aligned}
$$

Further,

$$
\sqrt{\sum_{s=t-D}^{t-1} \gamma^{-s} x^\top V_{t-1}^{-1} x} \leq \frac{1}{\sqrt{\lambda}} \|x\|_2 \frac{1}{\sqrt{1-\gamma}} \, .
$$

Bringing those bounds together, we get

$$
\|\theta_t^\star - \bar{\theta}_t\|_2 \leq \frac{2SL^2}{\lambda} \frac{\gamma^D}{1-\gamma} + \frac{\sqrt{d}}{\sqrt{\lambda}} \frac{L}{\sqrt{1-\gamma}} \sum_{p=t-D}^{t-1} \|\theta_p^\star - \theta_{p+1}^\star\|_2 \, . \tag{4.15}
$$

Combining the previous results and without the extra assumption on the action sets, we have the following regret guarantee:

Theorem 4.3. *Assuming that $\sum_{s=1}^{T-1} \|\theta_s^\star - \theta_{s+1}^\star\|_2 \leq B_T$, the regret of the* D-LinUCB *algorithm is bounded for all $\gamma \in (0,1)$ and integer $D \geq 1$, with probability at least $1-\delta$, by*

$$
R(T) \leq \frac{2L\sqrt{d}}{\sqrt{\lambda(1-\gamma)}} D B_T + \frac{4L^3 S}{\lambda} \frac{\gamma^D}{1-\gamma} T + 2\sqrt{2} \beta_T \sqrt{dT} \sqrt{T \log\left(\frac{1}{\gamma}\right) + \log\left(1 + \frac{L^2}{d\lambda(1-\gamma)}\right)} \, . \tag{4.16}
$$

Proof. Combining previous arguments, the instantaneous regret is bounded with high probability by:

$$
r_t \leq \frac{2L\sqrt{d}}{\sqrt{\lambda}} \frac{1}{\sqrt{1-\gamma}} \sum_{p=t-D}^{t-1} \|\theta_p^\star - \theta_{p+1}^\star\|_2 + \frac{4L^3 S}{\lambda} \frac{\gamma^D}{1-\gamma} + 2\beta_{t-1} \|A_t\|_{V_{t-1}^{-1} \widetilde{V}_{t-1} V_{t-1}^{-1}} \, .
$$

We then follow the steps from Theorem 4.2 and let $W_t = V_t^{-1}\widetilde{V}_t V_t^{-1}$.

$$R(T) = \sum_{t=1}^{T} r_t \leq \frac{2L\sqrt{d}}{\sqrt{\lambda(1-\gamma)}} \sum_{t=1}^{T} \sum_{p=t-D}^{t-1} \|\theta_p^\star - \theta_{p+1}^\star\|_2 + \frac{4L^3 S}{\lambda} \frac{\gamma^D}{1-\gamma} T + 2\beta_T \sum_{t=1}^{T} \min\left(1, \|A_t\|_{W_{t-1}}\right)$$
$$\leq \frac{2L\sqrt{d}}{\sqrt{\lambda(1-\gamma)}} B_T D + \frac{4L^3 S \gamma^D}{\lambda(1-\gamma)} T + 2\sqrt{2}\beta_T \sqrt{dT} \sqrt{T\log(1/\gamma) + \log\left(1 + \frac{L^2}{d\lambda(1-\gamma)}\right)}.$$

In the first inequality, we use that $t \mapsto \beta_t$ is increasing and the last inequality is an application of Corollary 4.15 reported in Appendix. □

Again here, the first two terms of the r.h.s. of Equation (4.16) are the result of the bias due to the non-stationary environment with the larger dependency in $1/(1-\gamma)$ for the first term now. In the next section, we show that this extra dependency is unfortunately necessary and is not a consequence of a loose upper-bound.

4.3.2.4 Controlling the Bias in Non-Stationary Environments

We recall the error that was made when controlling the bias term:

$$\left\| V_{t-1}^{-1} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top (\theta_p^\star - \theta_{p+1}^\star) \right\|_2 \leq \lambda_{\max}\left(V_{t-1}^{-1} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top \right) \|\theta_p^\star - \theta_{p+1}^\star\|_2 \,.$$

Without a symmetric matrix and denoting $\sigma_{\max}(M)$ the largest singular of a matrix M, the bound would hold with $\sigma_{\max}$ instead of $\lambda_{\max}$. Naturally, the initial bound for controlling the bias from Equation (4.9) would hold if it was possible to show that $\sigma_{\max}\left(V_{t-1}^{-1}\sum_{s=t-D}^{p}\gamma^{-s}A_s A_s^\top\right) \leq 1$. Let I denote an interval included in $[\![1; t-1]\!]$.

In [Cheung et al., 2019], the authors first show (up to the fact that they use a sliding window instead of weights) that $V_{t-1}^{-1}\sum_{s\in I}\gamma^{-s}A_s A_s^\top$ and $V_{t-1}^{-1/2}\sum_{s\in I}\gamma^{-s}A_s A_s^\top V_{t-1}^{-1/2}$ share the same characteristic polynomials. Secondly, as the matrix $V_{t-1}^{-1/2}\sum_{s\in I}\gamma^{-s}A_s A_s^\top V_{t-1}^{-1/2}$ is positive semidefinite, the authors claim that the matrix $V_{t-1}^{-1}\sum_{s\in I}\gamma^{-s}A_s A_s^\top$ is also positive semidefinite.

If $V_{t-1}^{-1}\sum_{s=t-D}^{p}\gamma^{-s}A_s A_s^\top$ was positive semidefinite, for $t-D \leq p \leq t-1$, we would conclude using the following arguments:

$$\sigma_{\max}\left(V_{t-1}^{-1} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top \right) = \max_{x:\|x\|_2=1} x^\top V_{t-1}^{-1} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top x$$
$$\leq \max_{x:\|x\|_2=1} x^\top V_{t-1}^{-1} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top x + x^\top V_{t-1}^{-1} \sum_{s=1}^{t-D-1} \gamma^{-s} A_s A_s^\top x \tag{4.17}$$
$$\leq \max_{x:\|x\|_2=1} x^\top V_{t-1}^{-1} \sum_{s=1}^{p} \gamma^{-s} A_s A_s^\top x + x^\top V_{t-1}^{-1} \sum_{s=p+1}^{t-1} \gamma^{-s} A_s A_s^\top x \tag{4.18}$$
$$\leq \max_{x:\|x\|_2=1} x^\top V_{t-1}^{-1} V_{t-1} x \leq 1 \,.$$

In Equation (4.17) and in Equation (4.18) we have used the semi-definite property that was assumed to hold. Unfortunately, if M_1 and M_2 are two matrices such that M_1 is positive semidefinite (PSD) and M_1 and M_2 have the same characteristic polynomial, it does not imply in general that M_2 is PSD. [Touati and Vincent, 2020] propose the following counterexample when $d = 2$, $M_1 = I_2$ the 2-dimensional identity matrix, $M_2 = ((1, 0)^\top, (-10, 1)^\top)$. Both matrix share the same characteristic polynomial $p(x) = (1 - x)^2$, I_2 is clearly PSD but M_2 is not PSD as with $x = (1, 1)^\top$, $x^\top M_2 x = -8 < 0$.

[Zhao and Zhang, 2021] consider the simpler restart setting and obtain similar regret guarantees than when using a sliding window or discount factors. They further show that establishing $\sigma_{\max}\left(V_{t-1}^{-1}\sum_{s=t-D}^{p}\gamma^{-s}A_sA_s^\top\right) \leq 1$ is not possible and construct a hard problem instance that shows that (the equivalent in the restart setting) the additional $1/\sqrt{1-\gamma}$ dependence when bounding $\left\|V_{t-1}^{-1}\sum_{s=1}^{t-1}\gamma^{-s}A_sA_s^\top(\theta_s^\star - \theta_t^\star)\right\|_2$ is indeed necessary. This shows that there is necessarily an additional cost for bounding the bias without assumption on the action sets.

4.3.3 Asymptotical Bound

It can be checked that, as T tends to infinity, the optimal choice of the analysis parameter D is to take $D = \log(T)/(1 - \gamma)$. Further assuming that one may tune γ as a function of the horizon T and the variation upper bound B_T yields the following result.

Corollary 4.4. *With orthogonal arm-sets, by choosing $\gamma = 1 - (B_T/(dT))^{2/3}$, the pseudo regret of the* D-LinUCB *algorithm is asymptotically upper bounded with high probability by a term $\widetilde{\mathcal{O}}(d^{2/3}B_T^{1/3}T^{2/3})$ when $T \to \infty$.*

Without this assumption, by choosing $\gamma = 1 - (B_T/(\sqrt{d}T))^{1/2}$, the pseudo regret of the D-LinUCB *algorithm is asymptotically upper bounded with high probability by a term $\widetilde{\mathcal{O}}(d^{7/8}B_T^{1/4}T^{3/4})$ when $T \to \infty$.*

With orthogonal arm-sets, this result is favorable as it corresponds to the same order as the lower bound established by [Besbes et al., 2014] and [Cheung et al., 2019]. On the other hand, the guarantee of Corollary 4.4 requires horizon-dependent tuning of the discount factor γ. A first approach for obtaining guarantees without this additional assumption is the Bandit-Over-Bandit mechanism introduced in [Cheung et al., 2019] and for which a regret guarantee of the same order can be obtained [Zhao and Zhang, 2021, Theorem 5].

4.4 Experiments

This section is devoted to the evaluation of the empirical performance of D-LinUCB. We first consider two simulated low-dimensional environments that illustrate the behavior of the algorithms when confronted to either abrupt changes or slow variations of the parameters. The analysis of the previous section suggests that D-LinUCB should behave properly in both situations. We then consider a more realistic scenario in Section 4.4.2, where the contexts are high-dimensional and extracted from a data set of actual user interactions with a web service.

For benchmarking purposes, we compare D-LinUCB to the Dynamic Linear Upper Confidence Bound (dLinUCB) algorithm proposed by [Wu et al., 2018] and with the Sliding Window Linear UCB (SW-LinUCB) of [Cheung et al., 2019]. The principle of the dLinUCB algorithm is that a master bandit algorithm is in charge of choosing the best LinUCB slave bandit for making the

recommendation. Each slave model is built to run in each one of the different environments. The choice of the slave model is based on a lower confidence bound for the so-called *badness* of the different models. The badness is defined as the number of times the expected reward was found to be far enough from the actual observed reward on the last τ_b steps, where τ_b is a parameter of the algorithm. When a slave is chosen, the action proposed to a user is the result of the LinUCB algorithm associated with this slave. When the action is made, all the slave models that were good enough are updated and the models whose badness were too high are deleted from the pool of slaves models. If none of the slaves were found to be sufficiently good, a new slave is added to the pool.

The other algorithm that we use for comparison is SW-LinUCB from [Cheung et al., 2019]. Rather than using exponentially increasing weights, a hard threshold is adopted. Indeed, the actions and rewards included in the τ-length sliding window are used to estimate the linear regression coefficients. We expect D-LinUCB and SW-LinUCB to behave similarly as they both may be shown to have the same sort of regret guarantees (see appendix).

In the case of abrupt changes, we also compare these algorithms to the Oracle Restart LinUCB (LinUCB-OR) strategy that would know the change-points and simply restart, after each change, a new instance of the LinUCB algorithm. The regret of this strategy may be seen as an empirical lower bound on the optimal behavior of an online learning algorithm in abruptly changing environments.

In the following figures, the vertical red dashed lines correspond to the change-points (in abrupt changes scenarios). They are represented to ease the understanding but except for LinUCB-OR, they are of course unknown to the learning algorithms. When applicable, the blue dashed lines correspond to the average detection time of the breakpoints with the dLinUCB algorithm. For D-LinUCB the discount parameter is chosen as $\gamma = 1-(\frac{B_T}{dT})^{2/3}$. For SW-LinUCB the window's length is set to $\tau = (\frac{dT}{B_T})^{2/3}$, where $d = 2$ in the experiment. Those values are theoretically supposed to minimize the asymptotic regret with the orthogonal arm-set assumption and provided good empirical results. For the Dynamic Linear UCB algorithm, the badness is estimated from $\tau_b = 200$ steps, as in the experimental section of [Wu et al., 2018].

4.4.1 Synthetic Data in Abruptly-Changing or Slowly-Varying Scenarios

In this first experiment, we observe the empirical performance of all algorithms in an abruptly changing environment of dimension 2 with 3 breakpoints. The number of rounds is set to $T = 6000$. The light blue triangles correspond to the different positions of the true unknown parameter $\theta_t^\star$: before $t = 1000$, $\theta_t^\star = (1, 0)$; for $t \in [\![1000, 2000]\!]$, $\theta_t^\star = (-1, 0)$; for $t \in [\![2000, 3000]\!]$, $\theta_t^\star = (0, 1)$; and, finally, for $t > 3000$, $\theta_t^\star = (0, -1)$. This corresponds to a hard problem as the sequence of parameters is widely spread in the unit ball. Indeed it forces the algorithm to adapt to big changes, which typically requires a longer adaptation phase. On the other hand, it makes the detection of changes easier, which is an advantage for dLinUCB. In the second half of the experiment (when $t \geq 3000$) there is no change, LinUCB struggles to catch up and suffers linear regret for long periods after the last change-point. The results of our simulations are shown in the left column of Figure 4.1. On the top row we show a 2-dimensional scatter plot of the estimate of the unknown parameters $\hat{\theta}_t$ every 1000 steps averaged on 100 independent experiment. The bottom row corresponds to the regret averaged over 100 independent experiments with the upper and the lower 5% quantiles. In this environment, with 1-subgaussian random noise, dLinUCB struggles to detect the change-points. Over the 100 experiments, the first change-point was

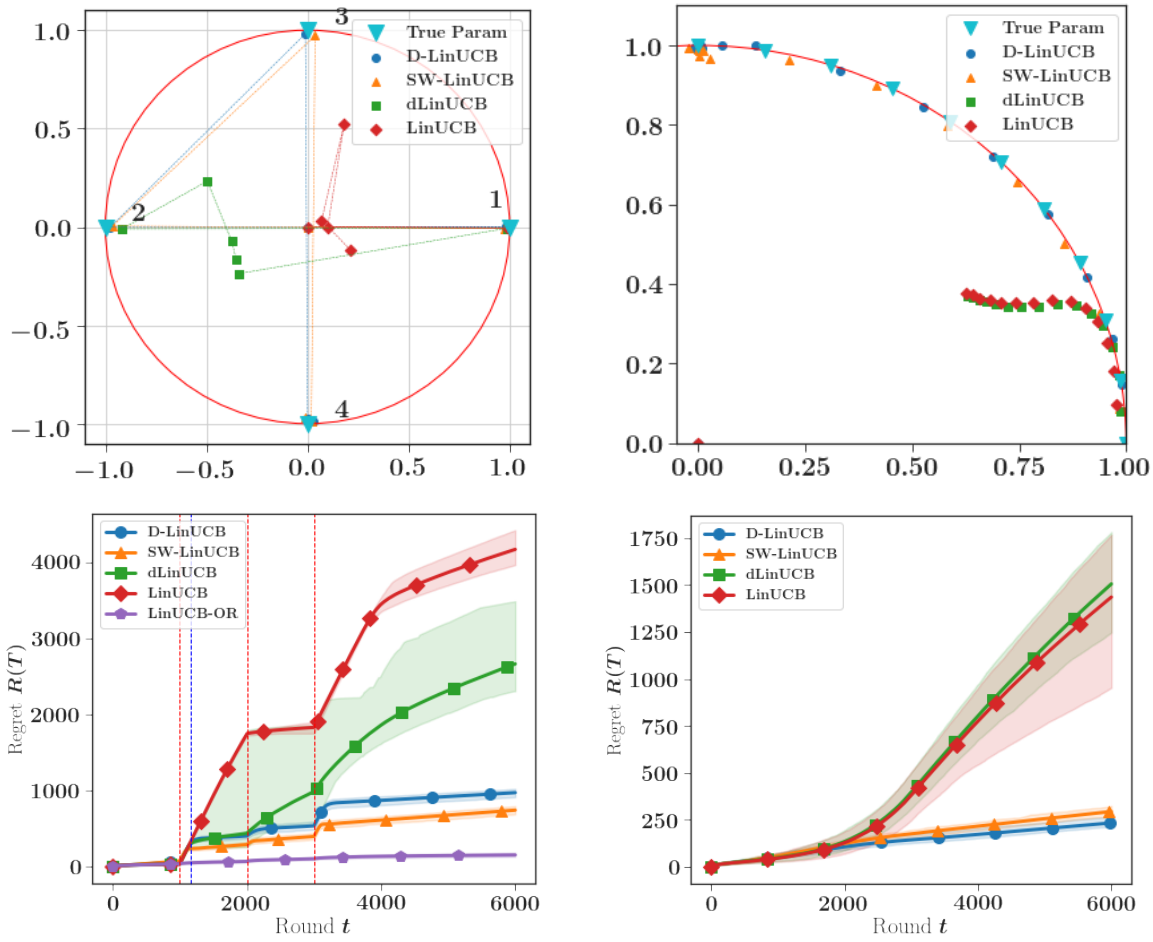


Figure 4.1: Performances of the algorithms in the abruptly-changing environment (on the left), and, the slowly-varying environment (on the right). The upper plots correspond to the estimated parameter and the lower ones to the accumulated regret, both are averaged on $N = 100$ independent experiments

detected in 95% of the runs, the second was never detected and the third only in 6% of the runs, thus limiting the effectiveness of the dLinUCB approach. When decreasing the variance of the noise, the performance of dLinUCB improves and gets closer to the performance of the oracle restart strategy LinUCB-OR. It is worth noting that for both SW-LinUCB and D-LinUCB, the estimator $\hat{\theta}_t$ adapts itself to non-stationarity and is able to follow $\theta_t^\star$ (with some delay), as shown on the scatter plot. Predictably, LinUCB-OR achieves the best performance by restarting exactly whenever a change-point happens.

The second experiment corresponds to a slowly-changing environment. It is easier for LinUCB to keep up with the adaptive policies in this scenario. Here, the parameter $\theta_t^\star$ starts at 1 and moves continuously counter-clockwise on the unit-circle up to the position $[0, 1]$ in 3000 steps. We then have a steady period of 3000 steps. For this sequence of parameters, $\mathcal{B}_T = \sum_{t=1}^{T-1} \|\theta_t^\star - \theta_{t+1}^\star\|_2 = 1.57$. The results are reported in the right column of Figure 4.1. Unsurprisingly, dLinUCB does not detect any change and thus displays the same performance as LinUCB. SW-LinUCB and D-LinUCB behaves similarly and are both robust to such an evolution in the regression parameters. The performance of LinUCB-OR is not reported here, as

restarting becomes ineffective when the changes are too frequent (here, during the first 3000 time steps, there is a change at every single step). The scatter plot also gives interesting information: $\hat{\theta}_t$ tracks $\theta_t^\star$ quite effectively for both SW-LinUCB and D-LinUCB but the two others algorithms lag behind. LinUCB will eventually catch up if the length of the stationary period becomes larger.

4.4.2 Simulation Based on a Real Dataset

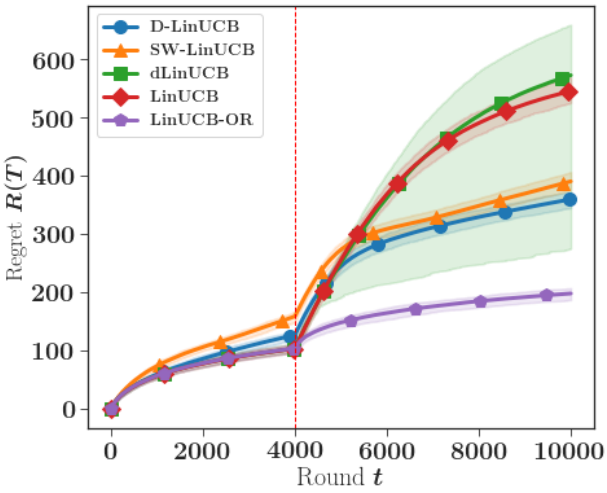


Figure 4.2: Behavior of the different algorithms on large-dimensional data

D-LinUCB also performs well in high-dimensional space ($d = 50$). For this experiment, a dataset providing a sample of 30 days of Criteo live traffic data [Diemert et al., 2017] was used. It contains banners that were displayed to different users and contextual variables, including the information of whether the banner was clicked or not. We kept the categorical variables $cat1$ to $cat9$, together with the variable *campaign*, which is a unique identifier of each campaign. Beforehand, these contexts have been one-hot encoded and 50 of the resulting features have been selected using a Singular Value Decomposition. $\theta^\star$ is obtained by linear regression. The rewards are then simulated using the regression model with an additional Gaussian noise of variance $\sigma^2 = 0.15$. At each time step, the different algorithms have the choice between two 50-dimensional contexts drawn at random from two separate pools of 10000 contexts corresponding, respectively, to clicked or not clicked banners. The non-stationarity is created by switching 60% of $\theta^\star$ coordinates to $-\theta^\star$ at time 4000, corresponding to a partial class inversion. The cumulative dynamic regret is then averaged over 100 independent replications. The results are shown on Figure 4.2. In the first stationary period, LinUCB and dLinUCB perform better than the adaptive policies by using all available data, whereas the adaptive policies only use the most recent events. After the breakpoint, LinUCB suffers a large regret, as the algorithm fails to adapt to the new environment. In this experiment, dLinUCB does not detect the change-point systematically and performs similarly as LinUCB on average, it can still outperform adaptive policies from time to time when the breakpoint is detected as can be seen with the 5% quantile. D-LinUCB and SW-LinUCB adapt more quickly to the change-point and perform significantly better than the non-adaptive policies after the breakpoint. Of course, the oracle policy LinUCB-OR is the best performing policy. The take-away message is that there is no free lunch: in a stationary period by using only the most recent events SW-LinUCB and D-LinUCB do not perform as good as a

policy that uses all the available information. Nevertheless, after a breakpoint, the recovery is much faster with the adaptive policies.

4.5 Conclusion

In this chapter, we considered the non-stationary linear bandit setting where the regression parameter $\theta^\star$ can vary over time. We measured the non-stationarity through the variation budget $\mathcal{B}_T$ a general setting that contains both abruptly changing and slowly drifting environments. We proposed D-LinUCB an adaptive linear bandit algorithm based on carefully designed exponential weights. With an additional assumption on the different action-sets and with the knowledge of an upper-bound B_T on the true variation budget, we established the asymptotic optimality of this algorithm. A regret of order $\mathcal{O}(B_T^{1/3}T^{2/3})$ was obtained, matching the existing lower bound for this setting (Theorem 1.20). In the general case however, we explained a technical flaw in existing approaches based on forgetting mechanisms, and obtained a degraded regret bound of order $\mathcal{O}(B_T^{1/4}T^{3/4})$.

Appendix

Appendix 4.A Confidence Bounds for Weighted Linear Bandits

4.A.1 Preliminary Results

In this section we give the main results for obtaining Theorem 4.1. For the sake of conciseness all the results will be stated with σ-subgaussian noises but the proofs will be done with the particular value of $\sigma = 1$. We recall that $(\eta_s)_s$ is, conditionally on the past, a sequence of σ-subgaussian random noises. The results of this section are close to the one proposed in [Abbasi-Yadkori et al., 2011] but our results are valid with a sequence of predictable weights.

We introduce the quantity $S_t = \sum_{s=1}^{t} w_s A_s \eta_s$ and $\widetilde{V}_t = \sum_{s=1}^{t} w_s^2 A_s A_s^\top + \mu_t I_d$. When the regularization term is omitted, we let $\widetilde{V}_t(0) = \sum_{s=1}^{t} w_s^2 A_s A_s^\top$. The filtration associated with the random observations is denoted $\mathcal{F}_t$ such that A_t is $\mathcal{F}_{t-1}$-measurable and η_t is $\mathcal{F}_t$-measurable. The weights are also assumed to be predictable. The following lemma is an extension to the weighted case of Lemma 8 of [Abbasi-Yadkori et al., 2011].

Lemma 4.5. *Let $(w_t)_{t\geq 1}$ be a sequence of predictable and positive weights. Let $x \in \mathbb{R}^d$ be arbitrary and consider for any $t \geq 1$*

$$M_t(x) \coloneqq \exp\left(\frac{1}{\sigma} x^\top S_t - \frac{1}{2} x^\top \widetilde{V}_t(0) x\right) .$$

Let τ be a stopping time with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Then $M_\tau(x)$ is almost surely well-defined and

$$\forall x \in \mathbb{R}^d, \quad \mathbb{E}[M_\tau(x)] \leq 1 .$$

Proof. First, we prove that $\forall x \in \mathbb{R}^d, (M_t(x))_{t=0}^{\infty}$ is a super-martingale.
Let $x \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbb{E}[M_t(x)|\mathcal{F}_{t-1}] &= \mathbb{E}\left[\exp\left(x^\top S_{t-1} + x^\top w_t A_t \eta_t - 1/2 x^\top (\widetilde{V}_{t-1}(0) + w_t^2 A_t A_t^\top) x\right) |\mathcal{F}_{t-1}\right] \\
&= M_{t-1}(x) \mathbb{E}\left[\exp(x^\top w_t A_t \eta_t - \frac{1}{2} w_t^2 x^\top A_t A_t^\top x) |\mathcal{F}_{t-1}\right] \\
&= M_{t-1}(x) \exp\left(-\frac{1}{2} w_t^2 x^\top A_t A_t^\top x\right) \mathbb{E}\left[\exp(x^\top w_t A_t \eta_t) |\mathcal{F}_{t-1}\right] \\
&\leq M_{t-1}(x) \exp\left(-\frac{1}{2} w_t^2 x^\top A_t A_t^\top x\right) \exp\left(\frac{1}{2} w_t^2 (x^\top A_t)^2\right) \\
&= M_{t-1}(x) .
\end{aligned}$$

The second equality comes from the fact that S_{t-1} and $\widetilde{V}_{t-1}$ are $\mathcal{F}_{t-1}$-measurable. The inequality comes from the definition of the conditional 1-subgaussianity where we also use the $\mathcal{F}_{t-1}$-measurability of w_t.

Using this supermartingale property, we have $\mathbb{E}[M_t(x)] \leq 1$. The convergence theorem for non-negative supermartingales ensures that $M_\infty(x) = \lim_{t\to\infty} M_t(x)$ is almost surely well defined. By introducing the stopped supermartingale $\mathcal{M}_t(x) = M_{\min(t,\tau)}(x)$, we have $M_\tau(x) = \lim_{t\to\infty} \mathcal{M}_t(x)$. Knowing that $\mathcal{M}_t(x)$ is also a supermartingale, we have

$$\mathbb{E}[\mathcal{M}_t(x)] = \mathbb{E}[M_{\min(t,\tau)}(x)] \leq \mathbb{E}[M_{\min(0,\tau)}(x)] = \mathbb{E}[M_0(x)] = 1 .$$

By using Fatou's lemma:

$$\mathbb{E}[M_\tau(x)] = \mathbb{E}[\liminf_{t\to\infty} \mathcal{M}_t(x)] \leq \liminf_{t\to\infty} \mathbb{E}[\mathcal{M}_t(x)] \leq 1 \ .$$

□

In the next lemma, we will integrate $M_t(x)$ with respect to a time-dependent probability measure. This is the key for allowing sequential regularizations in the concentration inequality stated in Theorem 4.1. This lemma is inspired by the method of mixtures first presented in [Peña et al., 2008]. The idea of using time-varying probability measures is inspired from the proof of Theorem 11 in [Kirschner and Krause, 2018]. The two following lemmas are included in the appendix so that the chapter is self-contained. There are not a mere consequence of the results in [Abbasi-Yadkori et al., 2011] because of the time-dependent regularization parameters.

Lemma 4.6. *Let $(h_t)_t$ be a sequence of probability measures on $\mathbb{R}^d$. We define $\widetilde{M}_t = \int_{\mathbb{R}^d} M_t(x) dh_t(x)$. Then,*

$$\forall t, \quad \mathbb{E}[\widetilde{M}_t] \leq 1 \ .$$

Proof.

$$\begin{aligned}
\mathbb{E}[\widetilde{M}_t] = \int \widetilde{M}_t \, d\mathbb{P} &= \int \left(\int_{\mathbb{R}^d} M_t(x) dh_t(x) \right) d\mathbb{P} \\
&= \int_{\mathbb{R}^d} \left(\int M_t(x) d\mathbb{P} \right) dh_t(x) \quad \text{(Fubini's theorem)} \\
&= \int_{\mathbb{R}^d} \mathbb{E}[M_t(x)] dh_t(x) \leq \int_{\mathbb{R}^d} dh_t(x) \quad \text{(Lemma 4.5)} \\
&\leq 1 \ . \quad (h_t \text{ probability measure.})
\end{aligned}$$

□

Lemma 4.6 is a warm-up for the next lemma and is helpful for understanding why Lemma 4.7 holds. It is valid for any fixed time t. The next step is to give its equivalent in a stopped version in the specific case of gaussian random vectors.

Lemma 4.7. *Let $(\mu_t)_t$ be a deterministic sequence of regularization parameters. Let $\mathcal{F}_\infty = \sigma\left(\cup_{t=1}^\infty \mathcal{F}_t\right)$ be the tail σ-algebra of the filtration $(\mathcal{F}_t)_t$. Let $X = (X_t)_{t\geq 1}$ be an independent sequence of gaussian random vectors such that $X_t \sim \mathcal{N}(0, \frac{1}{\mu_t} I_d) = h_t$ with X independent of $\mathcal{F}_\infty$. We define*

$$\bar{M}_t(\mu_t) = \mathbb{E}[M_t(X_t)|\mathcal{F}_\infty] = \int_{\mathbb{R}^d} M_t(x) f_{\mu_t}(x) dx \ ,$$

where f_{μ_t} is the probability density function associated with h_t defined as,

$$f_{\mu_t}(x) = \frac{1}{\sqrt{(2\pi)^d \det(1/\mu_t I_d)}} \exp\left(-\frac{\mu_t x^\top x}{2} \right) \ .$$

Let τ be a stopping time with respect to the filtration $(\mathcal{F}_t)_t$ then,

$$\mathbb{E}[\bar{M}_\tau(\mu_\tau)] \leq 1 \ .$$

Proof. We can use the result of Lemma 4.5 which gives $\forall x \in \mathbb{R}^d$, $\mathbb{E}[M_\tau(x)] \leq 1$. We have,

$$\begin{aligned}\mathbb{E}[\bar{M}_\tau(\mu_\tau)] &= \mathbb{E}[\mathbb{E}[M_\tau(X_\tau)|\mathcal{F}_\infty]] = \mathbb{E}[\mathbb{E}[\mathbb{E}[M_\tau(X_\tau)|\mathcal{F}_\infty]|(X_t)_{t\geq 1}]] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{E}[M_\tau(X_\tau)|(X_t)_{t\geq 1}]|\mathcal{F}_\infty]] \leq 1 \ .\end{aligned}$$

The inequality is a consequence of Lemma 4.5 as, conditionally to the sequence $(X_t)_t$, $M_\tau(X_\tau)$ is of the form $M_\tau(x)$ with a fixed x. □

We finally state the main result needed to obtain Theorem 4.1.

Proposition 4.8. *For $(w_s)_{s\geq 1}$ a sequence of predictable and positive weights, $\forall \delta > 0$, the following deviation inequality holds*

$$\mathbb{P}\left(\exists t \geq 0, \|S_t\|_{\widetilde{V}_t^{-1}} \geq \sigma\sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\widetilde{V}_t)}{\mu_t^d}\right)}\right) \leq \delta \ .$$

Proof. For a fixed t,

$$\begin{aligned}\bar{M}_t(\mu_t) &= \int_{\mathbb{R}^d} M_t(x) f_{\mu_t}(x)dx = \frac{1}{\sqrt{(2\pi)^d \det(1/\mu_t I_d)}} \int_{\mathbb{R}^d} \exp\left(x^\top S_t - \frac{1}{2}\|x\|^2_{\mu_t I_d} - \frac{1}{2}\|x\|^2_{\widetilde{V}_t(0)}\right) dx \\ &= \frac{1}{\sqrt{(2\pi)^d \det(1/\mu_t I_d)}} \int_{\mathbb{R}^d} \exp\left(x^\top S_t - \frac{1}{2}\|x\|^2_{\widetilde{V}_t}\right) dx \\ &= \frac{1}{\sqrt{(2\pi)^d \det(1/\mu_t I_d)}} \int_{\mathbb{R}^d} \exp\left(\frac{1}{2}\|S_t\|^2_{\widetilde{V}_t^{-1}} - \frac{1}{2}\|x - \widetilde{V}_t^{-1}S_t\|^2_{\widetilde{V}_t}\right) dx \\ &= \frac{\exp\left(\frac{1}{2}\|S_t\|^2_{\widetilde{V}_t^{-1}}\right)}{\sqrt{(2\pi)^d \det(1/\mu_t I_d)}} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\|x - \widetilde{V}_t^{-1}S_t\|^2_{\widetilde{V}_t}\right) dx \\ &= \frac{\exp\left(\frac{1}{2}\|S_t\|^2_{\widetilde{V}_t^{-1}}\right)}{\sqrt{(2\pi)^d \det(1/\mu_t I_d)}} \sqrt{(2\pi)^d \det\left(\widetilde{V}_t^{-1}\right)} = \exp\left(\frac{1}{2}\|S_t\|^2_{\widetilde{V}_t^{-1}}\right)\sqrt{\frac{\det(\mu_t I_d)}{\det(\widetilde{V}_t)}} \ .\end{aligned}$$

We introduce the particular stopping time,

$$\tau = \min\left\{t \geq 0, \|S_t\|_{\widetilde{V}_t^{-1}} \geq \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\widetilde{V}_t)}{\det(\mu_t I_d)}\right)}\right\} \ .$$

Thus,

$$\begin{aligned}&\mathbb{P}\left(\exists t \geq 0, \|S_t\|_{\widetilde{V}_t^{-1}} \geq \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\widetilde{V}_t)}{\det(\mu_t I_d)}\right)}\right) = \mathbb{P}(\tau < \infty) \\ &= \mathbb{P}\left(\tau < \infty, \|S_\tau\|_{\widetilde{V}_\tau^{-1}} \geq \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\widetilde{V}_\tau)}{\det(\mu_\tau I_d)}\right)}\right) \\ &\leq \mathbb{P}\left(\|S_\tau\|_{\widetilde{V}_\tau^{-1}} \geq \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\widetilde{V}_\tau)}{\det(\mu_\tau I_d)}\right)}\right) \\ &= \mathbb{P}\left(\exp\left(\frac{1}{2}\|S_\tau\|^2_{\widetilde{V}_\tau^{-1}}\right)\sqrt{\frac{\det(\mu_\tau I_d)}{\det(\widetilde{V}_\tau)}} \geq \frac{1}{\delta}\right) \leq \delta\mathbb{E}[\bar{M}_\tau(\mu_\tau)] \leq \delta \text{ (Lemma 4.7)} \ .\end{aligned}$$

□

Appendix 4.B D-LinUCB Analysis

In this section, the environment is non-stationary, which means that the unknown parameter $\theta^\star$ may evolve over time and is henceforth denoted $\theta_t^\star$. The reward generation process in the one presented in Equation (1.20).

4.B.1 Preliminary Results

In this section, V_t and $\widetilde{V}_t$ are defined by

$$V_t = \sum_{s=1}^{t} \gamma^{-s} A_s A_s^\top + \lambda \gamma^{-t} I_d, \quad \widetilde{V}_t = \sum_{s=1}^{t} \gamma^{-2s} A_s A_s^\top + \lambda \gamma^{-2t} I_d \,.$$

We recall the definition of β_t:

$$\beta_t = \sqrt{\lambda} S + \sigma \sqrt{2\log(1/\delta) + d\log\left(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)}\right)} \,.$$

With $\hat{\theta}_t$ defined in Equation (4.2), the confidence ellipsoid we consider is defined by

$$\mathcal{C}_t = \left\{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} \leq \beta_{t-1} \right\} . \tag{4.19}$$

Theorem 4.1 can be applied with this choice of weights and regularization. We combine it with an upper bound for $\det(\widetilde{V}_t)$ given below.

Proposition 4.9 (Determinant inequality for the weighted design matrix)**.** *Let $(\lambda_t)_t$ be a deterministic sequence of regularization parameters. Let $V_t = \sum_{s=1}^t w_s A_s A_s^\top + \lambda_t I_d$ be the weighted design matrix. Under the assumption $\forall t, \|A_t\|_2 \leq L$, the following holds*

$$\det(V_t) \leq \left(\lambda_t + \frac{L^2 \sum_{s=1}^t w_s}{d} \right)^d .$$

Proof.

$$\begin{aligned}
\det(V_t) &= \prod_{i=1}^{d} l_i \quad (l_i \text{ are the eigenvalues}) \\
&\leq \left(\frac{1}{d} \sum_{i=1}^{d} l_i \right)^d \quad \text{(AM-GM inequality)} \\
&\leq \left(\frac{1}{d} \mathrm{trace}(V_t) \right)^d \leq \left(\frac{1}{d} \sum_{s=1}^{t} w_s \mathrm{trace}(A_s A_s^\top) + \lambda_t \right)^d \\
&\leq \left(\frac{1}{d} \sum_{s=1}^{t} w_s \|A_s\|_2^2 + \lambda_t \right)^d \leq \left(\lambda_t + \frac{L^2}{d} \sum_{s=1}^{t} w_s \right)^d .
\end{aligned}$$

□

Corollary 4.10. *In the specific case where the weights are given by $w_t = \gamma^{-t}$ with $0 < \gamma < 1$. Proposition 4.9 can be rewritten*

$$\det(V_t) \leq \left(\lambda_t + \frac{L^2(\gamma^{-t}-1)}{d(1-\gamma)}\right)^d = \left(\lambda\gamma^{-t} + \frac{L^2(\gamma^{-t}-1)}{d(1-\gamma)}\right)^d .$$

We also have,

$$\det(\widetilde{V}_t) \leq \left(\mu_t + \frac{L^2(\gamma^{-2t}-1)}{d(1-\gamma^2)}\right)^d = \left(\lambda\gamma^{-2t} + \frac{L^2(\gamma^{-2t}-1)}{d(1-\gamma^2)}\right)^d .$$

Proof. Apply Proposition 4.9 and use $\sum_{s=1}^t \gamma^{-s} = \frac{\gamma^{-t}-1}{1-\gamma}$ and $\sum_{s=1}^t \gamma^{-2s} = \frac{\gamma^{-2t}-1}{1-\gamma^2}$. □

Corollary 4.10 and Proposition 4.8 yield the following result.

Corollary 4.11. *$\forall\delta > 0$, with the weights $w_t = \gamma^{-t}$ and $0 < \gamma < 1$, we have*

$$\mathbb{P}\left(\exists t \geq 0, \|S_t\|_{\widetilde{V}_t^{-1}} \geq \sigma\sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)}\right)}\right) \leq \delta.$$

Thanks to this corollary we are now ready to show that $\bar{\theta}_t$ belongs to $\mathcal{C}_{t-1}$ with high probability.

Proposition 4.12. *Let $\mathcal{C}_t = \left\{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} \leq \beta_{t-1}\right\}$ denote the confidence ellipsoid. Let $\bar{\theta}_t = V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{-s}A_sA_s^\top\theta_s^\star + \lambda\gamma^{-(t-1)}\theta_t^\star\right)$. Then, $\forall\delta > 0$,*

$$\mathbb{P}\left(\forall t \geq 1, \bar{\theta}_t \in \mathcal{C}_t\right) \geq 1 - \delta .$$

Proof.

$$\begin{aligned}
\bar{\theta}_t - \hat{\theta}_{t-1} &= V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{-s}A_sA_s^\top\theta_s^\star + \lambda\gamma^{-(t-1)}\theta_t^\star - \sum_{s=1}^{t-1}\gamma^{-s}A_sX_s\right) \\
&= V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{-s}A_sA_s^\top\theta_s^\star + \lambda\gamma^{-(t-1)}\theta_t^\star - \sum_{s=1}^{t-1}\gamma^{-s}A_sA_s^\top\theta_s^\star - \sum_{s=1}^{t-1}\gamma^{-s}A_s\eta_s\right) \\
&= -V_{t-1}^{-1}S_{t-1} + \lambda\gamma^{-(t-1)}V_{t-1}^{-1}\theta_t^\star .
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\bar{\theta}_t - \hat{\theta}_{t-1}\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} &\leq \|S_{t-1}\|_{\widetilde{V}_{t-1}^{-1}} + \lambda\gamma^{-(t-1)}\|\theta_t^\star\|_{\widetilde{V}_{t-1}^{-1}} \\
&\leq \|S_{t-1}\|_{\widetilde{V}_{t-1}^{-1}} + \sqrt{\lambda}S \quad (\widetilde{V}_{t-1}^{-1} \leq 1/(\gamma^{-2(t-1)}\lambda)I_d \text{ and } \|\theta_t^\star\|_2 \leq S) \\
&\leq \beta_{t-1} . \quad \text{(Corollary 4.11)}
\end{aligned}$$

□

4.B.2 Control of the Norm of Actions

Lemma 4.13. *Let* $V_t = \sum_{s=1}^{t} \gamma^{-s} A_s A_s^\top + \lambda\gamma^{-t} I_d$, $\widetilde{V}_t = \sum_{s=1}^{t} \gamma^{-2s} A_s A_s^\top + \lambda\gamma^{-2t} I_d$ *and* $0 < \gamma < 1$. *We have*

$$\forall t,\ V_t^{-1} \widetilde{V}_t V_t^{-1} \leq \gamma^{-t} V_t^{-1} .$$

Proof.

$$\widetilde{V}_t = \sum_{s=1}^{t} \gamma^{-2s} A_s A_s^\top + \lambda\gamma^{-2t} I_d \leq \gamma^{-t} \sum_{s=1}^{t} \gamma^{-s} A_s A_s^\top + \lambda\gamma^{-2t} I_d = \gamma^{-t} V_t .$$

Consequently,

$$V_t^{-1} \widetilde{V}_t V_t^{-1} \leq \gamma^{-t} V_t^{-1} V_t V_t^{-1} \leq \gamma^{-t} V_t^{-1} .$$

□

Thanks to Lemma 4.13 we establish the following proposition,

Proposition 4.14.

$$\sum_{t=1}^{T} \min\left(1, \|A_t\|^2_{V_{t-1}^{-1} \widetilde{V}_{t-1} V_{t-1}^{-1}}\right) \leq 2 \sum_{t=1}^{T} \log\left(1 + \gamma^{-t} \|A_t\|^2_{V_{t-1}^{-1}}\right) \leq 2 \log\left(\frac{\det(V_T)}{\lambda^d}\right) .$$

Proof. We first use the fact that: $\forall x \geq 0, \min(1, x) \leq 2\log(1 + x)$.

$$\begin{aligned}
\min\left(1, \|A_t\|^2_{V_{t-1}^{-1} \widetilde{V}_{t-1} V_{t-1}^{-1}}\right) &\leq 2 \log\left(1 + \|A_t\|^2_{V_{t-1}^{-1} \widetilde{V}_{t-1} V_{t-1}^{-1}}\right) \\
&\leq 2 \log\left(1 + \gamma^{-(t-1)} \|A_t\|^2_{V_{t-1}^{-1}}\right) \quad \text{(Lemma 4.13)} \\
&\leq 2 \log\left(1 + \gamma^{-t} \|A_t\|^2_{V_{t-1}^{-1}}\right) . \quad (\gamma \leq 1)
\end{aligned}$$

Furthermore,

$$V_t \geq \gamma^{-t} A_t A_t^\top + V_{t-1} \geq V_{t-1}^{1/2} (I_d + \gamma^{-t} V_{t-1}^{-1/2} A_t A_t^\top V_{t-1}^{-1/2}) V_{t-1}^{1/2} .$$

Given that all those matrices are symmetric positive definite, the previous inequality implies that

$$\begin{aligned}
\det(V_t) &\geq \det(V_{t-1}) \det(1 + (\gamma^{-t/2} V_{t-1}^{-1/2} A_t)(\gamma^{-t/2} V_{t-1}^{-1/2} A_t)^\top) \\
&\geq \det(V_{t-1}) \left(1 + \gamma^{-t} \|A_t\|^2_{V_{t-1}^{-1}}\right) . \quad \left(\text{Using } \det(I_d + xx^\top) = 1 + \|x\|_2^2\right)
\end{aligned}$$

Therefore,

$$\frac{\det(V_T)}{\det(V_0)} = \prod_{t=1}^{T} \frac{\det(V_t)}{\det(V_{t-1})} \geq \prod_{t=1}^{T} \left(1 + \gamma^{-t} \|A_t\|^2_{V_{t-1}^{-1}}\right) .$$

Finally by applying the log function to the previous inequality,

$$\sum_{t=1}^{T} \min\left(1, \|A_t\|^2_{V_{t-1}^{-1} \widetilde{V}_{t-1} V_{t-1}^{-1}}\right) \leq 2 \sum_{t=1}^{T} \log\left(1 + \gamma^{-t} \|A_t\|^2_{V_{t-1}^{-1}}\right) \leq 2 \log\left(\frac{\det(V_T)}{\det(V_0)}\right) .$$

□

Corollary 4.15.

$$\sqrt{\sum_{t=1}^{T} \min\left(1, \|A_t\|^2_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\right)} \leq \sqrt{2d}\sqrt{T\log\left(\frac{1}{\gamma}\right) + \log\left(1 + \frac{L^2}{d\lambda(1-\gamma)}\right)}\,.$$

Proof. The proof of this corollary is based on the previous lemma and on Corollary 4.10. We have

$$\begin{aligned}\log\left(\frac{\det(V_T)}{\det(V_0)}\right) &\leq \log\left(\frac{1}{\lambda^d}\left(\lambda\gamma^{-T} + \frac{L^2(\gamma^{-T}-1)}{d(1-\gamma)}\right)^d\right) \quad \text{(Corollary 4.10)}\\ &\leq dT\log\left(\frac{1}{\gamma}\right) + d\log\left(1 + \frac{L^2}{d\lambda(1-\gamma)}\right)\,.\end{aligned}$$

□

4.B.3 Proof of Corollary 4.4

Corollary 4.4. *With orthogonal arm-sets, by choosing $\gamma = 1 - (B_T/(dT))^{2/3}$, the pseudo regret of the* D-LinUCB *algorithm is asymptotically upper bounded with high probability by a term $\widetilde{\mathcal{O}}(d^{2/3}B_T^{1/3}T^{2/3})$ when $T \to \infty$.*

Without this assumption, by choosing $\gamma = 1 - (B_T/(\sqrt{d}T))^{1/2}$, the pseudo regret of the D-LinUCB *algorithm is asymptotically upper bounded with high probability by a term $\widetilde{\mathcal{O}}(d^{7/8}B_T^{1/4}T^{3/4})$ when $T \to \infty$.*

Proof. Let us start with the case where the actions sets are orthogonal. Using Theorem 4.2, the pseudo-regret of D-LinUCB can be bounded with high probability in the following way:

$$R(T) \leq 2LDB_T + \frac{4L^3S}{\lambda}\frac{\gamma^D}{1-\gamma}T + 2\sqrt{2}\beta_T\sqrt{dT}\sqrt{T\log(1/\gamma) + \log\left(1 + \frac{L^2}{d\lambda(1-\gamma)}\right)}\,.$$

Let γ be defined as $\gamma = 1 - (\frac{B_T}{dT})^{2/3}$ and $D = \frac{\log(T)}{(1-\gamma)}$. With this choice of γ, D is equivalent to $d^{2/3}B_T^{-2/3}T^{2/3}\log(T)$. Thus, DB_T is equivalent to $d^{2/3}B_T^{1/3}T^{2/3}\log(T)$.
In addition,

$$\gamma^D = \exp(D\log(\gamma)) = \exp\left(\frac{\log(\gamma)}{1-\gamma}\log(T)\right) \sim 1/T\,.$$

Hence, $T\gamma^D\frac{1}{1-\gamma}$ behaves as $d^{2/3}T^{2/3}B_T^{-2/3}$.
Furthermore, $\log(1/\gamma) \sim d^{-2/3}B_T^{2/3}T^{-2/3}$, implying that $T\log(1/\gamma) \sim d^{-2/3}B_T^{2/3}T^{1/3}$.
As a result, it holds that, $\beta_T\sqrt{dT}\sqrt{T\log(1/\gamma) + \log\left(1 + \frac{L^2}{d\lambda(1-\gamma)}\right)}$ is equivalent to $dT^{1/2}\sqrt{\log(T/B_T)}\sqrt{d^{-2/3}B_T^{2/3}T^{1/3}} = d^{2/3}B_T^{1/3}T^{2/3}\sqrt{\log(T/B_T)}$. By adding those three terms and neglecting the log factors, we obtain the desired result.

Let us now drop the additional assumption on the action-sets. Using Theorem 4.3, the

pseudo-regret now satisfies

$$R(T) \leq \frac{2L\sqrt{d}}{\sqrt{\lambda(1-\gamma)}} DB_T + \frac{4L^3 S}{\lambda} \frac{\gamma^D}{1-\gamma} T + 2\sqrt{2}\beta_T \sqrt{dT} \sqrt{T \log\left(\frac{1}{\gamma}\right) + \log\left(1 + \frac{L^2}{d\lambda(1-\gamma)}\right)} .$$

This time Let γ be defined as $\gamma = 1 - (\frac{B_T}{\sqrt{dT}})^{1/2}$ and D is unchanged. With this choice of γ,

$$\frac{\sqrt{d} D B_T}{\sqrt{1-\gamma}} \sim \frac{\sqrt{d} B_T \log T}{(1-\gamma)^{3/2}} \sim d^{7/8} B_T^{1/4} T^{3/4} .$$

In addition,

$$\gamma^D = \exp(D \log(\gamma)) = \exp\left(\frac{\log(\gamma)}{1-\gamma} \log(T)\right) \sim 1/T .$$

Hence, $T\gamma^D \frac{1}{1-\gamma}$ behaves as $d^{1/4} T^{1/2} B_T^{-1/2}$ and is negligible.
Furthermore, $\log(1/\gamma) \sim d^{-1/4} B_T^{1/2} T^{-1/2}$, implying that $T \log(1/\gamma) \sim d^{-1/4} B_T^{1/2} T^{1/2}$.
As a result, it holds that, $\beta_T \sqrt{dT} \sqrt{T \log(1/\gamma) + \log\left(1 + \frac{L^2}{d\lambda(1-\gamma)}\right)}$ is equivalent to $dT^{1/2} \sqrt{\log(T/B_T)} \sqrt{d^{-1/4} B_T^{1/2} T^{1/2}} = d^{7/8} B_T^{1/4} T^{3/4} \sqrt{\log(T/B_T)}$. By adding those three terms and neglecting the log factors, we obtain the desired result. □

5 | Generalized Linear Bandits in Abruptly Changing Environments

Contextual sequential decision problems with categorical or numerical observations are common and Generalized Linear Bandits (GLB) offer a solid theoretical framework to address them. In contrast to linear bandits that were discussed in the previous chapter, existing algorithms for GLB have two drawbacks undermining their applicability. First, they rely on excessively pessimistic concentration bounds due to the non-linear nature of the model. Second, they require either non-convex projection steps or burn-in phases to enforce boundedness of the estimators. Both of these issues are worsened when considering non-stationary models, in which the GLB parameter may vary with time. In this chapter, we focus on self-concordant GLB (which include logistic and Poisson regression) with forgetting achieved either by the use of a sliding window or exponential weights. We propose a novel confidence-based algorithm for the maximum-likelihood estimator with forgetting and analyze its performance in abruptly changing environments. The results from this chapter are based on [Russac et al., 2021a].

Outline

5.1 Introduction

Generalized linear bandits (GLB) have been introduced as a generalization of linear bandits, able to describe broader reward models of considerable practical relevance, in particular binary or categorical rewards [Filippi et al., 2010, Li et al., 2017]. Generalized linear bandits are for instance a natural option in online advertising applications where the rewards take the form of clicks [Chapelle and Li, 2011]. In this chapter, we focus on deterministic algorithms and refer to [Chapelle and Li, 2011, Kveton et al., 2020] for randomized algorithms applicable to GLB. Compared to the linear bandit case, there are two distinctive drawbacks of GLB algorithms. The first is **(1)** the presence of a problem-dependent constant, imposed by the non-linear nature of the model, that is possibly *prohibitively large* and has a negative impact both on the design of algorithms and on their analysis. The second is **(2)** the need to modify the Maximum Likelihood Estimator (MLE) to ensure that it has a bounded norm. Usually this is achieved by resorting to an additional *non-convex* projection program applied to the MLE [Filippi et al., 2010]. These distinctions correspond to a fundamental difference between the models, and explain why methods developed for linear bandits may fail in the case of GLB.

The first drawback **(1)** was recently addressed by [Faury et al., 2020], in the specific case of logistic bandits. They showed that in this particular setting, the regret bounds of carefully designed algorithms could be significantly improved only at the cost of minor algorithmic modifications. Their analysis tightens the gap with the linear case, and takes a significant step towards the development of efficient GLB algorithms.

The second drawback **(2)** has seen little treatment in the literature, except for the work of [Li et al., 2017] who proved that the projection step of [Filippi et al., 2010] could be avoided by resorting to random initialization phases. However, a careful examination of the required conditions shows that these initialization phases can be prohibitively long to be deployed in scenarios of practical interest.

The aforementioned improvements to the original GLB algorithm of [Filippi et al., 2010] were developed under a stationarity assumption. However relaxing this assumption is of interest in real-world applications of contextual bandits. In the linear bandit literature, this has motivated the development of adequate algorithms, able to handle changes in the structure of the reward signal as discussed in the previous chapter. [Russac et al., 2020] generalized such approaches to GLB, but without addressing neither **(1)** nor **(2)**. As a result, the practical relevance of their approach remains questionable and the development of *efficient* and *non-stationary* GLB algorithms stands incomplete.

This chapter aims at closing this gap. We study a broad family of GLB, known as *self-concordant* (which includes for instance the logistic and Poisson bandits), in environments where the parameter is allowed to switch arbitrarily over time. Under this setting, we answer **(1)** by providing a non-trivial extension of the concentration results from [Faury et al., 2020]. We also leverage the self-concordance property to *remove* the projection step, henceforth overcoming **(2)**. This is made possible by an improved characterization of the, possibly weighted, MLE in (self-concordant) generalized linear models. Combined together, these two contributions lead to the design of *efficient* GLB algorithms, with improved regret bounds and which do not require to solve hard (i.e. non-convex) optimization programs. In doing so, we also answer the long-standing issue of providing proper confidence regions centered around the pristine MLE in GLB.

5.2 Background

5.2.1 Setting and Assumptions

We consider the same setting than in Section 1.5.1 and we recall the main ingredients here. At each time step, the environment provides a (time-dependent) action set $\mathcal{A}_t$ and the agent plays a d-dimensional action $A_t \in \mathcal{A}_t$. We will assume that the reward's distribution belongs to a *canonical exponential family* with respect to a reference measure ξ. The conditional distribution of the reward x given some feature vector a satisfies

$$\frac{d\mathbb{P}_\theta(x|a)}{d\xi} = \exp\Big(xa^\top\theta - b(a^\top\theta)\Big) \ .$$

Thanks to the properties of exponential families, b is convex and can be related to the function $\mu = \dot{b}$, itself referred to as the *inverse link* or *mean* function.

Let $\mathcal{F}_t = \sigma(\mathcal{A}_1, A_1, X_1, \ldots, \mathcal{A}_t, A_t, X_t, A_{t+1}, \mathcal{A}_{t+1})$ denote the σ field containing the information available before obtaining the reward at time $t+1$. A key feature of this description is that given a ground-truth parameter $\theta^\star$, selecting an action A_t at time t yields a reward X_t conditionally independent on the past and such that $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = \mu(A_t^\top\theta^\star)$.

The non-stationary nature of the considered environments is characterized as follows: the bandit parameter $\theta^\star$ is allowed to change in an arbitrary fashion up to Γ_T times within the horizon T. In the following, $\theta^\star$ will be indexed by t to clearly exhibit its dependency w.r.t round t, and the reward signal will follow:

$$\mathbb{E}[X_t|\mathcal{F}_{t-1}] = \mu(A_t^\top\theta_t^\star) \ .$$

The focus of this chapter is the *dynamic regret* defined as

$$R(T) = \sum_{t=1}^{T} \max_{a\in\mathcal{A}_t} \mu(a^\top\theta_t^\star) - \sum_{t=1}^{T} \mu(A_t^\top\theta_t^\star) \ .$$

We will work under the following assumptions.

Assumption 5.1 (Bounded actions and bandit parameters)**.**

$$\forall t \geq 1, \|\theta_t^\star\|_2 \leq S \quad \textit{and} \quad \forall a \in \mathcal{A}_t, \|a\|_2 \leq 1 \ .$$

We define the admissible parameter space $\Theta = \{\theta \in \mathbb{R}^d, \|\theta\|_2 \leq S\}$.

Assumption 5.2 (Bounded rewards)**.**

$$\exists m \in \mathbb{R}^+ \textit{such that} \, \forall t \geq 1, 0 \leq X_t \leq m \ .$$

Assumption 5.3. *The mean function $\mu : \mathbb{R} \mapsto \mathbb{R}$ is continuously differentiable, Lipschitz with constant k_μ and such that*

$$c_\mu = \inf_{\theta\in\Theta, \|a\|_2\leq 1} \dot{\mu}(a^\top\theta) > 0 \ .$$

The quantity c_μ is crucial in the analysis, as it represents the (worst case) sensitivity of the mean function. Our last assumption differs from most of existing works as we focus here on *self-concordant* GLMs. This assumption on the curvature of the mean function is rather mild, and covers for instance the logistic and Poisson models.

Assumption 5.4 (Generalized self-concordance)**.** *The mean function verifies* $|\ddot{\mu}| \leq \dot{\mu}$.

In order to estimate the unknown bandit parameter $\theta_t^\star$, we will adopt a *weighted* regularized maximum-likelihood principle. Formally, we define $\hat{\theta}_t$ for $\lambda > 0$ and $\gamma \in (0, 1]$ as the solution of the strictly convex program

$$\hat{\theta}_t = \operatorname{argmin}_{\theta \in \mathbb{R}^d} - \sum_{s=1}^{t} \gamma^{t-s} \log \mathbb{P}_\theta(X_s | A_s) + \frac{\lambda}{2} \|\theta\|_2^2 \,. \tag{5.1}$$

Equivalently, $\hat{\theta}_t$ may be defined as the minimizer of $-\sum_{s=1}^{t} \gamma^{-s} \log \mathbb{P}_\theta(X_s|A_s) + \frac{\lambda \gamma^{-(t-1)}}{2} \|\theta\|_2^2$, with time-independent increasing weights γ^{-s} and time-varying regularization $\lambda \gamma^{-(t-1)}$, which is more handy for analysis purposes, see Chapter 4.

5.2.2 Stationary Generalized Linear Bandits

GLB were first considered in the seminal work of [Filippi et al., 2010] who proposed GLM-UCB, an optimistic algorithm with a regret upper bound of the form $\tilde{\mathcal{O}}(c_\mu^{-1} d\sqrt{T})$. A key characteristic of GLM-UCB is a *projection step*, used to map the MLE onto the set of admissible parameters Θ. Formally, when the MLE $\hat{\theta}_t$ is not in Θ, it needs to be replaced by

$$\tilde{\theta}_t = \operatorname{argmin}_{\theta \in \Theta} \left\| \sum_{s=1}^{t} \left[\mu(A_s^\top \theta) - \mu(A_s^\top \hat{\theta}_t) \right] A_s \right\|_{V_t^{-1}} \tag{5.2}$$

where V_t is an invertible $d \times d$ square matrix.

With GLM-UCB, both the size of the confidence set (thus the exploration bonus) and the regret bound scale as c_μ^{-1}. However, this constant can be prohibitively large. In the cases of the logistic and Poisson bandits, one has $c_\mu^{-1} \geq e^S$, revealing an *exponential* dependency on S. If we consider the example of click prediction in online advertising with the logistic GLB, c_μ^{-1} is of the order 10^3, corresponding to typical click rates of less than a percent.

This critical dependency was addressed by [Faury et al., 2020] for the logistic bandit. They introduce LogUCB1 and LogUCB2 for which they respectively prove $\widetilde{\mathcal{O}}(c_\mu^{-1/2} d\sqrt{T})$ and $\widetilde{\mathcal{O}}(d\sqrt{T} + c_\mu^{-1})$ regret upper bounds. Their analysis relies on the self-concordance property of the logistic log-likelihood. Self-concordance offers a refined way to control the curvature of the log-likelihood, and has been used in batch statistical learning [Bach, 2010] and online optimization [Bach and Moulines, 2013] (see also [Boyd and Vandenberghe, 2004, Section 9.6] for a broader picture). However, the analysis of [Faury et al., 2020] does not use the self-concordance to its fullest and a projection step is still required, as detailed in Section 5.5.

Since the mean function μ can be non-convex (as for example in the case of logistic regression), the projection step defined in Equation (**P0**) generally involves the minimization of a non-convex function. Solving this program can be arduous and finding ways to bypass it is desirable. This was achieved by [Li et al., 2017] using a *burn-in phase* corresponding to an initial number of rounds during which the agent plays randomly. This ensures that $\hat{\theta}_t$ stays in Θ for subsequent rounds and therefore avoids the projection step. This technique was re-used in other recent works, such as [Kveton et al., 2020, Zhou et al., 2019]. A major drawback of this approach however is the length of this burn-in phase, which typically grows with c_μ^{-2} [Kveton et al., 2020, Section 4.5]. In the previously cited example of click-prediction, this would lead the agent to act randomly for approximately 10^6 rounds.

5.2.3 Forgetting in Non-Stationary Environments

Motivated by the non-stationary nature of most real-life applications of contextual bandits, a consequent theory for linear bandits in non-stationary environments has been recently developed as pointed out in Chapter 4. We focus here on forgetting policies, a broader perspective is discussed in Section 5.5. In [Cheung et al., 2021], a sliding window is used and the estimator is constructed based on the most recent observations only. In [Russac et al., 2019] exponentially increasing weights are used to give more importance to most recent observations. In [Zhao et al., 2020] the algorithm is restarted on a regular basis. These contributions were generalized to GLB by [Russac et al., 2020, Cheung et al., 2021, Zhao et al., 2020]. However, the approach of [Russac et al., 2020] still suffers from the aforementioned limitations (dependency w.r.t. c_μ and need for a projection step) while the analysis of both [Cheung et al., 2021] and [Zhao et al., 2020] are missing key features of the problem at hand (see [Russac et al., 2020, Section 1] and discussion in Section 6.3.2 in the next Chapter).

The non-stationary nature of the problem rules out the use of burning phases as changes in the GLB parameter can lead $\hat{\theta}_t$ to leave Θ, even when well initialized. This also accentuates the inconveniences brought by the projection step, as $\hat{\theta}_t$ leaving Θ is more likely to happen. This is why finding alternatives without projection is even more attractive in this particular setting. Furthermore, a generalization of the improvements brought by [Faury et al., 2020] to non-stationary world is missing, and it is unclear if the dependency in c_μ can still be reduced in this harder setting.

5.2.4 Contributions

The present chapter addresses these challenges, focusing on the use of exponential weights to adapt to changes in the model. First, we extend in Theorem 5.3 the Bernstein-like tail-inequality of [Faury et al., 2020, Theorem 1] to *weighted* self-normalized martingales. We then leverage the self-concordance property (Assumption 5.4) to provide an improved characterization of the maximum-likelihood estimator (Proposition 5.4). This allows to provide concentration guarantees *without* projecting $\hat{\theta}_t$ back to Θ. Combining these results leads to the SC-D-GLUCB strategy (Algorithm 11), which does not resort to a non-convex projection step and enjoys an $\tilde{\mathcal{O}}(c_\mu^{-1/3} d^{2/3} \Gamma_T^{1/3} T^{2/3})$ worst case regret upper bound (Theorem 5.2). A $\mathcal{O}(c_\mu^{-1/2} \Delta^{-1} d \sqrt{\Gamma_T T})$ regret bound is also obtained (Theorem 5.1) under an additional minimal gap $\Delta > 0$ assumption (Assumption 5.5).

A summary of our contributions and comparison with prior work are given in Table 5.1.

5.3 Algorithm and Results

5.3.1 Algorithms

In this section, we consider the abruptly changing environments defined in Section 5.2. We propose two algorithms: SC-D-GLUCB, which is based on discount factors, and SC-SW-GLUCB using a sliding window. The pseudo-code of SC-SW-GLUCB and the corresponding theoretical results are reported in Appendix 5.C. Associated with the weighed MLE defined in Equation (5.1), define the weighted design matrix as

$$V_t = \sum_{s=1}^{t} \gamma^{t-s} A_s A_s^\top + \frac{\lambda}{c_\mu} I_d \,. \tag{5.3}$$

Algorithm	Setting	Projection	Regret Bound
GLM-UCB [Filippi et al., 2010]	Stationary GLM	Non-convex	$\widetilde{\mathcal{O}}\left(\boldsymbol{c_\mu^{-1}} \cdot d \cdot \sqrt{T}\right)$
LogUCB1 [Faury et al., 2020]	Stationary Logistic	Non-convex	$\widetilde{\mathcal{O}}\left(\boldsymbol{c_\mu^{-1/2}} d\sqrt{T}\right)$
D-GLUCB [Russac et al., 2020]	Non-Stationary GLM	Non-convex	$\widetilde{\mathcal{O}}\left(\boldsymbol{c_\mu^{-1}} d^{2/3}\Gamma_T^{1/3}T^{2/3}\right)$
SC-D-GLUCB **(this chapter)**	Non-Stationary GLM + SC + Ass. 5.5	**No projection**	$\widetilde{\mathcal{O}}\left(\boldsymbol{c_\mu^{-1/2}} d\sqrt{\Gamma_T T}\right)$
SC-D-GLUCB **(this chapter)**	Non-Stationary GLM + SC	**No projection**	$\widetilde{\mathcal{O}}\left(\boldsymbol{c_\mu^{-1/3}} d^{2/3}\Gamma_T^{1/3}T^{2/3}\right)$

Table 5.1: Comparison of regret guarantees for different algorithms in the GLM setting with respect to the degree of non-linearity c_μ, the dimension d, the horizon T and the number Γ_T of abrupt changes. In the table SC stands for self-concordant. Regret guarantees for SC-SW-GLUCB are the same than for SC-D-GLUCB.

The SC-D-GLUCB algorithm proceeds as follows. First, based on the previous rewards and actions, $\hat{\theta}_t$ is computed. After receiving the action set $\mathcal{A}_t$, the action A_t is chosen optimistically as the maximizer of the current estimate $\mu(a^\top\hat{\theta}_t)$ of each arm's reward inflated by the confidence bonus $c_\mu^{-1/2}\beta_T^\delta\|a\|_{V_t^{-1}}$. Finally, the reward X_t is received and the matrix V_t is updated. The expression of β_T^δ is a consequence of our novel concentration result and is defined in Equation (5.4). A pseudo-code of the algorithm is presented in Algorithm 11.

There are two differences between SC-D-GLUCB and the algorithm from [Russac et al., 2020]. First, we directly use $\hat{\theta}_t$ to make predictions about the arms' performances, whether it belongs to Θ or not. Second, the exploration term scales as $c_\mu^{-1/2}$ (instead of c_μ^{-1}), as in [Faury et al., 2020]. The latter has a direct impact on the regret-bound of SC-D-GLUCB, to be stated below.

Input: Failure probability δ, dimension d, regularization λ, upper bound for actions L, upper bound for parameters S, discount factor γ.
Initialization: $V_0 = (\lambda/c_\mu)I_d$, $\hat{\theta}_0 = 0_{\mathbb{R}^d}$
for $t = 1$ ***to*** T **do**
  Receive $\mathcal{A}_t$, compute $\hat{\theta}_{t-1}$ according to (5.1)
  Play $A_t = \text{argmax}_{a\in\mathcal{A}_t}\mu(a^\top\hat{\theta}_{t-1}) + \frac{\beta_T^\delta}{\sqrt{c_\mu}}\|a\|_{V_{t-1}^{-1}}$ with β_T^δ defined in Equation (5.4)
  Receive reward X_t
  Update: $V_t \leftarrow A_tA_t^\top + \gamma V_{t-1} + \frac{\lambda}{c_\mu}(1-\gamma)I_d$

Algorithm 11: SC-D-GLUCB

5.3.2 Regret Upper Bounds

We detail in this section the performance guarantees for SC-D-GLUCB. Define

$$\beta_T^\delta = k_\mu\sqrt{\lambda}\left(1 + \bar{S} + \sqrt{\frac{1+\bar{S}}{\lambda}}\rho_T^\delta + \left(\frac{\rho_T^\delta}{\sqrt{\lambda}}\right)^2\right)^{3/2} \tag{5.4}$$

with

$$\bar{S} = S + \frac{2Sk_\mu + m}{T\lambda(1-\gamma)}\,, \tag{5.5}$$

and where

$$\rho_T^\delta = \frac{\sqrt{\lambda}}{2m} + \frac{2m}{\sqrt{\lambda}}\log\left(\frac{T}{\delta}\right) + \frac{2m}{\sqrt{\lambda}}d\log(2) + \frac{dm}{\sqrt{\lambda}}\log\left(1 + \frac{k_\mu(1-T^{-2})}{d\lambda(1-\gamma^2)}\right)\,.$$

The latter expression is a direct consequence of the concentration result presented in Theorem 5.3 below. The difference between $\bar{S}$ and S is a bias term due to the non-stationarity.

Before stating our first theorem, we add an additional assumption on the minimal gap. This assumption is discussed in Section 5.5 and is only used in Theorem 5.1.

Assumption 5.5. *Let $A_{t,\star} = \operatorname{argmax}_{a\in\mathcal{A}_t}\mu(a^\top\theta_t^\star)$ denote the optimal action at time t. The reward gaps $\Delta_t = \min_{a\in\mathcal{A}_t,\mu(a^\top\theta_t^\star)<\mu(A_{t,\star}^\top\theta_t^\star)}\mu(A_{t,\star}^\top\theta_t^\star) - \mu(a^\top\theta_t^\star)$ satisfies*

$$\forall t \le T, \Delta_t \ge \Delta > 0\,.$$

Theorem 5.1. *Under Assumption 5.5, the regret of the* SC-D-GLUCB *algorithm is bounded for all $\gamma \in (1/2, 1)$ with probability at least $1-\delta$ by*

$$\begin{aligned} R(T) \le\ & C_1\frac{\Gamma_T}{1-\gamma} + C_2\frac{1}{T(1-\gamma)^2\Delta} \\ & + C_3\frac{\beta_T^\delta\sqrt{dT}}{\sqrt{c_\mu}\Delta}\sqrt{T\log(1/\gamma) + \log\left(1 + \frac{1}{d\lambda(1-\gamma)}\right)} \\ & + C_4\frac{d(\beta_T^\delta)^2}{c_\mu\Delta}\left(T\log(1/\gamma) + \log\left(1 + \frac{1}{d\lambda(1-\gamma)}\right)\right)\,, \end{aligned}$$

where C_1, C_2, C_3, C_4 are universal constants independent of c_μ, γ with only logarithmic terms in T.

In particular, setting $\gamma = 1 - \frac{\sqrt{c_\mu\Gamma_T}}{d\sqrt{T}}$ and $\lambda = d\log(T)$ leads to

$$R(T) = \widetilde{\mathcal{O}}(\Delta^{-1}c_\mu^{-1/2}d\sqrt{\Gamma_T T})\,.$$

There is a strong link between the cost of non-stationarity in the K-arm setting and the one observed in the more general GLB setting. In the K-arm setting, any sub-optimal arm i is played at most $\mathcal{O}(\Delta_i^{-2}\log(T))$ times (e.g [Munos, 2014, Proposition 1.1]), whereas in any abruptly changing environment, forgetting policies play a sub-optimal arm i at most $\widetilde{\mathcal{O}}((\Delta_T(i))^{-2}\sqrt{\Gamma_T T})$ [Garivier and Moulines, 2011]. $\Delta_T(i)$ is the minimum distance between the mean of the optimal arm and the mean of the suboptimal arm i over the entire time horizon. For GLBs, in the stationary case [Filippi et al., 2010, Theorem 1] give a gap-dependent bound on the regret scaling as $\mathcal{O}(\Delta^{-1}c_\mu^{-2}d^2\log(T))$. Here, the bound of Theorem 5.1 is of order $\mathcal{O}(\Delta^{-1}c_\mu^{-1/2}d\sqrt{\Gamma_T T})$. The reduced dependency in c_μ in the latter bound is a direct consequence of the use of self-concordance. Also note that when the inverse link function is the identity and the action set is the canonical basis, our analysis recovers the results of [Garivier and Moulines, 2011].

We give an upper bound for the worst case regret of Algorithm 11 in the following theorem; its proof is deferred to the appendix.

Theorem 5.2. *The regret of the* SC-D-GLUCB *algorithm is bounded for all* $\gamma \in (1/2, 1)$ *with probability at least* $1 - \delta$ *by*

$$R(T) \leq C_1 \frac{\Gamma_T}{1-\gamma} + C_2 \frac{\beta_T^\delta \sqrt{dT}}{\sqrt{c_\mu}} \sqrt{T \log\left(\frac{1}{\gamma}\right) + \log\left(1 + \frac{1}{d\lambda(1-\gamma)}\right)} \,,$$

where C_1 *and* C_2 *are universal constants independent of* c_μ *and* γ *with only logarithmic terms in* T*.*

In particular, setting $\gamma = 1 - \left(\frac{c_\mu^{1/2}\Gamma_T}{dT}\right)^{2/3}$ *and* $\lambda = d\log(T)$ *leads to*

$$R(T) = \widetilde{\mathcal{O}}(c_\mu^{-1/3} d^{2/3} \Gamma_T^{1/3} T^{2/3}) \,.$$

As in the linear case (see Chapter 4), this regret bound highlights the existence of two mechanisms of different nature. The first term is due to non-stationarity, the number of changes Γ_T being multiplied by $1/(1-\gamma)$, which is a rough measure of the forgetting time induced by the exponential weights. The second term characterizes the rate at which the weighted MLE $\hat{\theta}_t$ approaches $\theta_t^\star$. By balancing both terms, we can characterize the asymptotic behavior of the regret bound.

In Theorem 5.2, optimally tuning γ yields the asymptotic worst case rate of $T^{2/3}$. This is similar to the asymptotic rate achievable in the linear case with a different measure of non-stationarity (Corollary 4.4) the same dependency is attained with a sliding window for MDPs in abruptly changing environments [Gajane et al., 2018] and with restart factors [Auer et al., 2008].

Remark 5.1. *The proof of Theorem 5.2 reveals that for rounds* t *where* $\hat{\theta}_t$ *lies in* Θ*, it is possible to obtain a (usually) tighter concentration result (depending on the values of* λ *and* S*) by replacing* β_T^δ *with* $k_\mu \sqrt{1 + 2S}(\sqrt{\lambda}S + \rho_T^\delta)$*. This cannot be used to improve the result of Theorem 5.2, as one doesn't know in advance for which rounds the condition will be satisfied, but this minor modification of Algorithm 11 is most often advisable in practice. See Section 5.B.4 in Appendix for more details.*

5.4 Key Arguments

In this section, we detail some key elements of our analysis. First, we describe the concentration result in its most generic form. Then, we explain the main steps to derive the upper bound of the regret of SC-D-GLUCB.

5.4.1 A Tail-Inequality for Self-Normalized Weighted Martingales

To reduce the dependency in c_μ, it is essential to take into account the actual conditional variance of the generalized linear model [Faury et al., 2020]. With exponentially increasing weights, we also need time-dependent regularization parameters to avoid a vanishing effect of the regularization [Russac et al., 2019]. Carefully combining these two elements yields the following concentration result.

Theorem 5.3. *Let t be a fixed time instant. Let $\{\mathcal{F}_u\}_{u=0}^t$ be a filtration. Let $\{A_u\}_{u=0}^t$ be a stochastic process on $\mathbb{R}^d$ such that A_u is $\mathcal{F}_{u-1}$ measurable and $\|A_u\|_2 \leq 1$. Let $\{\epsilon_u\}_{u=0}^t$ be a martingale difference sequence such that ϵ_u is $\mathcal{F}_u$ measurable. Assume that the weights are non-decreasing, strictly positive and the time horizon is known. Furthermore, assume that conditionally on $\mathcal{F}_u$ we have $|\epsilon_u| \leq m$ a.s. Let $\{\lambda_u\}_{u=1}^t$ be a deterministic sequence of regularization terms and denote $\sigma_t^2 = \mathbb{E}\left[\epsilon_t^2 | \mathcal{F}_{t-1}\right]$. Let $\widetilde{H}_t = \sum_{s=1}^t w_s^2 \sigma_s^2 A_s A_s^\top + \lambda_t I_d$ and $S_t = \sum_{s=1}^t w_s \epsilon_s A_s$, then for any $\delta \in (0,1]$,*

$$\mathbb{P}\left(\|S_t\|_{\widetilde{H}_t^{-1}} \geq \frac{\sqrt{\lambda_t}}{2mw_t} + \frac{2mw_t}{\sqrt{\lambda_t}} \log\left(\frac{\det(\widetilde{H}_t)^{1/2}}{\delta \lambda_t^{d/2}} \right) + \frac{2mw_t}{\sqrt{\lambda_t}} d\log(2) \right) \leq \delta \, .$$

5.4.2 Upper Bounding the Regret of SC-D-GLUCB

In a non-stationary environment, each change in the parameter will necessarily result in a number of rounds where the bias of the weighted MLE estimator cannot be controlled. This gives rise to the first term in the upper bound in Theorem 5.2. To make this observation more explicit, for $D \geq 1$, define $\mathcal{T}(\gamma) = \{1 \leq t \leq T, \text{ such that } \theta_s^\star = \theta_t^\star \text{ for } t - D + 1 \leq s \leq t\}$ the set of time instants that are at least D steps away from the previous closest breakpoint. Central in the analysis of weighted GLBs is the matrix

$$G_t(\hat{\theta}_t, \theta_t^\star) = \sum_{s=1}^t \gamma^{t-s} \alpha(A_s, \hat{\theta}_t, \theta_t^\star) A_s A_s^\top + \lambda I_d \, ,$$

where

$$\alpha(A_s, \hat{\theta}_t, \theta_t^\star) = \int_0^1 \dot{\mu}(A_s^\top((1-v)\theta_t^\star + v\hat{\theta}_t))dv \, .$$

As in the linear case, we define its analogue with squared exponential weights,

$$\widetilde{G}_t(\hat{\theta}_t, \theta_t^\star) = \sum_{s=1}^t \gamma^{2(t-s)} \alpha(A_s, \hat{\theta}_t, \theta_t^\star) A_s A_s^\top + \lambda I_d \, .$$

We add the subscript $t-D:t$ to a quantity when the sum is for time instants between $t-D+1$ and t. In this subsection, for space constraints, we will denote equivalently $\widetilde{G}_t(\hat{\theta}_t, \theta_t^\star)$ (resp. $G_t(\hat{\theta}_t, \theta_t^\star)$) by $\widetilde{G}_t$ (resp. G_t). As for linear bandits, the exploration bonus is designed to mitigate the impact of prediction errors. We focus below on upper bounding the prediction error in $\hat{\theta}_t$ defined as $\Delta_t(a, \hat{\theta}_t) = |\mu(a^\top \hat{\theta}_t) - \mu(a^\top \theta_t^\star)|$. The exact link between the regret and this quantity is made explicit in Proposition 5.32 in Appendix. By defining $g_t(\theta) = \sum_{s=t-D+1}^t \gamma^{t-s} \mu(A_s^\top \theta) A_s + \lambda\theta$, when $t \in \mathcal{T}(\gamma)$ one can upper bound the prediction error in $\hat{\theta}_t$.

$$\Delta_t(a, \hat{\theta}_t) \leq \frac{c\gamma^D}{1-\gamma} + k_\mu \underbrace{\|g_t(\hat{\theta}_t) - g_t(\theta_t^\star)\|_{\widetilde{G}_{t-D:t}^{-1}}}_{①} \underbrace{\|a\|_{G_t^{-1}}}_{②} \, .$$

The first term corresponds to the bias due to non-stationarity. ① is a measure of the deviation of $\hat{\theta}_t$ from $\theta_t^\star$ adapted to the non-linear nature of the problem. Note that $g_t(\hat{\theta}_t) - g_t(\theta_t^\star)$ involves a martingale difference sequence (thanks to the optimality condition of the MLE) that can be controlled using Theorem 5.3. However, to bound ① using Theorem 5.3 one needs to link the matrix $\widetilde{G}_{t-D:t}$ with $\widetilde{H}_{t-D:t}$, the self-concordance allows exactly to do this.

Self-Concordance. More precisely, the use of self-concordance offers a sharp relation (independent of c_μ) between the first derivative of the mean function evaluated at different points. Using Lemma 5.23 reported in Appendix 5.D, standard calculations yield:

$$\widetilde{G}_{t-D:t} \geq \left(1 + C + \frac{1}{\sqrt{\lambda}} \|g_t(\hat{\theta}_t) - g_t(\theta_t^\star)\|_{\widetilde{G}_{t-D:t}^{-1}}\right) \widetilde{H}_{t-D:t} \,. \tag{5.6}$$

Note that Equation (5.6) involves the deviation term that we want to control. Here, C is a residual bias due to the non-stationarity of the environment.

Better Characterization of the MLE. By leveraging Equation (5.6) to bound the deviation $g_t(\hat{\theta}_t) - g_t(\theta_t^\star)$ in the $\widetilde{G}_{t-D:t}^{-1}$-norm, one obtains an implicit equation. Solving it leads to the following proposition.

Proposition 5.4. *When $t \in \mathcal{T}(\gamma)$, the following holds,*

$$\|g_t(\hat{\theta}_t) - g_t(\theta_t^\star)\|_{\widetilde{G}_{t-D:t}^{-1}(\hat{\theta}_t, \theta_t^\star)} \leq \sqrt{1+C}\rho_T^\delta + \frac{1}{\sqrt{\lambda}} \left(\rho_T^\delta\right)^2 \,,$$

where C is a residual term due to non-stationarity.

Remark 5.2. *In stark contrast with previously existing works (see [Filippi et al., 2010, Proposition 1]), deviations from the true parameter $\theta_t^\star$ are characterized uniquely by the MLE (and not by its projected counterpart). This can be done whether $\hat{\theta}_t$ belongs to Θ or not and without any projection. This is not specific to the non-stationary nature of the problem but fundamentally relies on an improved analysis of the MLE. Similar guarantees can be obtained in any stationary environment. See Section 5.5 for a more detailed comparison of the possible uses of the self-concordance property.*

① can be upper bounded using Proposition 5.4. To upper bound ② we use the following inequality.

$$G_t \geq \left(1 + C + \frac{1}{\sqrt{\lambda}} \|g_t(\hat{\theta}_t) - g_t(\theta_t^\star)\|_{\widetilde{G}_{t-D:t}^{-1}}\right)^{-1} c_\mu V_t \,. \tag{5.7}$$

Combining Proposition 5.4 with Equation (5.7) gives the upper bound for ②. Putting everything together, we obtain the form of β_T^δ given in Equation 5.4. The regret bound is then obtained by summing the exploration bonus for the different time instants. Applying the so-called elliptical lemma (see [Lattimore and Szepesvári, 2020, Chap. 19]) and letting $D = \log(T)/\log(1/\gamma)$ completes the proof.

5.5 Discussion

Assumption on the Gaps. Assumptions similar to our Assumption 5.5 requiring a minimum gap are frequent in non-stationary bandits. First, note that Δ is not required for the algorithm but only for the theoretical analysis. Second, similar assumptions can be found for K-arm bandits in several works to obtain the optimal $\widetilde{\mathcal{O}}(\sqrt{\Gamma_T T})$ regret bound. This is in particular the case for change-points detection methods: [Cao et al., 2019, Corollary 1] and [Zhou et al., 2020, Corollary 4.3] is proved under an assumption on the minimal gap. This remains true for forgetting strategies: the bound of [Garivier and Moulines, 2011] is gap-dependent, [Trovo et al., 2020]

achieve a $\mathcal{O}(\Delta^{-1}\sqrt{T\Gamma_T})$ regret. More demanding, the LM-DSEE and SW-UCB# algorithms from [Wei and Srivatsva, 2018] require the minimum gap as an input of the algorithm. Generally speaking, none of those works provide an analysis when the minimum gap can depend on the time horizon T and when the mean of different arms can be arbitrarily close. We suspect that forgetting policies would obtain a $\mathcal{O}(\Gamma_T^{1/3}T^{2/3})$ worst case dependency as in Theorem 5.2 and that changepoint detection methods are likely to fail in such a case.

Tightness of the Bound. For problems with a finite number of actions, [Auer et al., 2018] have developed an algorithm that does not require the knowledge of the number of breakpoints nor assumption on the gaps. This was extended to the K-arm setting by [Auer et al., 2019] and to the more general contextual bandits by [Chen et al., 2019]. Both works ([Auer et al., 2019, Chen et al., 2019]) achieve the optimal $\widetilde{\mathcal{O}}(\sqrt{\Gamma_T T})$ regret bound. Yet, their analysis does not apply to the GLB framework. Furthermore, both works rely on replaying phases that are incompatible with time-dependent action sets as considered here. Additionally, in [Chen et al., 2019] the regret is defined with respect to the best policy in some finite class, whereas our results apply to the general setting where actions can change over time and the regret benchmark is the ground-truth of the environment. The best lower-bound for forgetting policies in abruptly changing environments with time-dependent action sets remains unknown. While it is known that forgetting policies are minimax optimal when non-stationarity is measured through the so-called variational budget and adding structure on the action-sets, whether such methods are optimal in abruptly changing environments is unclear. Nonetheless, the bound obtained by [Garivier and Moulines, 2011] in the K-arm setting yields a worst case regret bound that can be shown to be of order $\mathcal{O}(\Gamma_T^{1/3}T^{2/3})$ (see appendix Section 5.E).

Knowledge of Γ_T. Optimizing the choice of the forgetting parameter γ (w.r.t. the regret bound) requires the knowledge of Γ_T. The Bandit over Bandit (BOB) framework introduced by [Cheung et al., 2019] can be used to circumvent this requirement. When the assumption 5.5 is satisfied, following the proof from [Cheung et al., 2021] one would obtain a regret bound of order $\widetilde{\mathcal{O}}(\Delta^{-1}dc_\mu^{-1/2}\sqrt{T\max(\Gamma_T, T^{1/2})})$ (see [Auer et al., 2019, Remark 2]). Similarly, in the absence of Assumption 5.5 an upper bound of order $\widetilde{\mathcal{O}}(c_\mu^{-1/3}d^{2/3}T^{2/3}\max(\Gamma_T, d^{-1/2}T^{1/4})^{1/3})$ can be achieved (see [Zhao et al., 2020, Theorem 4]).

Self-Concordance. The analysis of [Faury et al., 2020] does not use self-concordance to its fullest. We present an improved analysis valid in any stationary time frame, proving that a better treatment of the self-concordance removes the need for the inconvenient projection. Informally, the self-concordance links $\mu(a^\top\hat{\theta}_t)$ to $\mu(a^\top\theta^\star)$ without resorting to global bounds on $\dot{\mu}$ (e.g k_μ and c_μ). In [Faury et al., 2020], this takes the form of a Taylor-like expansion:

$$\mu(a^\top\theta_t) \leq \mu(a^\top\theta^\star) + \frac{|a^\top(\theta^\star - \theta_t)|}{1+2S}\dot{\mu}(a^\top\theta^\star)\ ,$$

where θ_t is a projected version of $\hat{\theta}_t$ in Θ. The denominator of the r.h.s. is reminiscent of this projection step. We show here that a finer analysis yields the following, more implicit but powerful bound:

$$\mu(a^\top\hat{\theta}_t) \leq \mu(a^\top\theta^\star) + \frac{|a^\top(\theta^\star - \hat{\theta}_t)|}{1+|a^\top(\theta^\star - \hat{\theta}_t)|}\dot{\mu}(a^\top\theta^\star)\ .$$

We illustrate the Figure 5.1 the difference between the two bounds in the logistic case with scalar values.

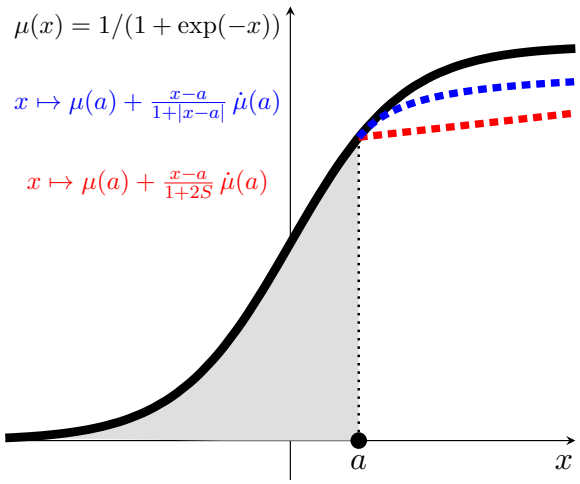


Figure 5.1: Illustration of the tighter bound that we use in the logistic case when $d = 1$

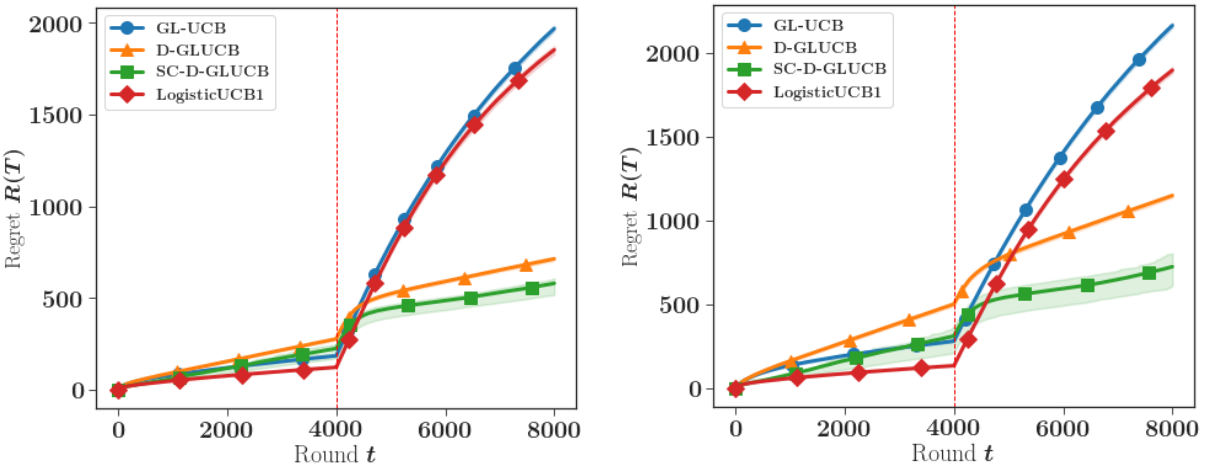
Note that when $\hat{\theta}_t \in \Theta$ (i.e there is no need for a projection), our bound implies the one of [Faury et al., 2020]. The kind of relationship displayed in the above equation allows us to derive a tail inequality for the deviation from $\hat{\theta}_t$ to $\theta^\star$ without projecting $\hat{\theta}_t$, by solving an implicit equation. We believe that this new approach is of interest in other settings involving self-concordant GLBs. The self-concordance assumption (Assumption 5.4) is not particularly restrictive and goes beyond logistic functions. Under the classical Assumption 5.1 (i.e. bounded features) all GLMs are self-concordant (cf. Sec. 2 of [Bach, 2014]) with constants that depend on the link function.

5.6 Experiments

In this section, we illustrate the empirical performance of SC-D-GLUCB in a simulated, abruptly changing environment with a logistic link function $\mu(x) = 1/(1+\exp(-x))$. In this two-dimensional problem, there is a switch in the reward distribution at $t = 4000$ (red dashed line on Figure 6.2).

SC-D-GLUCB (Algorithm 11) is compared with GLM-UCB from [Filippi et al., 2010], LogUCB1 from [Faury et al., 2020] and with D-GLUCB from [Russac et al., 2020]. SC-D-GLUCB (resp. D-GLUCB) is related with LogUCB1 (resp. GLM-UCB) in the sense that the exploration terms have the same scaling but the former incorporate the exponential weights making it possible to adapt to changes. The average regret of the different policies together with their central 50% quantiles, averaged on 200 independent runs, are reported in Figure 6.2 for two different parameter values.

In Fig. 6.2a, $\theta^\star$ starts on the circle of radius $S = 6$ (corresponding to $c_\mu^{-1} = \exp(S) \approx 400$) with an angle of $2\pi/3$ and jumps at $t = 4000$ to an angle of $4\pi/3$. The experiment reported on Fig. 6.2b is identical with a radius $S = 7$ corresponding to a $c_\mu^{-1} \approx 1000$. As previously discussed, using such values of S is required in situation where the actions return binary rewards with expected values in the range 10^{-3} – 10^{-2}, which is typically the case in web advertising or recommendation applications.



(a) $c_\mu^{-1} = 400$ (b) $c_\mu^{-1} = 1000$

Figure 5.2: Regret of the different algorithms in a 2D abruptly changing environment averaged on 200 independent experiments and the 25% associated quantiles.

For both experiments, at every time steps, 50 randomly generated actions in the unit circle are proposed to the learner. For SC-D-GLUCB and D-GLUCB the asymptotically optimal choice of the discount factors is used: $\gamma = 1 - (\Gamma_T/(d \times T))^{2/3}$ with $d = 2$, $\Gamma_T = 2$ and $T = 8000$. To speed up the learning that is hard with those values of c_μ, all the algorithms have their exploration bonus divided by 5.

As expected, the algorithms tuned for non stationary situations (SC-D-GLUCB, D-GLUCB) perform worse than their stationary counterparts (LogUCB1 and GLM-UCB) during the first stationary phase. More precisely, with the choice made for γ the estimation of $\hat{\theta}_t$ for algorithms that use exponential weights is roughly based on the $1/(1-\gamma) \approx 400$ most recent observations. In contrast, LogUCB1 and GLM-UCB use all the observations from the start to compute the MLE, which eventually leads to a more precise estimation. Right after the change, the bias caused by the non-stationarity results in a significant increase in regret. Unweighted algorithms are affected much more deeply by this phenomenon that will eventually cause large losses in performance due to the persistence of obsolete information.

The theoretical analysis of Section 5.3.2 suggests that the advantage of SC-D-GLUCB is all the more significant in strongly non-linear (large c_μ^{-1}) non-stationary environments. This is obvious in Figure 6.2, particularly when comparing Fig. 6.2a and Fig. 6.2b, which differ by the range on which the logistic function is used for making reward predictions. Note that, on average, for these two simulated scenarios the fact that the MLE $\hat{\theta}_t$ does not belong to Θ happens for several hundred of rounds. All the algorithms except SC-D-GLUCB would require non convex projection steps at these instants, or equivalently, one should inflate S (and thus c_μ^{-1}) to ensure the compliance of these algorithms with the associated theory. In producing Figure 6.2, this projection step was simply bypassed, which provides an optimistic evaluation of the performance of the competitors of SC-D-GLUCB. Interestingly, the observation that the dispersion of performance of SC-D-GLUCB is slightly higher than that of D-GLUCB can be traced back to the use of Remark 5.1 in these simulations: SC-D-GLUCB adapts to the events $\{\hat{\theta}_t \notin \Theta\}$ (rather than pretending that these did not happen) and thus its performance is made somewhat dependent on the actual occurrence of these events.

5.7 Conclusion

In this chapter, we considered self-concordant generalized linear bandits in abruptly changing environments. Thanks to the self-concordance assumption, we were able to remove the projection step that was required for obtaining estimators in the admissible space. In doing so, we proposed an improved characterization of the weighted maximum likelihood estimator. Furthermore, we extended an existing concentration inequality tailored to self-concordant generalized linear models to more general estimators. Thanks to this concentration inequality, we obtained a reduced dependency in a problem-dependent constant coming from the non-linearity of the model. Studying generalized linear bandits in more general drifting environments is the topic of the next chapter.

Appendix

The Appendix is structured as follows. In Section 5.A, our new concentration result for self-normalized weighted martingales with time dependent regularization parameters is presented. In Section 5.A.3, similar concentration results are established when a sliding window is used. Section 5.B studies the regret with discount factors through our improved characterization of the MLE. Section 5.C gives similar results with a sliding window. Section 5.D gathers some technical results, in particular the main properties resulting from the self-concordance assumption. Finally in Section 5.E, a worst case bound for a sliding window policy in the K-arm setting is presented.

Appendix 5.A Tail-inequality for Self-normalized Weighted Martingales

While keeping in mind our objective of obtaining a deviation inequality with exponentially increasing weights, we give more generic results under two assumptions on the weights.

Assumption 5.6. *The time horizon T is known in advance.*

Assumption 5.7. *The weights are deterministic, strictly positive and non-decreasing, i.e,*

$$\forall 1 \leq t \leq T, 0 < w_0 \leq w_t \leq w_{t+1} \leq w_T \ .$$

We recall the statement of the corresponding concentration result.

Theorem 5.3. *Let t be a fixed time instant. Let $\{\mathcal{F}_u\}_{u=0}^t$ be a filtration. Let $\{A_u\}_{u=0}^t$ be a stochastic process on $\mathbb{R}^d$ such that A_u is $\mathcal{F}_{u-1}$ measurable and $\|A_u\|_2 \leq 1$. Let $\{\epsilon_u\}_{u=0}^t$ be a martingale difference sequence such that ϵ_u is $\mathcal{F}_u$ measurable. Assume that the weights are non-decreasing, strictly positive and the time horizon is known. Furthermore, assume that conditionally on $\mathcal{F}_u$ we have $|\epsilon_u| \leq m$ a.s. Let $\{\lambda_u\}_{u=1}^t$ be a deterministic sequence of regularization terms and denote $\sigma_t^2 = \mathbb{E}\left[\epsilon_t^2 | \mathcal{F}_{t-1}\right]$. Let $\widetilde{H}_t = \sum_{s=1}^t w_s^2 \sigma_s^2 A_s A_s^\top + \lambda_t I_d$ and $S_t = \sum_{s=1}^t w_s \epsilon_s A_s$, then for any $\delta \in (0,1]$,*

$$\mathbb{P}\left(\|S_t\|_{\widetilde{H}_t^{-1}} \geq \frac{\sqrt{\lambda_t}}{2mw_t} + \frac{2mw_t}{\sqrt{\lambda_t}} \log\left(\frac{\det(\widetilde{H}_t)^{1/2}}{\delta \lambda_t^{d/2}} \right) + \frac{2mw_t}{\sqrt{\lambda_t}} d\log(2) \right) \leq \delta \ .$$

Theorem 5.3 is a non-trivial extension of [Faury et al., 2020, Theorem 1] allowing for the use of time-dependent regularization parameters and weights. We now state several lemmas that are useful for establishing Theorem 5.3.

5.A.1 Useful Lemmas

As a first step **we fix a time instant** t. Let $M_u^t(\xi)$ for $\xi \in \mathbb{R}^d$ and $0 \leq u \leq t$ be defined as

$$M_u^t(\xi) = \exp\left(\frac{1}{mw_t} \xi^\top S_u - \frac{1}{m^2 w_t^2} \xi^\top \widetilde{H}_u(0) \xi \right) \ , \tag{5.8}$$

with $S_u = \sum_{s=1}^u w_s \epsilon_s A_s$ and $\widetilde{H}_u(0) = \sum_{s=1}^u w_s^2 \sigma_s^2 A_s A_s^\top$ where $\sigma_s^2 = \mathbb{E}[\epsilon_s^2 | \mathcal{F}_{s-1}]$.

We prefer the notation M_u^t to M_u to clearly indicate the dependency on the weight w_{t-1}. When $u = t$, we prefer the notation M_t to M_t^t. For the entire appendix, we use the notation $\mathcal{B}_2(d) = \{a \in \mathbb{R}^d, \|a\|_2 \leq 1\}$.

Lemma 5.5. *For all $\xi \in \mathcal{B}_2(d)$ and $1 \leq u \leq t$, under Assumption 5.6 and 5.7, we have*

$$\mathbb{E}\left[M_u^t(\xi)|\mathcal{F}_{u-1}\right] \leq M_{u-1}^t(\xi)\ . \quad a.s.$$

Proof.

$$\mathbb{E}\left[M_u^t(\xi)|\mathcal{F}_{u-1}\right] = M_{u-1}^t(\xi)\exp\left(-\frac{1}{m^2w_t^2}\xi^\top w_u^2\sigma_u^2 A_u A_u^\top \xi\right) \times \mathbb{E}\left[\exp\left(\frac{1}{mw_t}\xi^\top w_u\epsilon_u a_u\right)|\mathcal{F}_{u-1}\right]\ .$$

The equality holds because A_u is $\mathcal{F}_{u-1}$ measurable and ϵ_{u-1} is $\mathcal{F}_{u-1}$ measurable. With $\tilde{\epsilon}_u = \epsilon_u/m$ and $v = \frac{w_u}{w_t}\xi^\top A_u$, the conditions of Lemma 5.7 (stated below) are met and:

$$\mathbb{E}\left[\exp\left(\frac{1}{mw_t}\xi^\top w_u\epsilon_u A_u\right)|\mathcal{F}_{u-1}\right] = \mathbb{E}\left[\exp(v\tilde{\epsilon}_u)|\mathcal{F}_{u-1}\right] \leq 1 + \frac{v^2}{m^2}\sigma_u^2\ .$$

$|v| \leq 1$ holds because of Assumption 5.7 and both ξ and $A_{u-1} \in \mathcal{B}_2(d)$. Therefore,

$$\begin{aligned}\mathbb{E}\left[M_u^t(\xi)|\mathcal{F}_{u-1}\right] &\leq M_{u-1}^t(\xi)\exp\left(-\frac{1}{m^2w_t^2}\xi^\top w_u^2\sigma_u^2 A_u A_u^\top\xi\right) \times \left(1 + \frac{w_u^2}{m^2w_t^2}\sigma_u^2\xi^\top A_u A_u^\top\xi\right)\\ &\leq M_{u-1}^t(\xi) \quad \text{(a.s.)}\ ,\end{aligned}$$

where the last inequality uses $1 + x \leq \exp(x)$. □

Hence, for all $0 \leq u \leq t$ and $\xi \in \mathcal{B}_2(d)$, $\mathbb{E}\left[M_t(\xi)\right] \leq \mathbb{E}\left[M_u^t(\xi)\right] \leq \mathbb{E}\left[M_0^t(\xi)\right] = 1$.

For $0 \leq u \leq t$ we define,

$$\bar{M}_u^t = \int_\xi M_u^t(\xi)dh_u(\xi)\ . \tag{5.9}$$

Here, h_u is the density of an isotropic normal distribution of precision $\frac{2\lambda_u}{m^2w_t^2}$ truncated on $\mathcal{B}_2(d)$. We will denote $N(h_u)$ its normalization constant.

Lemma 5.6. *Let t be a fixed time instant, for all $0 \leq u \leq t$, under assumptions 5.6 and 5.7, with $\{h_u\}_{u=1}^t$ the density of an isotropic normal distribution of precision $\frac{2\lambda_u}{m^2w_t^2}$ truncated on $\mathcal{B}_2(d)$ we have,*

$$\mathbb{E}\left[\bar{M}_u^t\right] \leq 1\ .$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\bar{M}_u^t\right] &= \int_\Omega \bar{M}_u^t d\mathbb{P}(w) = \int_\Omega\left(\int_{\mathbb{R}^d} M_u^t(\xi)dh_u(\xi)\right)d\mathbb{P}(w)\\ &\leq \int_{\mathbb{R}^d}\left(\int_\Omega M_u^t(\xi)d\mathbb{P}(w)\right)dh_u(\xi) \quad \text{(Fubini)}\\ &\leq \int_{\mathbb{R}^d}\left(\int_\Omega 1 d\mathbb{P}(w)\right)dh_u(\xi) \quad \text{(Lemma 5.5 + } h_u \text{ defined on } \mathcal{B}_2(d))\\ &\leq \int_{\mathbb{R}^d} dh_u(\xi) = 1\ . \quad (h_u \text{ is a probability density function})\end{aligned}$$

□

Remark 5.3. *Allowing time-dependent regularization parameters is essential in our analysis to avoid the vanishing effect of the regularization with exponentially increasing weights for example. This is a fundamental difference with the deviation result provided in [Faury et al., 2020]. Furthermore, allowing the regularization parameters to be time-dependent comes at a cost here, we loose the property* $\mathbb{E}\left[\bar{M}_u^t | \mathcal{F}_{u-1}\right] \leq \bar{M}_{u-1}^t$ *that would hold with a fixed regularization parameter (as in [Faury et al., 2020]). In the linear bandit setting, this issue was discussed in Lemma 2 in [Russac et al., 2019].*

In particular, applying Lemma 5.6 for $u = t$ gives,

$$\mathbb{E}\left[\bar{M}_t\right] = \mathbb{E}\left[\bar{M}_t^t\right] \leq 1 \,. \tag{5.10}$$

Lemma 5.7 (Lemma 7 of [Faury et al., 2020])**.** *Let* ε *be a centered random variable of variance* σ^2 *and such that* $|\varepsilon| \leq 1$ *almost surely. Then for all* $v \in [-1, 1]$*,*

$$\mathbb{E}\left[\exp(v\varepsilon)\right] \leq 1 + v^2\sigma^2 \,.$$

Remark 5.4. *We stress out that* $v \in [-1, 1]$ *is required for Lemma 5.7 to hold. It has strong consequences in our setting with the weights as the normalization* $1/w_t$ *and* $1/w_t^2$ *in the definition of* M_u^t *are needed to ensure that* $v = (w_u/w_t)\xi^\top A_u$ *that appears in the proof of Lemma 5.5 will be smaller than 1. As a consequence, the stopping trick presented in [Abbasi-Yadkori et al., 2011] can not be applied to* $\bar{M}_u^t$ *because of its dependency on* t*. For this reason, the deviation result presented in Theorem 5.3 is only valid for a fixed time instant* t*. To obtain a deviation result on the entire trajectory a union bound is required.*

5.A.2 Proof of Theorem 5.3

The proof of this theorem follows the line of proof of [Faury et al., 2020]. The main differences are the time-dependent regularization parameters and the presence of weights. We recall that in Equation (5.9) h_t is the density of an isotropic normal distribution of precision $\frac{2\lambda_t}{m^2 w_t^2}$ truncated on $\mathcal{B}_2(d)$ and denote $N(h_t)$ its normalization constant.

The following holds,

$$\bar{M}_t = \frac{1}{N(h_t)} \int_{\mathbb{R}^d} \mathbb{1}\left[\xi \in \mathcal{B}_2(d)\right] \exp\left(\frac{1}{mw_t}\xi^\top S_t - \frac{1}{m^2 w_t^2}\xi^\top \widetilde{H}_t \xi\right) d\xi \,. \tag{5.11}$$

Let $f_t : \mathbb{R}^d \mapsto \mathbb{R}$ be defined as $f_t(\xi) = \frac{1}{mw_t}\xi^\top S_t - \frac{1}{m^2 w_t^2}\xi^\top \widetilde{H}_t \xi$. As a quadratic function, f_t can be rewritten for $\xi^\star = \operatorname{argmax}_{\|\xi\|_2 \leq 1/2} f_t(\xi)$,

$$f_t(\xi) = f_t(\xi^\star) + \nabla f_t(\xi^\star)^\top (\xi - \xi^\star) + \frac{1}{2}(\xi - \xi^\star)^\top \nabla^2 f_t(\xi^\star)(\xi - \xi^\star) \,.$$

Using $\forall \xi \in \mathcal{B}_2(d),\ \nabla^2 f_t(\xi) = -\frac{2}{m^2 w_t^2}\widetilde{H}_t$,

$$\begin{aligned}
\bar{M}_t &= \frac{e^{f_t(\xi^\star)}}{N(h_t)} \int_{\mathbb{R}^d} \mathbb{1}\left[\|\xi\|_2 \le 1\right] \exp\left(\nabla f_t(\xi^\star)^\top(\xi - \xi^\star) - \frac{1}{m^2 w_t^2}\|\xi - \xi^\star\|^2_{\widetilde{H}_t}\right) d\xi \\
&= \frac{e^{f_t(\xi^\star)}}{N(h_t)} \int_{\mathbb{R}^d} \mathbb{1}\left[\|\xi + \xi^\star\|_2 \le 1\right] \exp\left(\nabla f_t(\xi^\star)^\top \xi - \frac{1}{m^2 w_t^2}\|\xi\|^2_{\widetilde{H}_t}\right) d\xi \\
&\ge \frac{e^{f_t(\xi^\star)}}{N(h_t)} \int_{\mathbb{R}^d} \mathbb{1}\left[\|\xi\|_2 \le 1/2\right] \exp\left(\nabla f_t(\xi^\star)^\top \xi - \frac{1}{m^2 w_t^2}\|\xi\|^2_{\widetilde{H}_t}\right) d\xi \\
&\ge \frac{e^{f_t(\xi^\star)} N(g_t)}{N(h_t)} \mathbb{E}_{\xi \sim e_t}\left[\exp\left(\nabla f_t(\xi^\star)^\top \xi\right)\right] .
\end{aligned}$$

The second equality is obtained after a change of variable $\xi \mapsto \xi - \xi^\star$. In the last inequality, e_t is the density of a d-dimensional normal distribution with precision matrix $\frac{2}{m^2 w_t^2}\widetilde{H}_t$ truncated on $\{a \in \mathbb{R}^d, \|a\|_2 \le 1/2\}$.

$$\bar{M}_t \ge \frac{e^{f_t(\xi^\star)} N(e_t)}{N(h_t)} \exp\left(\mathbb{E}_{\xi \sim e_t}\left[\nabla f_t(\xi^\star)^\top \xi\right]\right) . \quad \text{(Jensen's inequality)}$$

e_t is symmetric which implies $\mathbb{E}_{\xi \sim e_t}[\xi] = 0$. Hence,

$$\bar{M}_t \ge \frac{e^{f_t(\xi^\star)} N(e_t)}{N(h_t)} . \tag{5.12}$$

Therefore,

$$\begin{aligned}
\delta &\ge \mathbb{P}\left(\bar{M}_t \ge \frac{1}{\delta}\right) \quad \text{(Equation (5.10) + Markov's Inequality)} \\
&\ge \mathbb{P}\left(f_t(\xi^\star) \ge \log\left(\frac{1}{\delta}\right) + \log\left(\frac{N(h_t)}{N(e_t)}\right)\right) \quad \text{(Equation (5.12))} \\
&= \mathbb{P}\left(\max_{\|\xi\|_2 \le 1/2} f_t(\xi) \ge \log\left(\frac{1}{\delta}\right) + \log\left(\frac{N(h_t)}{N(e_t)}\right)\right) \\
&\ge \mathbb{P}\left(f_t(\xi_0) \ge \log\left(\frac{1}{\delta}\right) + \log\left(\frac{N(h_t)}{N(e_t)}\right)\right) .
\end{aligned}$$

In the last inequality ξ_0 is defined as $\xi_0 = \frac{\sqrt{\lambda_t}}{2}\frac{\widetilde{H}_t^{-1} S_t}{\|S_t\|_{\widetilde{H}_t^{-1}}}$, such that $\|\xi_0\|_2 \le 1/2$ holds. This can be seen by using $\widetilde{H}_t \ge \lambda_t I_d$. We also have,

$$f_t(\xi_0) = \frac{1}{m w_t}\xi_0^\top S_t - \frac{1}{m^2 w_t^2}\xi_0^\top \widetilde{H}_t \xi_0 = \frac{\sqrt{\lambda_t}}{2 m w_t}\|S_t\|_{\widetilde{H}_t^{-1}} - \frac{\lambda_t}{4 m^2 w_t^2} .$$

Therefore,

$$\mathbb{P}\left(\|S_t\|_{\widetilde{H}_t^{-1}} \ge \frac{\sqrt{\lambda_t}}{2 m w_t} + \frac{2 m w_t}{\sqrt{\lambda_t}}\log(1/\delta) + \frac{2 m w_t}{\sqrt{\lambda_t}}\log\left(\frac{N(h_t)}{N(e_t)}\right)\right) \le \delta . \tag{5.13}$$

We conclude using Proposition 5.8 stated below.

Proposition 5.8. *Let h_t be the density of a d-dimensional isotropic normal distribution of precision $\frac{2\lambda_t}{m^2 w_t^2}$ truncated on $\mathcal{B}_2(d)$. Let e_t be the density of a d-dimensional normal distribution with precision matrix $\frac{2}{m^2 w_t^2}\widetilde{H}_t$ truncated on $\{a \in \mathbb{R}^d, \|a\|_2 \leq 1/2\}$. The following inequality holds,*

$$\log\left(\frac{N(h_t)}{N(e_t)}\right) \leq \log\left(\frac{\det(\widetilde{H}_t)}{\lambda_t^{d/2}}\right) + d\log(2)\ . \tag{5.14}$$

Proof.

$$\begin{aligned} N(h_t) &= \int_{\mathbb{R}^d} \mathbb{1}\left[\|\xi\|_2 \leq 1\right] \exp\left(-\frac{1}{2}\frac{2\lambda_t}{m^2 w_t^2}\|\xi\|_2^2\right) d\xi \\ &= \left(\frac{m^2 w_t^2}{2\lambda_t}\right)^{d/2} \int_{\mathbb{R}^d} \mathbb{1}\left[\|\xi\|_2 \leq \frac{\sqrt{2\lambda_t}}{m w_t}\right] \exp\left(-\frac{1}{2}\|\xi\|_2^2\right) d\xi\ . \end{aligned}$$

$$\begin{aligned} N(e_t) &= \int_{\mathbb{R}^d} \mathbb{1}\left[\|\xi\|_2 \leq 1/2\right] \exp\left(-\frac{1}{2}\frac{2}{m^2 w_t^2}\xi^\top \widetilde{H}_t \xi\right) d\xi \\ &= \frac{1}{\left|\det\left(\frac{\sqrt{2}}{m w_t}\widetilde{H}_t^{1/2}\right)\right|} \int_{\mathbb{R}^d} \mathbb{1}\left[\|\xi\|_2 \leq \frac{1}{2}\frac{\sqrt{2\lambda_t}}{m w_t}\right] \exp\left(-\frac{1}{2}\|\xi\|_2^2\right) d\xi \\ &\geq \left(\frac{m^2 w_t^2}{2}\right)^{d/2} \det(\widetilde{H}_t)^{-1/2} \int_{\mathbb{R}^d} \mathbb{1}\left[\|\xi\|_2 \leq \frac{1}{2}\frac{\sqrt{2\lambda_t}}{m w_t}\right] \exp\left(-\frac{1}{2}\|\xi\|_2^2\right) d\xi\ . \end{aligned}$$

Therefore,

$$\frac{N(h_t)}{N(e_t)} \leq \frac{\det(\widetilde{H}_t)}{\lambda_t^{d/2}} \underbrace{\frac{\int_{\mathbb{R}^d} \mathbb{1}\left[\|\xi\|_2 \leq \frac{\sqrt{2\lambda_t}}{m w_t}\right] \exp\left(-\frac{1}{2}\|\xi\|_2^2\right) d\xi}{\int_{\mathbb{R}^d} \mathbb{1}\left[\|\xi\|_2 \leq \frac{1}{2}\frac{\sqrt{2\lambda_t}}{m w_t}\right] \exp\left(-\frac{1}{2}\|\xi\|_2^2\right) d\xi}}_{R}\ . \tag{5.15}$$

The last step consists in upper bounding the ratio of the integrals R. Following, [Faury et al., 2020, Lemma 6], one gets $R = 2^d$.
We conclude by using this equality in Equation (5.15) and applying the logarithm on both sides. □

5.A.3 A Unifying Concentration Result for Discount Factors and Sliding-Window

In this section, we explain how Theorem 5.3 can be used with self-concordant GLBs to obtain a concentration inequality that encapsulates the analysis for both discount-factors and the sliding-window.

Up to now, we have stated the results in the most generic way. Actually, in our analysis we will use a weaker version of the concentration inequality established in Theorem 5.3.

Theorem 5.9. *Let t be a fixed time instant. Let $\{\mathcal{F}_u\}_{u=0}^t$ be a filtration. Let $\{A_u\}_{u=0}^t$ be a stochastic process on $\mathbb{R}^d$ such that A_u is $\mathcal{F}_{u-1}$ measurable and $\|A_u\|_2 \leq 1$. Let $\{\epsilon_u\}_{u=0}^t$ be a martingale difference sequence such that ϵ_u is $\mathcal{F}_u$ measurable. Assume that the weights are non-decreasing, positive and the time horizon is known. Furthermore, assume*

that conditionally on $\mathcal{F}_u$ we have $|\epsilon_u| \leq m$ a.s. Let $\{\lambda_u\}_{u=1}^t$ be a deterministic sequence of regularization terms and denote $\sigma_t^2 = \mathbb{E}\left[\epsilon_t^2 | \mathcal{F}_{t-1}\right]$. Let $\widetilde{H}_{t-t_0:t} = \sum_{s=t-t_0+1}^t w_s^2 \sigma_s^2 A_s A_s^\top + \lambda_t I_d$ and $S_{t-t_0:t} = \sum_{s=t-t_0+1}^t w_s \epsilon_s A_s$. Then for any $\delta \in (0,1]$,

$$\mathbb{P}\left(\|S_{t-t_0:t}\|_{\widetilde{H}^{-1}_{t-t_0:t}} \geq \frac{\sqrt{\lambda_t}}{2mw_t} + \frac{2mw_t}{\sqrt{\lambda_t}} \log\left(\frac{\det(\widetilde{H}_{t-t_0:t})^{1/2}}{\delta \lambda_t^{d/2}} \right) + \frac{2mw_t}{\sqrt{\lambda_t}} d\log(2) \right) \leq \delta \, .$$

Proof. The arguments used to establish Theorem 5.9 are the same than for Theorem 5.3. We only give the main term that differs from the proof of Theorem 5.3.
With t a fixed time instant, for any u such that $t - t_0 + 1 \leq u \leq t$, M_u^t is defined as

$$M_u^t(\xi) = \exp\left(\frac{1}{mw_t} \xi^\top S_{t-t_0:u} - \frac{1}{m^2 w_t^2} \xi^\top \sum_{s=t-t_0}^{u} w_s^2 A_s A_s^\top \xi \right) \, ,$$

with $S_{t-t_0:u} = \sum_{s=t-t_0+1}^u w_s \epsilon_s A_s$. Following the steps of the proof of Theorem 5.3 with these slight differences gives the result. □

Discount Factors Let $t_0 = D$ be the equivalent of the sliding window length with exponential weights, $w_t = \gamma^{-t}$ and $\lambda_t = \lambda \gamma^{-2t}$ for $0 < \gamma < 1$. Even when γ depends on T, the weights satisfy the assumptions 5.6 and 5.7. We can obtain:

Corollary 5.10 (Concentration result with discount factors)**.** *Under the same assumption than Theorem 5.9, when defining $\widetilde{H}_{t-D:t} = \sum_{s=t-D+1}^t \gamma^{2(t-s)} \dot{\mu}(A_s^\top \theta_s^\star) A_s A_s^\top + \lambda I_d$ and $S_{t-D:t} = \sum_{s=t-D+1}^t \gamma^{-s} \epsilon_s A_s$. For any $\delta \in (0,1]$,*

$$\mathbb{P}\left(\left\| \gamma^t S_{t-D:t} \right\|_{\widetilde{H}^{-1}_{t-D:t}} \geq \frac{\sqrt{\lambda}}{2m} + \frac{2m}{\sqrt{\lambda}} \log\left(\frac{\det(\widetilde{H}_{t-D:t})^{1/2}}{\delta \lambda^{d/2}} \right) + \frac{2m}{\sqrt{\lambda}} d\log(2) \right) \leq \delta \, .$$

Sliding Window With $t_0 = \tau$ the length of the sliding window, with the weights satisfying $w_t = 1$ for $t - \tau + 1 \leq s \leq t$ and $\lambda_t = \lambda$, we have:

Corollary 5.11 (Concentration result with a sliding window)**.** *Under the same assumption than Theorem 5.9, when defining $H_t = \sum_{s=\max(1,t-\tau+1)}^t \dot{\mu}(A_s^\top \theta_s^\star) A_s A_s^\top + \lambda I_d$ and $S_t = \sum_{s=\max(1,t-\tau+1)}^t \epsilon_s A_s$. For any $\delta \in (0,1]$,*

$$\mathbb{P}\left(\|S_t\|_{H_t^{-1}} \geq \frac{\sqrt{\lambda}}{2m} + \frac{2m}{\sqrt{\lambda}} \log\left(\frac{\det(H_t)^{1/2}}{\delta \lambda^{d/2}} \right) + \frac{2m}{\sqrt{\lambda}} d\log(2) \right) \leq \delta \, .$$

Appendix 5.B Regret Analysis with Discount Factors

In this section we detail the regret analysis of SC-D-GLUCB. First we recall the main notation.

5.B.1 Notation

For any $\theta \in \mathbb{R}^d$,

$$\widetilde{H}_t(\theta) = \sum_{s=1}^{t} \gamma^{2(t-s)} \dot{\mu}(A_s^\top \theta) A_s A_s^\top + \lambda I_d \ . \tag{5.16}$$

$$H_t(\theta) = \sum_{s=1}^{t} \gamma^{t-s} \dot{\mu}(A_s^\top \theta) A_s A_s^\top + \lambda I_d \ . \tag{5.17}$$

$$\widetilde{V}_t = \sum_{s=1}^{t} \gamma^{2(t-s)} A_s A_s^\top + \frac{\lambda}{c_\mu} I_d \ . \tag{5.18}$$

$$V_t = \sum_{s=1}^{t} \gamma^{t-s} A_s A_s^\top + \frac{\lambda}{c_\mu} I_d \ . \tag{5.19}$$

$$g_t(\theta) = \sum_{s=1}^{t} \gamma^{t-s} \mu(A_s^\top \theta) A_s + \lambda \theta \ . \tag{5.20}$$

$$S_t = \sum_{s=1}^{t} \gamma^{-s} \epsilon_s A_s \ . \tag{5.21}$$

For any $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\alpha(a, \theta_1, \theta_2) = \int_0^1 \dot{\mu}(v a^\top \theta_2 + (1-v) a^\top \theta_1) dv \ .$$

$$G_t(\theta_1, \theta_2) = \sum_{s=1}^{t} \gamma^{t-s} \alpha(A_s, \theta_1, \theta_2) A_s A_s^\top + \lambda I_d \ .$$

$$\widetilde{G}_t(\theta_1, \theta_2) = \sum_{s=1}^{t} \gamma^{2(t-s)} \alpha(A_s, \theta_1, \theta_2) A_s A_s^\top + \lambda I_d \ . \tag{5.22}$$

Let $\widetilde{H}_t$ be defined as

$$\widetilde{H}_t = \sum_{s=1}^{t} \gamma^{2(t-s)} \dot{\mu}(A_s^\top \theta_s^\star) A_s A_s^\top + \lambda I_d \ . \tag{5.23}$$

Let us define $\mathcal{T}(\gamma)$ as

$$\mathcal{T}(\gamma) = \{1 \leq t \leq T, \text{such that} \, \forall s, \, t - D + 1 \leq s \leq t, \theta_s^\star = \theta_t^\star\} \ . \tag{5.24}$$

Remark 5.5. *$t \in \mathcal{T}(\gamma)$ when t is a least D steps away from the closest previous breakpoint. On the contrary to the analysis with the sliding window (see Appendix 5.C) the bias does not completely cancel out when we are far enough from a breakpoint.*

D is an analysis parameter and will be specified later in the different theorems. For the entire section we will use the notation $t-D:t$ when the sum concerns time instants s such that $t-D+1 \leq s \leq t$. In the weighted setting, we construct an estimator based on a weighted penalized log-likelihood. $\hat{\theta}_t$ is defined as the unique maximizer of

$$\sum_{s=1}^{t} \gamma^{t-s} \log \mathbb{P}_\theta(X_s|A_s) - \frac{\lambda}{2}\|\theta\|_2^2 \, .$$

By using the definition of the GLM and thanks to the concavity of this equation in θ, $\hat{\theta}_t$ is the unique solution of

$$\sum_{s=1}^{t} \gamma^{t-s}(X_s - \mu(A_s^\top \theta))A_s - \lambda\theta = 0 \, .$$

This can be summarized with

$$g_t(\hat{\theta}_t) = \sum_{s=1}^{t} \gamma^{t-s} X_s A_s = \gamma^t S_t + \sum_{s=1}^{t} \gamma^{t-s} \mu(A_s^\top \theta_s^\star) A_s \, . \tag{5.25}$$

5.B.2 Analysis of the Regret of SC-D-GLUCB

In this section, we present the main ideas to obtain an analysis of the regret of the SC-D-GLUCB algorithm when the projection step is avoided.

We define

$$\rho_T^\delta = \left(\frac{\sqrt{\lambda}}{2m} + \frac{2m}{\sqrt{\lambda}} \log\left(\frac{T}{\delta}\right) + \frac{dm}{\sqrt{\lambda}} \log\left(1 + \frac{k_\mu(1-\gamma^{2D})}{d\lambda(1-\gamma^2)}\right) + \frac{2m}{\sqrt{\lambda}} d\log(2) \right) \, , \tag{5.26}$$

and also,

$$\bar{S} = S + \frac{\gamma^D(2Sk_\mu + m)}{\lambda(1-\gamma)} \, . \tag{5.27}$$

The expression of ρ_T^δ and $\bar{S}$ given here coincide with the expression in the main content of the chapter when $D = \log(T)/\log(1/\gamma)$. ρ_T^δ is defined such that thanks to Corollary 5.10 with high probability for all t in $\mathcal{T}(\gamma)$, $\|\gamma^t S_{t-D:t}\|_{\widetilde{H}_{t-D:t}^{-1}} \leq \rho_T^\delta$ holds.

The next result uses the self-concordance to relate the first derivative of the link function evaluated at different points. This relation is independent of c_μ and only depends on the distance between the parameters.

Proposition 5.12. *When $\hat{\theta}_t$ is the maximum likelihood as defined in Equation (5.1) and $t \in \mathcal{T}(\gamma)$, we have*

$$\alpha(a, \theta_t^\star, \hat{\theta}_t) \geq \left(1 + \bar{S} + \frac{1}{\sqrt{\lambda}} \|\gamma^t S_{t-D:t}\|_{\widetilde{G}_{t-D:t}^{-1}(\theta_t^\star, \hat{\theta}_t)}\right)^{-1} \dot{\mu}(a^\top \theta_t^\star) \, ,$$

where $\bar{S}$ is defined in Equation (5.27).

Proof. In the proof, we will replace the notation $\widetilde{G}_{t-D:t}(\theta_t^\star, \hat{\theta}_t)$ with $\widetilde{G}_{t-D:t}$ and $\widetilde{G}_t(\theta_t^\star, \hat{\theta}_t)$

with $\widetilde{G}_t$ but also $G_t(\theta_t^\star, \hat{\theta}_t)$ with G_t. Using Lemma 5.23 we have,

$$\alpha(a, \theta_t^\star, \hat{\theta}_t) \geq \left(1 + \left|a^\top(\hat{\theta}_t - \theta_t^\star)\right|\right)^{-1} \dot{\mu}(a^\top \theta_t^\star) \, .$$

Combining this with the mean value theorem gives

$$\alpha(a, \theta_t^\star, \hat{\theta}_t) \geq \left(1 + \left|a^\top G_t^{-1}\left(g_t(\hat{\theta}_t) - g_t(\theta_t^\star)\right)\right|\right)^{-1} \dot{\mu}(a^\top \theta_t^\star) \, .$$

Next, it is possible to upper bound $|a^\top G_t^{-1}\left(g_t(\hat{\theta}_t) - g_t(\theta_t^\star)\right)|$ using the triangle inequality and Equation (5.25).

$$\begin{aligned}
\left|a^\top G_t^{-1}\left(g_t(\hat{\theta}_t) - g_t(\theta_t^\star)\right)\right| \leq & \underbrace{\left|a^\top G_t^{-1} \sum_{s=1}^{t} \gamma^{t-s}(\mu(A_s^\top \theta_s^\star) - \mu(A_s^\top \theta_t^\star))A_s\right|}_{b_{1,t}(a)} \\
& + \underbrace{\left|a^\top G_t^{-1}\left(-\lambda\theta_t^\star + \sum_{s=1}^{t-D} \gamma^{t-s}\epsilon_s A_s\right)\right|}_{b_{2,t}(a)} \\
& + \underbrace{\left|a^\top G_t^{-1}\gamma^t S_{t-D:t}\right|}_{b_{3,t}(a)}
\end{aligned}$$

The first term is controlled as follows,

$$\begin{aligned}
b_{1,t}(a) &= \left|a^\top G_t^{-1} \sum_{s=1}^{t} \gamma^{t-s}\left(\mu(A_s^\top \theta_s^\star) - \mu(A_s^\top \theta_t^\star)\right) A_s\right| \\
&\leq \|a\|_{G_t^{-1}} \left\|\sum_{s=1}^{t} \gamma^{t-s}(\mu(A_s^\top \theta_s^\star) - \mu(A_s^\top \theta_t^\star))A_s\right\|_{G_t^{-1}} \quad \text{(Cauchy-Schwarz ineq.)} \\
&\leq \frac{1}{\sqrt{\lambda}} \left\|\sum_{s=1}^{t-D} \gamma^{t-s}(\mu(A_s^\top \theta_s^\star) - \mu(A_s^\top \theta_t^\star))A_s\right\|_{G_t^{-1}} \quad (G_t \geq \lambda I_d \text{ and } t \in \mathcal{T}(\gamma)) \\
&\leq \frac{1}{\lambda} \sum_{s=1}^{t-D} \gamma^{t-s} |\alpha(A_s, \theta_s^\star, \theta_t^\star)| \times |A_s^\top(\theta_t^\star - \theta_s^\star)| \times \|A_s\|_2 \quad \text{(Triangle ineq.} + G_t \geq \lambda I_d) \\
&\leq \frac{2Sk_\mu}{\lambda} \sum_{s=1}^{t-D} \gamma^{t-s} \quad (\theta_s^\star \text{ and } \theta_t^\star \in \Theta) \\
&\leq \frac{2Sk_\mu}{\lambda} \frac{\gamma^D}{1-\gamma} \, .
\end{aligned}$$

Using similar arguments, one can upper bound $b_{2,t}(a)$.

$$\begin{aligned}
b_{2,t}(a) &= \left|a^\top G_t^{-1}\left(-\lambda\theta_t^\star + \sum_{s=1}^{t-D} \gamma^{t-s}\epsilon_s A_s\right)\right| \\
&\leq S + \|\sum_{s=1}^{t-D} \gamma^{t-s}\epsilon_s A_s\|_{G_t^{-2}} \\
&\leq S + \frac{m}{\lambda}\frac{\gamma^D}{1-\gamma} \, . \quad (|\epsilon_s| \leq m)
\end{aligned}$$

Before upper bounding, $b_{3,t}(a)$, we need the following relation.
When $0 < \gamma < 1$, $\gamma^{2(t-s)} \leq \gamma^{t-s}$ for s smaller than t which implies

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, \ \widetilde{G}_t(\theta_1, \theta_2) \leq G_t(\theta_1, \theta_2) \ . \tag{5.28}$$

We have,

$$\begin{aligned} b_{3,t}(a) &= |a^\top G_t^{-1} \widetilde{G}_t^{1/2} \widetilde{G}_t^{-1/2} \gamma^t S_{t-D:t}| \\ &\leq \|a\|_{G_t^{-1} \widetilde{G}_t G_t^{-1}} \|\gamma^t S_{t-D:t}\|_{\widetilde{G}_t^{-1}} \quad \text{(Cauchy-Schwarz ineq.)} \\ &\leq \|a\|_{G_t^{-1}} \|\gamma^t S_{t-D:t}\|_{\widetilde{G}_t^{-1}} \quad \text{(Equation (5.28))} \\ &\leq \frac{1}{\sqrt{\lambda}} \|\gamma^t S_{t-D:t}\|_{\widetilde{G}_{t-D:t}^{-1}} \ . \quad (G_t \geq \lambda I_d) \end{aligned}$$

By combining all the results we have,

$$\alpha(a, \theta_t^\star, \hat{\theta}_t) \geq \left(1 + \bar{S} + \frac{1}{\sqrt{\lambda}} \|\gamma^t S_{t-D:t}\|_{\widetilde{G}_{t-D:t}^{-1}}\right)^{-1} \dot{\mu}(a^\top \theta_t^\star) \ .$$

□

Corollary 5.13. *When $\hat{\theta}_t$ is the maximum likelihood as defined in Equation (5.1), and $t \in \mathcal{T}(\gamma)$, we have*

$$\widetilde{G}_{t-D:t}(\theta_t^\star, \hat{\theta}_t) \geq \left(1 + \bar{S} + \frac{1}{\sqrt{\lambda}} \|\gamma^t S_{t-D:t}\|_{\widetilde{G}_{t-D:t}^{-1}(\theta_t^\star, \hat{\theta}_t)}\right)^{-1} \widetilde{H}_{t-D:t} \ .$$

This proposition establishes a useful link between $\widetilde{G}_{t-D:t}(\theta_t^\star, \hat{\theta}_t)$ and $\widetilde{H}_{t-D:t}$.

Proof. Thanks to Proposition 5.12,

$$\alpha(A_s, \theta_t^\star, \hat{\theta}_t) \geq \left(1 + \bar{S} + \frac{1}{\sqrt{\lambda}} \|\gamma^t S_{t-D:t}\|_{\widetilde{G}_t^{-1}(\hat{\theta}_t, \theta_t^\star)}\right)^{-1} \dot{\mu}(A_s^\top \theta_t^\star) \ .$$

Therefore,

$$\begin{aligned} \sum_{s=t-D+1}^{t} \gamma^{2(t-s)} \alpha(A_s, \theta_t^\star, \hat{\theta}_t) A_s A_s^\top \geq \left(1 + \bar{S} + \frac{1}{\sqrt{\lambda}} \|\gamma^t S_{t-D:t}\|_{\widetilde{G}_t^{-1}(\hat{\theta}_t, \theta_t^\star)}\right)^{-1} \\ \times \sum_{s=t-D+1}^{t} \gamma^{2(t-s)} \dot{\mu}(A_s^\top \theta_t^\star) A_s A_s^\top \ . \end{aligned}$$

We obtain the announced result by using $\theta_s^\star = \theta_t^\star$ for $t - D + 1 \leq s \leq t$ because $t \in \mathcal{T}(\gamma)$ and by adding the regularization terms. □

Using Proposition 5.12 and Corollary 5.13, we can now prove Proposition 5.4. The proposition establishes an upper bound for the deviation of the MLE (through $\gamma^t S_{t-D:t}$) that only depends on ρ_T^δ the high probability upper bound obtained using Corollary 5.10.

Proposition 5.4. *For any $\delta \in (0,1]$, with probability higher than $1-\delta$,*

$$\forall t \in \mathcal{T}(\gamma),\ \|\gamma^t S_{t-D:t}\|_{\widetilde{G}^{-1}_{t-D:t}(\hat{\theta}_t,\theta^\star_t)} \leq \sqrt{1+\bar{S}}\rho^\delta_T + \frac{1}{\sqrt{\lambda}}\left(\rho^\delta_T\right)^2 ,$$

where ρ^δ_T is defined in Equation (5.26).

Remark 5.6. *Here, note that the left-hand side is controlled under the norm $\widetilde{G}^{-1}_{t-D:t}(\hat{\theta}_t,\theta^\star_t)$, whereas the right hand side is the consequence of the upper bound of the same term controlled in the $\widetilde{H}^{-1}_{t-D:t}$-norm (Corollary 5.10). Linking those two matrices independently from c_μ is not-straightforward. The self-concordance is the key ingredient to obtain this bound.*

Proof. Applying Corollary 5.13,

$$\|\gamma^t S_{t-D:t}\|^2_{\widetilde{G}^{-1}_{t-D:t}(\hat{\theta}_t,\theta^\star_t)} \leq \left(1+\bar{S}+\frac{1}{\sqrt{\lambda}}\|\gamma^t S_{t-D:t}\|_{\widetilde{G}^{-1}_{t-D:t}(\hat{\theta}_t,\theta^\star_t)}\right)\|\gamma^t S_{t-D:t}\|^2_{\widetilde{H}^{-1}_{t-D:t}} .$$

Let $X = \|\gamma^t S_{t-D:t}\|_{\widetilde{G}^{-1}_{t-D:t}(\hat{\theta}_t,\theta^\star_t)}$, it gives the following constraint,

$$\forall X,\ X^2 - \frac{1}{\sqrt{\lambda}}\|\gamma^t S_{t-D:t}\|^2_{\widetilde{H}^{-1}_{t-D:t}} X - \left(1+\bar{S}\right)\|\gamma^t S_{t-D:t}\|^2_{\widetilde{H}^{-1}_{t-D:t}} \leq 0 .$$

Solving this polynomial inequality yields

$$\|\gamma^t S_{t-D:t}\|_{\widetilde{G}^{-1}_t(\theta^\star_t,\hat{\theta}_t)} \leq \frac{1}{\sqrt{\lambda}}\|\gamma^t S_{t-D:t}\|^2_{\widetilde{H}^{-1}_{t-D:t}} + \sqrt{1+\bar{S}}\|\gamma^t S_{t-D:t}\|_{\widetilde{H}^{-1}_{t-D:t}} .$$

The result is then obtained by applying Corollary 5.10. □

Corollary 5.14. *When $\hat{\theta}_t$ is the maximum likelihood as defined in Equation (5.1) and $t \in \mathcal{T}(\gamma)$, we have*

$$G_t(\theta^\star_t,\hat{\theta}_t) \geq \left(1+\bar{S}+\frac{1}{\sqrt{\lambda}}\|\gamma^t S_{t-D:t}\|_{\widetilde{G}^{-1}_{t-D:t}(\theta^\star_t,\hat{\theta}_t)}\right)^{-1} c_\mu V_t .$$

Proof. Similar to the proof of Corollary 5.13. □

In the next proposition, we give an upper bound for $\Delta_t(a,\hat{\theta}_t)$ the prediction error in $\hat{\theta}_t$ which is directly connected to the instantaneous regret. Here, β^δ_T is defined as in the main content of the chapter in Equation (5.4) but we replace ρ^δ_T and $\bar{S}$ with the expressions stated in Equation (5.26) and Equation (5.27).

Proposition 5.15. *For any $\delta \in (0,1]$, with probability higher than $1-\delta$,*

$$\forall t \in \mathcal{T}(\gamma),\ \Delta_t(a,\hat{\theta}_t) \leq \frac{k_\mu}{\lambda}\frac{\gamma^D}{1-\gamma}(2Sk_\mu+m) + \frac{\beta^\delta_T}{\sqrt{c_\mu}}\|a\|_{V_t^{-1}} .$$

Proof. We denote $G_t = G_t(\theta_t^\star, \hat{\theta}_t)$ and we have,

$$\begin{aligned}
\Delta_t(a, \hat{\theta}_t) &= |\mu(a^\top \theta_t^\star) - \mu(a^\top \hat{\theta}_t)| \\
&\leq k_\mu |a^\top (\theta_t^\star - \hat{\theta}_t)| \\
&= k_\mu |a^\top G_t^{-1} (g_t(\theta_t^\star) - g_t(\hat{\theta}_t))| \quad \text{(Mean-Value Theorem)} \\
&= k_\mu \left| a^\top G_t^{-1} \left(\sum_{s=1}^{t} \gamma^{t-s} (\mu(A_s^\top \theta_t^\star) - \mu(A_s^\top \theta_s^\star)) A_s + \lambda \theta_t^\star - \gamma^t S_t \right) \right| .
\end{aligned}$$

In the last equality, we have used the characterization of the MLE (Equation (5.25)).

$$\begin{aligned}
\Delta_t(a, \hat{\theta}_t) \leq k_\mu \underbrace{\left| a^\top G_t^{-1} \sum_{s=1}^{t} \gamma^{t-s} (\mu(A_s^\top \theta_t^\star) - \mu(A_s^\top \theta_s^\star)) A_s \right|}_{c_{1,t}(a)} \\
+ k_\mu \underbrace{\left| a^\top G_t^{-1} \sum_{s=1}^{t-D} \gamma^{t-s} \epsilon_s A_s \right|}_{c_{2,t}(a)} + k_\mu \underbrace{\left| a^\top G_t^{-1} \left(\gamma^t S_{t-D:t} - \lambda \theta_t^\star \right) \right|}_{c_{3,t}(a)} .
\end{aligned}$$

We will bound the different terms.

$c_{1,t}(a)$ can be bounded like $b_{1,t}(a)$ in the proof of Proposition 5.12.

$$c_{1,t}(a) \leq \frac{2Sk_\mu}{\lambda} \frac{\gamma^D}{1-\gamma} .$$

$c_{2,t}(a)$ can be bounded like $b_{2,t}(a)$ in the proof of the same proposition.

$$c_{2,t}(a) \leq \frac{m}{\lambda} \frac{\gamma^D}{1-\gamma} .$$

The last term requires more work. $\widetilde{G}_t(\theta_t^\star, \hat{\theta}_t)$ will be denoted $\widetilde{G}_t$ for simplicity.

$$\begin{aligned}
c_{3,t}(a) &= \left| a^\top G_t^{-1} \left(\gamma^t S_{t-D:t} - \lambda \theta_t^\star \right) \right| = \left| a^\top G_t^{-1} \widetilde{G}_t^{1/2} \widetilde{G}_t^{-1/2} \left(\gamma^t S_{t-D:t} - \lambda \theta_t^\star \right) \right| \\
&\leq \|a\|_{G_t^{-1} \widetilde{G}_t G_t^{-1}} \|\gamma^t S_{t-D:t} - \lambda \theta_t^\star\|_{\widetilde{G}_t^{-1}} \\
&\leq \|a\|_{G_t^{-1}} \|\gamma^t S_{t-D:t} - \lambda \theta_t^\star\|_{\widetilde{G}_t^{-1}} \quad \text{(Equation (5.28))} \\
&\leq \|a\|_{G_t^{-1}} \left(\sqrt{\lambda} S + \|\gamma^t S_{t-D:t}\|_{\widetilde{G}_t^{-1}} \right) \quad (\widetilde{G}_t \geq \lambda I_d \text{ and Assumption 5.1}) \\
&\leq \frac{\|a\|_{V_t^{-1}}}{\sqrt{c_\mu}} \sqrt{1 + \bar{S} + \frac{1}{\sqrt{\lambda}} \|\gamma^t S_{t-D:t}\|_{\widetilde{G}_{t-D:t}^{-1}}} \left(\sqrt{\lambda} S + \|\gamma^t S_{t-D:t}\|_{\widetilde{G}_{t-D:t}^{-1}} \right) .
\end{aligned}$$

In the last inequality we used Corollary 5.14. The next step consists in upper bounding $\|\gamma^t S_{t-D:t}\|_{\widetilde{G}_{t-D:t}^{-1}}$ with Proposition 5.4 and to combine this with the high probability upper bound from Corollary 5.10. Therefore, with probability higher than $1-\delta$,

$$
\begin{aligned}
c_{3,t}(a) &\leq \frac{\|a\|_{V_t^{-1}}}{\sqrt{c_\mu}} \sqrt{1+\bar{S}+\sqrt{\frac{1+\bar{S}}{\lambda}}\rho_T^\delta + \frac{1}{\lambda}(\rho_T^\delta)^2} \left(\sqrt{\lambda}S + \|\gamma^t S_{t-D:t}\|_{\widetilde{G}_{t-D:t}^{-1}}\right) \\
&\leq \frac{\sqrt{\lambda}}{\sqrt{c_\mu}}\|a\|_{V_t^{-1}} \sqrt{1+\bar{S}+\sqrt{\frac{1+\bar{S}}{\lambda}}\rho_T^\delta + \frac{1}{\lambda}(\rho_T^\delta)^2} \left(S + \sqrt{\frac{1+\bar{S}}{\lambda}}\rho_T^\delta + \frac{1}{\lambda}(\rho_T^\delta)^2\right) \\
&\leq \frac{\sqrt{\lambda}}{\sqrt{c_\mu}}\|a\|_{V_t^{-1}} \left(1+\bar{S}+\sqrt{\frac{1+\bar{S}}{\lambda}}\rho_T^\delta + \frac{1}{\lambda}(\rho_T^\delta)^2\right)^{3/2} .
\end{aligned}
$$

□

The first term of the right hand side of Proposition 5.15 is a bias term resulting from the non-stationarity of the environment. The second term results from the concentration results we have established in Section 5.A combined with the self-concordance assumption.

With β_T^δ defined in Equation (5.4), the algorithm SC-D-GLUCB selects the action at time t as follows,

$$
\begin{aligned}
A_t &= \operatorname{argmax}_{a\in\mathcal{A}_t} \left(\mu(a^\top \hat{\theta}_t) + \frac{\beta_T^\delta}{\sqrt{c_\mu}}\|a\|_{V_t^{-1}} + \frac{k_\mu}{\lambda}\frac{\gamma^D}{1-\gamma}(2Sk_\mu + m)\right) \\
&= \operatorname{argmax}_{a\in\mathcal{A}_t} \left(\mu(a^\top \hat{\theta}_t) + \frac{\beta_T^\delta}{\sqrt{c_\mu}}\|a\|_{V_t^{-1}}\right) . \qquad (5.29)
\end{aligned}
$$

Note that the bias term is independent of the action. Nevertheless, this term will appear in the upper bound for the regret. Equation (5.29) explains how the actions are chosen in Algorithm 11.

We can now give the main theorem.

Theorem 5.2. *The regret of the* SC-D-GLUCB *algorithm is bounded for all* $\gamma \in (1/2, 1)$ *with probability at least* $1-\delta$ *by*

$$
\begin{aligned}
R(T) \leq{}& \frac{2\log(T)}{1-\gamma}\Gamma_T + \frac{2k_\mu(2Sk_\mu+m)}{\lambda}\frac{1}{1-\gamma} \\
&+ \frac{2\beta_T^\delta}{\sqrt{c_\mu}}\sqrt{dT}\sqrt{2\max\left(1,\frac{1}{\lambda}\right)}\sqrt{T\log(1/\gamma) + \log\left(1+\frac{1}{d\lambda(1-\gamma)}\right)} .
\end{aligned}
$$

In particular, setting $\gamma = 1 - \left(\frac{c_\mu^{1/2}\Gamma_T}{dT}\right)^{2/3}$ *and* $\lambda = d\log(T)$ *leads to*

$$
R(T) = \widetilde{\mathcal{O}}(c_\mu^{-1/3} d^{2/3} \Gamma_T^{1/3} T^{2/3}) .
$$

Proof. Using Proposition 5.15, we obtain a high probability upper bound for $\Delta_t(a, \hat{\theta}_t)$. We recall that the exploration bonus of SC-D-GLUCB is defined as,

$$
\frac{1}{\sqrt{c_\mu}}\beta_T^\delta \|a_t\|_{V_t^{-1}} + \frac{k_\mu}{\lambda}\frac{\gamma^D}{1-\gamma}(2Sk_\mu + m) .
$$

Furthermore, the estimator used by SC-D-GLUCB is the MLE $\hat{\theta}_t$ as defined in Equation (5.1),

all the conditions required for applying Proposition 5.32 are met. Hence when $t \in \mathcal{T}(\gamma)$,

$$r_t \leq \frac{2}{\sqrt{c_\mu}} \beta_T^\delta \|a_t\|_{V_t^{-1}} + \frac{2k_\mu}{\lambda} \frac{\gamma^D}{1-\gamma}(2Sk_\mu + m) \, .$$

The dynamic pseudo-regret can then be upper bounded by,

$$\begin{aligned}
R(T) = \sum_{t=1}^{T} r_t &= \sum_{t \in \mathcal{T}(\gamma)} r_t + \sum_{t \notin \mathcal{T}(\gamma)} r_t \leq \Gamma_T D + \sum_{t \in \mathcal{T}(\gamma)} r_t \\
&\leq \Gamma_T D + \frac{2k_\mu}{\lambda} \frac{\gamma^D}{1-\gamma}(2Sk_\mu + m)T + \frac{2\beta_T^\delta}{\sqrt{c_\mu}} \sum_{t \in \mathcal{T}(\gamma)} \|A_t\|_{V_t^{-1}} \\
&\leq \Gamma_T D + \frac{2k_\mu}{\lambda} \frac{\gamma^D}{1-\gamma}(2Sk_\mu + m)T + \frac{2\beta_T^\delta}{\sqrt{c_\mu}} \sqrt{T} \sqrt{\sum_{t \in \mathcal{T}(\gamma)} \|A_t\|^2_{V_t^{-1}}} \quad \text{(Cauchy-Schwarz ineq.)} \\
&\leq \Gamma_T D + \frac{2k_\mu}{\lambda} \frac{\gamma^D}{1-\gamma}(2Sk_\mu + m)T + \frac{2\beta_T^\delta}{\sqrt{c_\mu}} \sqrt{T} \sqrt{\sum_{t=1}^{T} \|A_t\|^2_{V_t^{-1}}} \\
&\leq \Gamma_T D + \frac{2k_\mu}{\lambda} \frac{\gamma^D}{1-\gamma}(2Sk_\mu + m)T + \frac{2\beta_T^\delta}{\sqrt{c_\mu}} \sqrt{T} \sqrt{2 \max\left(1, \frac{1}{\lambda}\right) \log\left(\frac{\det(V_T)}{\gamma^{dT}\lambda^d}\right)} \, .
\end{aligned}$$

The last inequality uses Lemma 5.30. Next, we use Corollary 5.28 to upper bound the determinant,

$$\frac{\det(V_T)}{\gamma^{dT}\lambda^d} \leq \gamma^{-dT} \left(1 + \frac{1-\gamma^T}{\lambda d(1-\gamma)}\right)^d \, .$$

Applying the logarithm function on both sides yields

$$\begin{aligned}
R_T \leq \Gamma_T D &+ \frac{2k_\mu(2Sk_\mu + m)}{\lambda} \frac{\gamma^D}{1-\gamma} T \\
&+ \frac{2\beta_T^\delta}{\sqrt{c_\mu}} \sqrt{dT} \sqrt{2 \max\left(1, \frac{1}{\lambda}\right)} \sqrt{T \log(1/\gamma) + \log\left(1 + \frac{1}{d\lambda(1-\gamma)}\right)} \, .
\end{aligned}$$

With the additional constraint $1/2 < \gamma < 1$, by setting $D = \log(T)/\log(1/\gamma)$, noticing that $0 < 1/\gamma - 1 < 1$ and using $\log(1+x) \geq x/2$ for $0 < x < 1$, we have

$$\log(1/\gamma) = \log(1 + 1/\gamma - 1) \geq \frac{1-\gamma}{2\gamma} \, .$$

Therefore, we have $D \leq \frac{2\gamma \log(T)}{1-\gamma}$.
By properly balancing the bias term due to the non-stationarity and the rate at which the weighted MLE approaches the true bandit parameter, the asymptotic behavior of SC-D-GLUCB can be characterized as follows: By setting $\gamma = 1 - \left(\frac{c_\mu^{1/2}\Gamma_T}{dT}\right)^{2/3}$ and $\lambda = d\log(T)$, we have:

- $\frac{2\log(T)}{1-\gamma}\Gamma_T$ scales as $\widetilde{\mathcal{O}}(c_\mu^{-1/3} d^{2/3} \Gamma_T^{1/3} T^{2/3})$.
- $\frac{2k_\mu(2Sk_\mu+m)}{\lambda} \frac{1}{1-\gamma}$ scales as $\widetilde{\mathcal{O}}(c_\mu^{-1/3} d^{2/3} \Gamma_T^{-2/3} T^{2/3})$.

- $\frac{2\beta_T^\delta}{\sqrt{c_\mu}}\sqrt{dT}\sqrt{2\max\left(1,\frac{1}{\lambda}\right)}\sqrt{T\log(1/\gamma)+\log\left(1+\frac{1}{d\lambda(1-\gamma)}\right)}$ scales as $\frac{1}{\sqrt{c_\mu}}dT\sqrt{\log(1/\gamma)}$ when omitting logarithmic factors and constant terms.

Using $-\log(1-x)\leq\frac{x-1}{x}$ for $0\leq x<1$, we also have

$$\sqrt{\log(1/\gamma)}=\sqrt{-\log(1-(1-\gamma))}\leq\sqrt{\frac{1-\gamma}{\gamma}}\leq\sqrt{2(1-\gamma)}\,.$$

$\sqrt{\log(1/\gamma)}$ scales as $\widetilde{\mathcal{O}}(c_\mu^{1/6}d^{-1/3}\Gamma_T^{1/3}T^{-1/3})$. Hence $c_\mu^{-1/2}dT\sqrt{\log(1/\gamma)}$ scales as $\widetilde{\mathcal{O}}(c_\mu^{-1/3}d^{2/3}\Gamma_T^{1/3}T^{2/3})$. Combining the different terms concludes the proof. □

Using Assumption 5.5, we can obtain refined regret bounds.

5.B.3 Gap-Dependent Bound

Theorem 5.1. *Under Assumption 5.5, the regret of the* SC-D-GLUCB *algorithm is bounded for all $\gamma\in(1/2,1)$ with probability at least $1-\delta$ by*

$$R(T)\leq C_1\frac{\Gamma_T}{1-\gamma}+C_2\frac{1}{T(1-\gamma)^2\Delta}+C_3\frac{\beta_T^\delta\sqrt{dT}}{\sqrt{c_\mu}\Delta}\sqrt{T\log(1/\gamma)+\log\left(1+\frac{1}{d\lambda(1-\gamma)}\right)}$$
$$+C_4\frac{d(\beta_T^\delta)^2}{c_\mu\Delta}(T\log(1/\gamma)+\log(1+\frac{1}{d\lambda(1-\gamma)}))\,,$$

where C_1, C_2, C_3, C_4 are universal constants independent of c_μ, γ with only logarithmic terms in T.

In particular, setting $\gamma=1-\frac{\sqrt{c_\mu\Gamma_T}}{d\sqrt{T}}$ and $\lambda=d\log(T)$ leads to

$$R(T)=\widetilde{\mathcal{O}}(\Delta^{-1}c_\mu^{-1/2}d\sqrt{\Gamma_T T})\,.$$

Proof. First note that for any suboptimal action $a\in\mathcal{A}_t$,

$$\mu(A_{t,\star}^\top\theta_t^\star)-\mu(a^\top\theta_t^\star)\geq\Delta\,.$$

This implies

$$r_t=\mu(A_{t,\star}^\top\theta_t^\star)-\mu(A_t^\top\theta_t^\star)\leq\frac{\left(\mu(A_{t,\star}^\top\theta_t^\star)-\mu(A_t^\top\theta_t^\star)\right)^2}{\Delta}=\frac{r_t^2}{\Delta}\,.\tag{5.30}$$

Using Proposition 5.32 one has,

$$r_t\leq\frac{2}{\sqrt{c_\mu}}\beta_T^\delta\|A_t\|_{V_t^{-1}}+\frac{2k_\mu}{\lambda}\frac{\gamma^D}{1-\gamma}(2Sk_\mu+m)\,.$$

This implies in particular,

$$r_t^2\leq\underbrace{\frac{4}{c_\mu}(\beta_T^\delta)^2\|A_t\|_{V_t^{-1}}^2}_{r_{1,t}}+\underbrace{\frac{4k_\mu^2}{\lambda^2}\frac{\gamma^{2D}}{(1-\gamma)^2}(2Sk_\mu+m)^2}_{r_{2,t}}+\underbrace{\frac{8k_\mu}{\lambda}\frac{\beta_T^\delta}{\sqrt{c_\mu}}\frac{\gamma^D}{1-\gamma}(2Sk_\mu+m)\|A_t\|_{V_t^{-1}}}_{r_{3,t}}\,.\tag{5.31}$$

The dynamic regret can then be upper bounded by,

$$R(T) = \sum_{t=1}^{T} r_t = \sum_{t\in\mathcal{T}(\gamma)} r_t + \sum_{t\notin\mathcal{T}(\gamma)} r_t \leq \Gamma_T D + \sum_{t\in\mathcal{T}(\gamma)} \left(\mu(A_{t,\star}^\top \theta_t^\star) - \mu(A_t^\top \theta_t^\star)\right)$$
$$\leq \Gamma_T D + \frac{1}{\Delta} \sum_{t\in\mathcal{T}(\gamma)} r_t^2 \,. \quad \text{(Equation (5.30))}$$

By applying Equation (5.31), the regret can be separated in 4 different terms.
When summing for the different time instants $r_{1,t}$ becomes

$$\sum_{t=1}^{T} r_{1,t} \leq \frac{8}{c_\mu}(\beta_T^\delta)^2 \max\left(1, \frac{1}{\lambda}\right) \log\left(\frac{\det(V_T)}{\gamma^{dT}\lambda^d}\right) \quad \text{(Lemma 5.30)}$$
$$\leq \frac{8d}{c_\mu}(\beta_T^\delta)^2 \max\left(1, \frac{1}{\lambda}\right) \left(T\log(1/\gamma) + \log\left(1 + \frac{1}{d\lambda(1-\gamma)}\right)\right) \,. \quad \text{(Corollary 5.28)}$$

For $r_{2,t}$, we have

$$\sum_{t=1}^{T} r_{2,t} \leq \frac{4k_\mu^2}{\lambda^2} \frac{\gamma^{2D}T}{(1-\gamma)^2}(2Sk_\mu + m)^2 \,.$$

Furthermore, $r_{3,t}$ is treated as follows:

$$\sum_{t=1}^{T} r_{3,t} \leq \frac{8k_\mu}{\lambda} \frac{\beta_T^\delta}{\sqrt{c_\mu}} \frac{\gamma^D}{1-\gamma}(2Sk_\mu + m) \sum_{t=1}^{T} \|A_t\|_{V_t^{-1}}$$
$$\leq \frac{8k_\mu}{\lambda} \frac{\beta_T^\delta}{\sqrt{c_\mu}} \frac{\gamma^D}{1-\gamma}(2Sk_\mu + m)\sqrt{T}\sqrt{\sum_{t=1}^{T} \|A_t\|_{V_t^{-1}}^2}$$
$$\leq \frac{8k_\mu \beta_T^\delta}{\lambda\sqrt{c_\mu}} \frac{\gamma^D}{1-\gamma}(2Sk_\mu + m)\sqrt{2dT\max\left(1, \frac{1}{\lambda}\right)}\sqrt{T\log\left(\frac{1}{\gamma}\right) + \log\left(1 + \frac{1}{d\lambda(1-\gamma)}\right)}.$$

When $\lambda = d\log(T)$, $D = \frac{\log(T)}{\log(1/\gamma)}$ and $\gamma = 1 - \frac{\sqrt{c_\mu \Gamma_T}}{d\sqrt{T}}$, we can upper bound the different terms following the proof of Theorem 5.2.
With those choices,

1. $\Gamma_T D$ scales as $\widetilde{\mathcal{O}}(c_\mu^{-1/2} d \Gamma_T^{1/2} T^{1/2})$
2. $\sum_{t=1}^{T} r_{1,t}$ scales as $\widetilde{\mathcal{O}}(c_\mu^{-1/2} d \Gamma_T^{1/2} T^{1/2})$
3. $\sum_{t=1}^{T} r_{2,t}$ scales as $\widetilde{\mathcal{O}}(c_\mu^{-1} \Gamma_T^{-1})$
4. $\sum_{t=1}^{T} r_{3,t}$ scales as $\widetilde{\mathcal{O}}(d^{1/4} c_\mu^{-3/4} \Gamma_T^{-1/4} T^{1/4})$

Keeping the highest order term in T and dividing by Δ yields the announced result. □

5.B.4 Refined Exploration Bonus when $\hat{\theta}_t \in \Theta$

As briefly explained in Remark 5.1 in the main text, when the MLE is an admissible parameter ($\hat{\theta}_t \in \Theta$) it is possible to obtain a usually tighter concentration result. In this section, we explain exactly how this can be done. Note that this improvement is mostly useful for the design of the algorithm and has no impact on the regret guarantees.

We define

$$\bar{\beta}_T^\delta = k_\mu \sqrt{1+2S}\left(\sqrt{\lambda}S + \rho_T^\delta\right) , \tag{5.32}$$

where ρ_T^δ is defined in Equation (5.26).

Proposition 5.16. *For any $\delta \in (0,1]$, with probability higher than $1-\delta$,*

$$\forall t \in \mathcal{T}(\gamma)\ s.t\ \hat{\theta}_t \in \Theta,\ \Delta_t(a,\hat{\theta}_t) \leq \frac{k_\mu}{\lambda}\frac{\gamma^D}{1-\gamma}(2Sk_\mu + m) + \frac{\bar{\beta}_T^\delta}{\sqrt{c_\mu}}\|a\|_{V_t^{-1}} .$$

Proof. We use the notation G_t (respectively $\widetilde{G}_t$) instead of $G_t(\theta_t^\star, \hat{\theta}_t)$ (respectively $\widetilde{G}_t(\theta_t^\star, \hat{\theta}_t)$). Following the same steps as for the proof of Proposition 5.15, one gets

$$\begin{aligned}
\Delta_t(a,\hat{\theta}_t) &\leq \frac{k_\mu}{\lambda}\frac{\gamma^D}{1-\gamma}(2Sk_\mu + m) + k_\mu |a^\top G_t^{-1}(\gamma^t S_{t-D:t} - \lambda\theta_t^\star)| \\
&\leq \frac{k_\mu}{\lambda}\frac{\gamma^D}{1-\gamma}(2Sk_\mu + m) + \|a\|_{G_t^{-1}\widetilde{G}_t G_t^{-1}}\|\gamma^t S_{t-D:t} - \lambda\theta_t^\star\|_{\widetilde{G}_t^{-1}} \\
&\leq \frac{k_\mu}{\lambda}\frac{\gamma^D}{1-\gamma}(2Sk_\mu + m) + \|a\|_{G_t^{-1}}\|\gamma^t S_{t-D:t} - \lambda\theta_t^\star\|_{\widetilde{G}_t^{-1}} . \quad \text{(Equation (5.28))}
\end{aligned}$$

Here, with the additional assumption $\hat{\theta}_t \in \Theta$, the self-concordance can be used to obtain an easier relation between $\widetilde{G}_t$ and $\widetilde{H}_t$ as stated in Lemma 5.25.

$$\begin{aligned}
\Delta_t(a,\hat{\theta}_t) &\leq \frac{k_\mu}{\lambda}\frac{\gamma^D}{1-\gamma}(2Sk_\mu + m) + \sqrt{1+2S}\|a\|_{G_t^{-1}}\|\gamma^{t-1} S_{t-D:t} - \lambda\theta_t^\star\|_{\widetilde{H}_t^{-1}} \quad \text{(Lemma 5.25)} \\
&\leq \frac{k_\mu}{\lambda}\frac{\gamma^D}{1-\gamma}(2Sk_\mu + m) + \sqrt{1+2S}\|a\|_{G_t^{-1}}\|\gamma^{t-1} S_{t-D:t} - \lambda\theta_t^\star\|_{\widetilde{H}_{t-D:t}^{-1}} .
\end{aligned}$$

The last inequality uses $\widetilde{H}_{t-D:t} \leq \widetilde{H}_t$. Now by applying Corollary 5.10, $\Delta_t(a,\hat{\theta}_t)$ can be further upper bounded.

$$\Delta_t(a,\hat{\theta}_t) \leq \frac{k_\mu}{\lambda}\frac{\gamma^D}{1-\gamma}(2Sk_\mu + m) + \sqrt{1+2S}\|a\|_{G_t^{-1}}\left(\sqrt{\lambda}S + \rho_T^\delta\right) .$$

The final step consists in using $G_t := G_t(\theta_t^\star, \hat{\theta}_t) \geq c_\mu V_t$ which holds because both $\hat{\theta}_t$ and $\theta_t^\star$ are in Θ. □

Consequently, when $\hat{\theta}_t \in \Theta$, the action a_t at time t can be chosen according to:

$$\begin{aligned}
A_t &= \text{argmax}_{a\in\mathcal{A}_t}\left(\mu(a^\top\hat{\theta}_t) + \frac{\bar{\beta}_T^\delta}{\sqrt{c_\mu}}\|a\|_{V_t^{-1}} + \frac{k_\mu}{\lambda}\frac{\gamma^D}{1-\gamma}(2Sk_\mu + m)\right) \\
&= \text{argmax}_{a\in\mathcal{A}_t}\left(\mu(a^\top\hat{\theta}_t) + \frac{\bar{\beta}_T^\delta}{\sqrt{c_\mu}}\|a\|_{V_t^{-1}}\right) .
\end{aligned} \tag{5.33}$$

Appendix 5.C Regret Analysis with a Sliding Window

In the main text only the analysis with discount factors is discussed. However as in the linear bandit literature, the analysis with exponential weights and a sliding window share similarities, in particular they have the same form of guarantees for the regret. For the sake of completeness, we give a detailed analysis of the results achievable with a sliding window.

5.C.1 Notation

Let us first introduce the main notations. For any value of $\theta \in \mathbb{R}^d$, we define,

$$H_t(\theta) = \sum_{s=\max(1,t-\tau+1)}^{t} \dot{\mu}(A_s^\top \theta) A_s A_s^\top + \lambda I_d \ . \tag{5.34}$$

$$V_t = \sum_{s=\max(1,t-\tau+1)}^{t} A_s A_s^\top + \frac{\lambda}{c_\mu} I_d \ . \tag{5.35}$$

$$g_t(\theta) = \sum_{s=\max(1,t-\tau+1)}^{t} \mu(A_s^\top \theta) A_s + \lambda \theta \ . \tag{5.36}$$

$$S_t = \sum_{s=\max(1,t-\tau+1)}^{t} \epsilon_s A_s \ . \tag{5.37}$$

For any $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\alpha(a, \theta_1, \theta_2) = \int_0^1 \dot{\mu}(v a^\top \theta_2 + (1-v) a^\top \theta_1) dv \ .$$

$$G_t(\theta_1, \theta_2) = \sum_{s=\max(1,t-\tau+1)}^{t} \alpha(A_s, \theta_1, \theta_2) A_s A_s^\top + \lambda I_d \ . \tag{5.38}$$

Let H_t be defined as

$$H_t = \sum_{s=\max(1,t-\tau+1)}^{t} \dot{\mu}(A_s^\top \theta_s^\star) A_s A_s^\top + \lambda I_d \ . \tag{5.39}$$

Let us define $\mathcal{T}(\tau)$ as

$$\mathcal{T}(\tau) = \{1 \leq t \leq T, \forall s, \text{such that } t - \tau + 1 \leq s \leq t, \theta_s^\star = \theta_t^\star\} \ . \tag{5.40}$$

$t \in \mathcal{T}(\tau)$ when t is a least τ steps away from the closest previous breakpoint. When focusing on time instants in $\mathcal{T}(\tau)$ the bias due to non-stationarity disappears. In the sliding window setting, we construct an estimator based on a truncated penalized log-likelihood.

In this section, $\hat{\theta}_t$ is defined as the unique maximizer of

$$\sum_{s=\max(1,t-\tau+1)}^{t} \log \mathbb{P}_\theta(X_s | A_s) - \frac{\lambda}{2} \|\theta\|_2^2 \ . \tag{5.41}$$

By using the definition of the GLM and thanks to the concavity of this equation in θ, $\hat{\theta}_t$ is the unique solution of

$$\sum_{s=\max(1,t-\tau+1)}^{t} (X_s - \mu(A_s^\top \theta))A_s - \lambda\theta = 0 \,.$$

This can be summarized with

$$g_t(\hat{\theta}_t) = \sum_{s=\max(1,t-\tau+1)}^{t} X_s A_s = S_t + \sum_{s=\max(1,t-\tau+1)}^{t} \mu(A_s^\top \theta_s^\star)A_s \,.$$

5.C.2 Algorithm

The SC-SW-GLUCB algorithm proceeds as follows. First, based on the τ last rewards and actions, $\hat{\theta}_{t-1}$ is computed using Equation (5.41). Then, after receiving the action set $\mathcal{A}_t$ the action A_t is chosen optimistically. Finally, by proposing this action a reward X_t is received and the design matrix is updated. The pseudo code of SC-SW-GLUCB is reported in Algorithm 12.

Input: Failure probability δ, dimension d, regularization λ, upper bound for actions L, upper bound for parameters S, sliding window τ.
Initialization: $V_0 = (\lambda/c_\mu)I_d$, $\hat{\theta}_0 = 0_{\mathbb{R}^d}$
for $t = 1$ ***to*** T **do**
  Receive $\mathcal{A}_t$, compute $\hat{\theta}_{t-1}$ according to (5.41)
  Play $A_t = \text{argmax}_{a\in\mathcal{A}_t} \mu(a^\top \hat{\theta}_{t-1}) + \frac{\beta_t^\delta}{\sqrt{c_\mu}} \|a\|_{V_{t-1}^{-1}}$ with β_t^δ defined in Equation (5.43)
  Receive reward X_t
  Update:
  if $t < \tau$ **then**
    $V_{t+1} \leftarrow A_t A_t^\top + V_t$
  else
    $V_{t+1} \leftarrow A_t A_t^\top - A_{t-\tau} A_{t-\tau}^\top + V_t$

Algorithm 12: SC-SW-GLUCB

5.C.3 Analysis of the Regret of SC-SW-GLUCB

In Section 5.B, the self-concordance is the key tool to obtain an analysis without using a projection step. In the next proposition, we link the matrix $G_t(\hat{\theta}_t, \theta_t^\star)$ with $H_t(\theta_t^\star)$ independently from c_μ.

Proposition 5.17. *When $\hat{\theta}_t$ is the maximum likelihood estimator as defined in Equation (5.41) and $t \in \mathcal{T}(\tau)$, we have:*

$$\alpha(a, \theta_t^\star, \hat{\theta}_t) \geq \left(1 + S + \frac{1}{\sqrt{\lambda}} \|S_t\|_{G_t^{-1}(\theta_t^\star, \hat{\theta}_t)}\right)^{-1} \dot{\mu}(a^\top \theta_t^\star) \,.$$

Note that the main difference with Proposition 5.12 is that $\bar{S}$ is now replaced by S. This is due to the fact that the bias disappears when using a sliding window for $t \in \mathcal{T}(\tau)$.

Proof. Thanks to Lemma 5.23, we have:

$$\begin{aligned}
\alpha(a,\theta_t^\star,\hat{\theta}_t) &\geq \left(1+\left|a^\top(\theta_t^\star-\hat{\theta}_t)\right|\right)^{-1}\dot{\mu}(a^\top\theta_t^\star)\\
&\geq \left(1+\left|a^\top G_t^{-1}(\theta_t^\star,\hat{\theta}_t)(g_t(\theta_t^\star)-g_t(\hat{\theta}_t))\right|\right)^{-1}\dot{\mu}(a^\top\theta_t^\star) \quad \text{(Mean-Value Theorem)}\\
&\geq \left(1+\|a\|_{G_t^{-1}(\theta_t^\star,\hat{\theta}_t)}\left\|g_t(\theta_t^\star)-g_t(\hat{\theta}_t)\right\|_{G_t^{-1}(\theta_t^\star,\hat{\theta}_t)}\right)^{-1}\dot{\mu}(a^\top\theta_t^\star) \quad \text{(Cauchy–Schwarz)}\\
&\geq \left(1+\lambda^{-1/2}\left\|g_t(\theta_t^\star)-g_t(\hat{\theta}_t)\right\|_{G_t^{-1}(\theta_t^\star,\hat{\theta}_t)}\right)^{-1}\dot{\mu}(a^\top\theta_t^\star) \quad (G_t(\theta_t^\star,\hat{\theta}_t)\geq\lambda I_d)\\
&\geq \left(1+\lambda^{-1/2}\left\|S_t-\lambda\theta_t^\star\right\|_{G_t^{-1}(\theta_t^\star,\hat{\theta}_t)}\right)^{-1}\dot{\mu}(a^\top\theta_t^\star) \quad (t\in\mathcal{T}(\tau))\\
&\geq \left(1+S+\lambda^{-1/2}\left\|S_t\right\|_{G_t^{-1}(\theta_t^\star,\hat{\theta}_t)}\right)^{-1}\dot{\mu}(a^\top\theta_t^\star)\,.
\end{aligned}$$

□

Corollary 5.18. *When $\hat{\theta}_t$ is the maximum likelihood estimator as defined in Equation (5.41), when $t\in\mathcal{T}(\tau)$ and H_t is defined in Equation (5.39), we have,*

$$G_t(\theta_t^*,\hat{\theta}_t)\geq\left(1+S+\frac{1}{\sqrt{\lambda}}\left\|S_t\right\|_{G_t^{-1}(\theta_t^*,\hat{\theta}_t)}\right)^{-1}H_t\,.$$

Furthermore,

$$\forall t\leq T,\ \left\|S_t\right\|_{G_t^{-1}(\theta_t^\star,\hat{\theta}_t)}\leq\sqrt{1+S}\left\|S_t\right\|_{H_t^{-1}}+\frac{1}{\sqrt{\lambda}}\left\|S_t\right\|_{H_t^{-1}}^2\,.$$

Proof. Using Proposition 5.17 and summing for time instants s such that $\max(1,t-\tau+1)\leq s\leq t$,

$$\sum_{s=t-\tau+1}^{t}\alpha(A_s,\theta_t^\star,\hat{\theta}_t)A_sA_s^\top\geq\left(1+S+\lambda^{-1/2}\left\|S_t\right\|_{G_t^{-1}(\theta_t^\star,\hat{\theta}_t)}\right)^{-1}\sum_{s=t-\tau+1}^{t}\dot{\mu}(A_s^\top\theta_s^\star)A_sA_s^\top\,.$$

Where we use $\theta_s^\star=\theta_t^\star$ for $t-\tau+1\leq s\leq t$ thanks to the assumption $t\in\mathcal{T}(\tau)$. The next step consists in adding the regularization term on both sides. Note that $\left(1+S+\lambda^{-1/2}\left\|S_t\right\|_{G_t^{-1}(\theta_t^\star,\hat{\theta}_t)}\right)\lambda\geq\lambda$ and obtain,

$$G_t(\theta_t^\star,\hat{\theta}_t)\geq\left(1+S+\lambda^{-1/2}\left\|S_t\right\|_{G_t^{-1}(\theta_t^\star,\hat{\theta}_t)}\right)^{-1}H_t\,.$$

This in turn implies,

$$\begin{aligned}
&\left\|S_t\right\|_{G_t^{-1}(\theta_t^\star,\hat{\theta}_t)}^2\leq\left(1+S+\lambda^{-1/2}\left\|S_t\right\|_{G_t^{-1}(\theta_t^\star,\hat{\theta}_t)}\right)\left\|S_t\right\|_{H_t^{-1}}^2\\
\iff&\left\|S_t\right\|_{G_t^{-1}(\theta_t^\star,\hat{\theta}_t)}^2-\lambda^{-1/2}\left\|S_t\right\|_{H_t^{-1}}^2\left\|S_t\right\|_{G_t^{-1}(\theta_t^\star,\hat{\theta}_t)}-(1+S)\left\|S_t\right\|_{H_t^{-1}}^2\leq 0\,.
\end{aligned}$$

Solving this polynomial inequality (in $\left\|S_t\right\|_{G_t^{-1}(\theta_t^\star,\hat{\theta}_t)}$) finally gives,

$$\left\|S_t\right\|_{G_t^{-1}(\theta_t^\star,\hat{\theta}_t)}\leq\sqrt{1+S}\left\|S_t\right\|_{H_t^{-1}}+\frac{1}{\sqrt{\lambda}}\left\|S_t\right\|_{H_t^{-1}}^2\,.$$

□

Using this technique, we have established an explicit link between $G_t(\theta_t^\star, \hat{\theta}_t)$ and H_t without the need to project $\hat{\theta}_t$ on Θ when $t \in \mathcal{T}(\tau)$.

We define

$$\rho_t^\delta = \left(\frac{\sqrt{\lambda}}{2m} + \frac{2m}{\sqrt{\lambda}} \log\left(\frac{T}{\delta}\right) + \frac{dm}{\sqrt{\lambda}} \log\left(1 + \frac{k_\mu \min(t,\tau)}{d\lambda}\right) + \frac{2m}{\sqrt{\lambda}} d\log(2) \right) , \tag{5.42}$$

and

$$\beta_t^\delta = k_\mu \sqrt{\lambda} \left(1 + S + \sqrt{\frac{1+S}{\lambda}} \rho_t^\delta + \left(\frac{\rho_t^\delta}{\sqrt{\lambda}} \right)^2 \right)^{3/2} . \tag{5.43}$$

In the next proposition, we give an upper bound for $\Delta_t(a, \hat{\theta}_t)$.

Proposition 5.19. *For any $\delta \in (0,1]$, with probability higher than $1-\delta$,*

$$\forall t \in \mathcal{T}(\tau),\ \Delta_t(a, \hat{\theta}_t) \leq \frac{\beta_t^\delta}{\sqrt{c_\mu}} \|a\|_{V_t^{-1}} .$$

Proof.

$$\begin{aligned}
\Delta_t(a, \hat{\theta}_t) &= |\mu(a^\top \theta_t^\star) - \mu(a^\top \hat{\theta}_t)| \leq k_\mu |a^\top (\theta_t^\star - \hat{\theta}_t)| \\
&= k_\mu |a^\top G_t^{-1}(\theta_t^\star, \hat{\theta}_t)(g_t(\theta_t^\star) - g_t(\hat{\theta}_t))| \quad \text{(Mean-Value Theorem)} \\
&\leq k_\mu \|a\|_{G_t^{-1}(\theta_t^\star, \hat{\theta}_t)} \|g_t(\theta_t^\star) - g_t(\hat{\theta}_t)\|_{G_t^{-1}(\theta_t^\star, \hat{\theta}_t)} \quad \text{(Cauchy-Schwarz ineq.)} \\
&\leq k_\mu \|a\|_{G_t^{-1}(\theta_t^\star, \hat{\theta}_t)} \|S_t - \lambda \theta_t^\star\|_{G_t^{-1}(\theta_t^\star, \hat{\theta}_t)} . \quad (t \in \mathcal{T}(\tau))
\end{aligned}$$

We can use Corollary 5.18 to link $\|a\|_{G_t^{-1}(\theta_t^\star, \hat{\theta}_t)}$ with $\|a\|_{H_t^{-1}}$.

$$\begin{aligned}
\Delta_t(a, \hat{\theta}_t) &\leq k_\mu \sqrt{1 + S + \frac{1}{\sqrt{\lambda}} \|S_t\|_{G_t^{-1}(\theta_t^\star, \hat{\theta}_t)}} \|a\|_{H_t^{-1}} \left(\sqrt{\lambda} S + \|S_t\|_{G_t^{-1}(\theta_t^\star, \hat{\theta}_t)} \right) \\
&\leq k_\mu \sqrt{\lambda} \sqrt{1 + S + \frac{1}{\sqrt{\lambda}} \|S_t\|_{G_t^{-1}(\theta_t^\star, \hat{\theta}_t)}} \|a\|_{H_t^{-1}} \left(S + \frac{1}{\sqrt{\lambda}} \|S_t\|_{G_t^{-1}(\theta_t^\star, \hat{\theta}_t)} \right) \\
&\leq k_\mu \sqrt{\lambda} \left(1 + S + \frac{1}{\sqrt{\lambda}} \|S_t\|_{G_t^{-1}(\theta_t^\star, \hat{\theta}_t)} \right)^{3/2} \|a\|_{H_t^{-1}} .
\end{aligned}$$

Then, using Corollary 5.18 we can upper bound $\|S_t\|_{G_t^{-1}(\theta_t^\star, \hat{\theta}_t)}$ with a combination of terms depending on $\|S_t\|_{H_t^{-1}}$. Recall that Corollary 5.11 gives with probability higher than $1-\delta$, for all t in $\mathcal{T}(\tau)$, $\|S_t\|_{H_t^{-1}} \leq \rho_t^\delta$.

$$\Delta_t(a, \hat{\theta}_t) \leq k_\mu \sqrt{\lambda} \left(1 + S + \sqrt{\frac{1+S}{\lambda}} \rho_t^\delta + \frac{1}{\lambda} (\rho_t^\delta)^2 \right)^{3/2} \|a\|_{H_t^{-1}} .$$

The proof is completed using $H_t \geq c_\mu V_t$, which holds thanks to Assumption 5.1 on the bandit parameters. □

Finally, we give an upper bound for the regret enjoyed by SC-SW-GLUCB.

Theorem 5.20. *The regret of the* SC-SW-GLUCB *algorithm is bounded with probability at least* $1-\delta$ *by,*

$$R(T) \leq \Gamma_T \tau + \frac{2\beta_T^\delta}{\sqrt{c_\mu}}\sqrt{dT}\sqrt{\lceil T/\tau \rceil}\sqrt{2\max\left(1, \frac{1}{\lambda}\right)}\sqrt{\log\left(1+\frac{\tau}{d\lambda}\right)},$$

where β_t^δ *is defined in Equation* (5.43).

Proof. The proof essentially follows the steps of the proof of Theorem 5.2. The main difference is that β_t^δ from Equation (5.43) is used and the elliptical lemma is different because the design matrix used a sliding window instead of weights.
Applying Proposition 5.32 when $t \in \mathcal{T}(\tau)$, with probability higher than $1-\delta$,

$$r_t \leq \frac{2}{\sqrt{c_\mu}}\beta_t^\delta \|a_t\|_{V_t^{-1}}. \tag{5.44}$$

The dynamic regret can then be upper bounded by,

$$\begin{aligned}
R(T) &= \sum_{t=1}^{T} r_t = \sum_{t\in\mathcal{T}(\tau)} r_t + \sum_{t\notin\mathcal{T}(\tau)} r_t \leq \Gamma_T\tau + \sum_{t\in\mathcal{T}(\tau)} r_t \\
&\leq \Gamma_T\tau + \frac{2\beta_T^\delta}{\sqrt{c_\mu}} \sum_{t\in\mathcal{T}(\tau)} \|A_t\|_{V_t^{-1}} \quad \text{(Equation (5.44))} \\
&\leq \Gamma_T\tau + \frac{2\beta_T^\delta}{\sqrt{c_\mu}}\sqrt{T}\sqrt{\sum_{t\in\mathcal{T}(\tau)} \|A_t\|^2_{V_t^{-1}}} \quad \text{(Cauchy-Schwarz ineq.)} \\
&\leq \Gamma_T\tau + \frac{2\beta_T^\delta}{\sqrt{c_\mu}}\sqrt{T}\sqrt{\sum_{t=1}^{T} \|A_t\|^2_{V_t^{-1}}} \\
&\leq \Gamma_T\tau + + \frac{2\beta_T^\delta}{\sqrt{c_\mu}}\sqrt{dT}\sqrt{\lceil T/\tau \rceil}\sqrt{2\max\left(1, \frac{1}{\lambda}\right)\log\left(1+\frac{\tau}{d\lambda}\right)}. \quad \text{(Lemma 5.31)}
\end{aligned}$$

□

Corollary 5.21 (Asymptotic bound)**.** *If* Γ_T *is known, by choosing* $\tau = \left(\frac{dT}{c_\mu^{1/2}\Gamma_T}\right)^{2/3}$ *and* $\lambda = d\log(T)$*, the regret of* SC-SW-GLUCB *scales as*

$$R(T) = \widetilde{\mathcal{O}}(c_\mu^{-1/3} d^{2/3} \Gamma_T^{1/3} T^{2/3}).$$

If Γ_T *is unknown, by choosing* $\tau = \left(\frac{dT}{c_\mu^{1/2}}\right)^{2/3}$*, the regret of* SC-SW-GLUCB *scales as*

$$R(T) = \widetilde{\mathcal{O}}(c_\mu^{-1/3} d^{2/3} \Gamma_T T^{2/3}).$$

Proof. When Γ_T is known, we set $\lambda = d\log(T)$ and $\tau = \left(\frac{dT}{\sqrt{c_\mu}\Gamma_T}\right)^{2/3}$. With those choices,

1. β_T^δ scales as $\sqrt{d\log(T)}$.

2. $\Gamma_T\tau$ scales as $\widetilde{\mathcal{O}}(c_\mu^{-1/3}d^{2/3}\Gamma_T^{2/3}T^{2/3})$.
3. $\frac{\beta_T^\delta}{\sqrt{c_\mu}}\sqrt{T}\sqrt{d\frac{T}{\tau}}$ scales as $\widetilde{\mathcal{O}}(c_\mu^{-1/3}d^{2/3}\Gamma_T^{1/3}T^{2/3})$.

The proof is similar when Γ_T is unknown. □

When the reward gaps are bounded from below we can obtain the following gap-dependent upper bound:

Theorem 5.22. *Under Assumption 5.5, when setting* $\tau = \frac{d\sqrt{T}}{\sqrt{c_\mu \Gamma_T}}$ *the regret of the algorithm* SC-SW-GLUCB *satisfies:*

$$R_T = \widetilde{\mathcal{O}}(\Delta^{-1}c_\mu^{-1/2}d\sqrt{\Gamma_T T}) \ .$$

Proof. First note that for any suboptimal action $a \in \mathcal{A}_t$,

$$\mu(a_{\star,t}^\top \theta_t^\star) - \mu(a^\top \theta_t^\star) \geq \Delta \ .$$

This implies

$$r_t = \mu(a_{\star,t}^\top \theta_t^\star) - \mu(a_t^\top \theta_t^\star) \leq \frac{\left(\mu(a_{\star,t}^\top \theta_t^\star) - \mu(a_t^\top \theta_t^\star)\right)^2}{\Delta} = \frac{r_t^2}{\Delta} \ . \tag{5.45}$$

Using Proposition 5.32 one has,

$$r_t \leq \frac{2}{\sqrt{c_\mu}}\beta_t^\delta \|a_t\|_{V_t^{-1}} \ .$$

The dynamic regret can then be upper bounded by,

$$\begin{aligned} R(T) &\leq \Gamma_T\tau + \frac{1}{\Delta}\sum_{t\in\mathcal{T}(\tau)} r_t^2 \quad \text{(Equation (5.45))} \\ &\leq \Gamma_T\tau + \frac{4(\beta_T^\delta)^2}{c_\mu\Delta}\sum_{t=1}^{T}\|a_t\|_{V_t^{-1}}^2 \\ &\leq \Gamma_T\tau + \frac{8(\beta_T^\delta)^2}{c_\mu\Delta}\max\left(1,\frac{1}{\lambda}\right) d\lceil T/\tau\rceil \log\left(1+\frac{\tau}{\lambda d}\right) \ . \quad \text{(Lemma 5.31)} \end{aligned}$$

We set $\lambda = d\log(T)$ and $\tau = \frac{d\sqrt{T}}{\sqrt{c_\mu\Gamma_T}}$. With those choices,

1. β_T^δ scales as $\sqrt{d\log(T)}$.
2. $\Gamma_T\tau$ scales as $\widetilde{\mathcal{O}}(c_\mu^{-1/2}d\Gamma_T^{1/2}T^{1/2})$.
3. $\frac{(\beta_T^\delta)^2}{c_\mu}d\frac{T}{\tau}$ scales as $\widetilde{\mathcal{O}}(c_\mu^{-1/2}d\Gamma_T^{1/2}T^{1/2})$.

Dividing by Δ yields the announced result. □

When $\hat{\theta}_t$ is in Θ it is also possible with a sliding window to obtain a usually better concentration result. This discussion is not reported here, but can be easily adapted from Proposition 5.16.

Appendix 5.D Useful Results

5.D.1 Self-Concordant Properties

In this section we state the main properties and lemma that can be obtained with the self-concordance assumption.

Lemma 5.23 (Lemma 9 in [Faury et al., 2020]). *For any $z_1, z_2 \in \mathbb{R}$, we have the following inequality*

$$\dot{\mu}(z_1)\frac{1-\exp(-|z_1-z_2|)}{|z_1-z_2|} \leq \int_0^1 \dot{\mu}(z_1+v(z_2-z_1))dv \leq \dot{\mu}(z_1)\frac{\exp(|z_1-z_2|)-1}{|z_1-z_2|} \ .$$

Furthermore,

$$\int_0^1 \dot{\mu}(z_1+v(z_2-z_1))dv \geq \dot{\mu}(z_1)(1+|z_1-z_2|)^{-1} \ .$$

Thanks to the self-concordance property we have an interesting relation between $G_t(\theta_1,\theta_2)$ and $H_t(\theta_1)$ or $H_t(\theta_2)$ when both θ_1 and $\theta_2 \in \Theta$. This relation is made explicit in the next lemma.

Lemma 5.24 (Self-concordance and sliding window). *For all $\theta_1, \theta_2 \in \Theta$, with G_t defined in Equation (5.38) and H_t defined in Equation (5.34) the following inequalities hold*

$$G_t(\theta_1,\theta_2) \geq (1+2S)^{-1}H_t(\theta_1) \ , \quad G_t(\theta_1,\theta_2) \geq (1+2S)^{-1}H_t(\theta_2) \ .$$

Proof. Applying Lemma 5.23, for any $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\alpha(a,\theta_1,\theta_2) \geq \frac{\dot{\mu}(a^\top\theta_1)}{1+|a^\top(\theta_1-\theta_2)|} \quad \text{and} \quad \alpha(a,\theta_1,\theta_2) \geq \frac{\dot{\mu}(a^\top\theta_2)}{1+|a^\top(\theta_1-\theta_2)|} \ .$$

Furthermore, if θ_1 and $\theta_2 \in \Theta$, then

$$|a^\top(\theta_1-\theta_2)| \leq 2S \ .$$

□

Lemma 5.25 (Self-concordance and discount factors). *For all $\theta_1, \theta_2 \in \Theta$, with $\widetilde{H}_t(\theta_1)$ defined in Equation (5.16) and $\widetilde{G}_t(\theta_1,\theta_2)$ defined in Equation (5.22) the following inequalities hold:*

$$\widetilde{G}_t(\theta_1,\theta_2) \geq (1+2S)^{-1}\widetilde{H}_t(\theta_1) \ , \quad \widetilde{G}_t(\theta_1,\theta_2) \geq (1+2S)^{-1}\widetilde{H}_t(\theta_2) \ .$$

Proof. Same arguments than for Lemma 5.24 □

5.D.2 Determinant Inequalities

Proposition 5.26 (Determinant inequality)**.** *Let $(\lambda_t)_t$ be a deterministic sequence of regularization parameters. Let $H_t \coloneqq \sum_{s=1}^{t} w_s^2 \sigma_s^2 A_s A_s^\top + \lambda_t I_d$. Under the Assumption 5.1 and $\forall t, \sigma_t^2 \leq k_\mu$, the following holds*

$$\det(H_t) \leq \left(\lambda_t + \frac{k_\mu \sum_{s=1}^{t} w_s^2}{d}\right)^d .$$

Proof.

$$\begin{aligned}
\det(H_t) = \prod_{i=1}^{d} l_i \quad (l_i \text{ are the eigenvalues}) &\leq \left(\frac{1}{d}\sum_{i=1}^{d} l_i\right)^d \quad \text{(AM-GM inequality)} \\
&\leq \left(\frac{1}{d}\text{trace}(H_t)\right)^d \leq \left(\frac{1}{d}\sum_{s=1}^{t} w_s^2\sigma_s^2 \text{trace}(A_s A_s^\top) + \lambda_t\right)^d \\
&\leq \left(\frac{1}{d}\sum_{s=1}^{t} w_s^2\sigma_s^2 \|A_s\|_2^2 + \lambda_t\right)^d \leq \left(\lambda_t + \frac{k_\mu}{d}\sum_{s=1}^{t} w_s^2\right)^d .
\end{aligned}$$

□

Corollary 5.27. *Let $\{A_s\}_{s=1}^{\infty}$ a sequence in $\mathbb{R}^d$ such that $\|A_s\|_2 \leq L$ for all $s \in \mathbb{N}^*$, and let λ be a non-negative scalar. In the specific case where the weights are given by $w_t = \gamma^{-t}$ with $0 < \gamma < 1$, under the same assumptions than Proposition 5.26, with $\widetilde{H}_t \coloneqq \sum_{s=t-t_0+1}^{t} \gamma^{2(t-s)} \sigma_s^2 A_s A_s^\top + \lambda I_d$, one has*

$$\det(\widetilde{H}_t) \leq \left(\lambda + \frac{k_\mu L(1-\gamma^{2t_0})}{d(1-\gamma^2)}\right)^d .$$

Corollary 5.28. *Let $\{A_s\}_{s=1}^{\infty}$ a sequence in $\mathbb{R}^d$ such that $\|A_s\|_2 \leq L$ for all $s \in \mathbb{N}^*$, and let λ be a non-negative scalar. For $t \geq 1$ define $V_t \coloneqq \sum_{s=1}^{t} \gamma^{t-s} A_s A_s^\top + \lambda I_d$. The following inequality holds:*

$$\det(V_t) \leq \left(\lambda + \frac{L^2(1-\gamma^t)}{d(1-\gamma)}\right)^d .$$

Corollary 5.29. *Let $\{A_s\}_{s=1}^{\infty}$ a sequence in $\mathbb{R}^d$ such that $\|A_s\|_2 \leq L$ for all $s \in \mathbb{N}^*$, and let λ be a non-negative scalar. With $H_t \coloneqq \sum_{s=\max(1,t-\tau+1)}^{t} \sigma_s^2 A_s A_s^\top + \lambda I_d$, one has*

$$\det(H_t) \leq \left(\lambda + \frac{k_\mu L \min(t,\tau)}{d}\right)^d .$$

5.D.3 Elliptical Lemma

The following lemma is a version of the Elliptical Lemma when discount factors are used. It comes from Proposition 4.14 from Chapter 4 and is stated here for the sake of completeness.

Lemma 5.30 (Elliptical potential with discount factors (based on Proposition 4.14))**.** *Let $\{A_s\}_{s=1}^{\infty}$ a sequence in $\mathbb{R}^d$ such that $\|A_s\|_2 \leq 1$ for all $s \in \mathbb{N}$, and let λ be a non-negative scalar. For $t \geq 1$ define $V_t \coloneqq \sum_{s=1}^{t} \gamma^{t-s} A_s A_s^\top + \lambda I_d$, the following inequality holds*

$$\sum_{t=1}^{T} \|A_t\|^2_{V_{t-1}^{-1}} \leq 2 \max\left(1, \frac{1}{\lambda}\right) \log\left(\frac{\det(V_T)}{\lambda^d \gamma^{dT}}\right) .$$

Proof. In the proof we introduce the matrix $W_t = \sum_{s=1}^{t} \gamma^{-s} A_s A_s^\top + \gamma^{-t}\lambda I_d$ such that $V_t = \gamma^t W_t$. We have,

$$\begin{aligned}
W_t &= \sum_{s=1}^{t} \gamma^{-s} A_s A_s^\top + \gamma^{-t}\lambda I_d \\
&= \gamma^{-t} A_t A_t^\top + \sum_{s=1}^{t-1} \gamma^{-s} A_s A_s^\top + \gamma^{-(t-1)}\lambda I_d + \gamma^{-t}\lambda I_d - \gamma^{-(t-1)}\lambda I_d \\
&= \gamma^{-t} A_t A_t^\top + \gamma^{-t}(1-\gamma)\lambda I_d + W_{t-1} \\
&\geq \gamma^{-t} A_t A_t^\top + W_{t-1} \geq W_{t-1}^{1/2}\left(I_d + \gamma^{-t} W_{t-1}^{-1/2} A_t A_t^\top W_{t-1}^{-1/2}\right) W_{t-1}^{1/2} .
\end{aligned}$$

This implies,

$$\begin{aligned}
\det(W_t) &\geq \det(W_{t-1}) \det\left(I_d + (\gamma^{-t/2} W_t^{-1/2} A_t)(\gamma^{-t/2} W_t^{-1/2} A_t)^\top\right) \\
&\geq \det(W_{t-1})\left(1 + \gamma^{-t}\|A_t\|^2_{W_t^{-1}}\right) \quad (\det(I_d + xx^\top) = 1 + \|x\|_2^2) .
\end{aligned}$$

This in turn gives,

$$\frac{\det(W_T)}{\det(W_0)} = \prod_{t=0}^{T-1} \frac{\det(W_{t+1})}{\det(W_t)} \geq \prod_{t=1}^{T}\left(1 + \gamma^{-(t+1)}\|A_{t+1}\|^2_{W_t^{-1}}\right) .$$

Taking the logarithm on both sides gives:

$$\begin{aligned}
\log\left(\frac{\det(W_T)}{\lambda^d}\right) &\geq \sum_{t=0}^{T-1} \log(1 + \gamma^{-(t+1)}\|A_{t+1}\|^2_{W_t^{-1}}) \geq \sum_{t=0}^{T-1} \log(1 + \gamma^{-t}\|A_{t+1}\|^2_{W_t^{-1}}) \\
&\geq \sum_{t=0}^{T-1} \log\left(1 + \frac{\gamma^{-t}\|A_{t+1}\|^2_{W_t^{-1}}}{\max\left(1, \frac{1}{\lambda}\right)}\right) .
\end{aligned}$$

Next, by using $W_t \geq \gamma^{-t}\lambda I_d$, we see that

$$\gamma^{-t}\|A_{t+1}\|^2_{W_t^{-1}} \leq \frac{1}{\lambda} .$$

Which ensures that

$$0 \leq \frac{\gamma^{-t}\|A_{t+1}\|^2_{W_t^{-1}}}{\max\left(1, \frac{1}{\lambda}\right)} \leq 1 .$$

Finally, using $\log(1+x) \geq x/2$ which is valid when $0 \leq x \leq 1$, we get:

$$\log\left(\frac{\det(W_T)}{\lambda^d}\right) \geq \frac{1}{2\max\left(1, \frac{1}{\lambda}\right)} \sum_{t=0}^{T-1} \gamma^{-t}\|A_{t+1}\|^2_{W_t^{-1}} .$$

□

The following lemma is a version of the Elliptical Lemma when a sliding window is used and can be extracted from [Russac et al., 2019, Proposition 9]. The proof is included here for the sake of completeness.

Lemma 5.31 (Elliptical potential with sliding window)**.** *Let $\{A_s\}_{s=1}^{\infty}$ a sequence in $\mathbb{R}^d$ such that $\|A_s\|_2 \leq 1$ for all $s \in \mathbb{N}$, and let λ be a non-negative scalar. For $t \geq 1$ define $V_t = \sum_{s=\max(1,t-\tau+1)}^{t} A_s A_s^\top + \lambda I_d$. The following inequality holds:*

$$\sum_{t=1}^{T} \|A_t\|^2_{V_{t-1}^{-1}} \leq 2d \max\left(1, \frac{1}{\lambda}\right) \lceil T/\tau \rceil \log\left(1 + \frac{\tau}{\lambda d}\right) .$$

Proof. We start by rewriting the sum as follows.

$$\sum_{t=1}^{T} \|A_t\|^2_{V_{t-1}^{-1}} = \sum_{k=0}^{\lceil T/\tau \rceil - 1} \sum_{t=k\tau+1}^{(k+1)\tau} \|A_t\|^2_{V_{t-1}^{-1}} .$$

For the k-th block of length τ we define the matrix $W_t^{(k)} = \sum_{s=k\tau+1}^{t} A_s A_s^\top + \lambda I_d$. We also have $\forall t \in [\![k\tau, (k+1)\tau]\!], V_t \geq W_t^{(k)}$ as every term in $W_t^{(k)}$ is contained in V_t and the extra-terms in V_t correspond to positive definite matrices.

$$\sum_{k=0}^{\lceil T/\tau \rceil - 1} \sum_{t=k\tau+1}^{(k+1)\tau} \|A_t\|^2_{V_{t-1}^{-1}} \leq \sum_{k=0}^{\lceil T/\tau \rceil - 1} \sum_{t=k\tau+1}^{(k+1)\tau} \|A_t\|^2_{(W_{t-1}^{(k)})^{-1}} .$$

Furthermore, $\forall t \in [\![k\tau + 1, (k+1)\tau]\!]$ we have,

$$\det(W_{t+1}^{(k)}) = \det(W_t^{(k)}) \left(1 + \|A_{t+1}\|^2_{(W_t^{(k)})^{-1}}\right) .$$

With positive definitive matrices whose determinants are strictly positive, this implies that

$$\frac{\det(W_{(k+1)\tau}^{(k)})}{\det(W_{k\tau}^{(k)})} = \prod_{t=k\tau}^{(k+1)\tau-1} \frac{\det(W_{t+1}^{(k)})}{\det(W_t^{(k)})} = \prod_{t=k\tau}^{(k+1)\tau-1} \left(1 + \|A_{t+1}\|^2_{(W_t^{(k)})^{-1}}\right) .$$

By definition we have $W_{k\tau}^{(k)} = \sum_{t=k\tau+1}^{k\tau} A_t A_t^\top + \lambda I_d = \lambda I_d$.

$$\begin{aligned}
\log\left(\frac{\det\left(W_{(k+1)\tau}^{(k)}\right)}{\lambda^d}\right) &= \sum_{t=k\tau}^{(k+1)\tau-1} \log\left(1 + \|A_{t+1}\|^2_{(W_t^{(k)})^{-1}}\right) \\
&\geq \sum_{t=k\tau}^{(k+1)\tau-1} \log\left(1 + \frac{1}{\max(1, 1/\lambda)} \|A_{t+1}\|^2_{(W_t^{(k)})^{-1}}\right) .
\end{aligned}$$

In the next step we use, $\forall 0 \leq x \leq 1, \log(1+x) \geq x/2$.

$$\log\left(\frac{\det\left(W_{(k+1)\tau}^{(k)}\right)}{\lambda^d}\right) \geq \frac{1}{2\max(1, 1/\lambda)} \sum_{t=k\tau}^{(k+1)\tau-1} \|A_{t+1}\|^2_{(W_t^{(k)})^{-1}} .$$

By summing, over the different blocks, we obtain

$$\sum_{k=0}^{\lceil T/\tau\rceil-1}\sum_{t=k\tau+1}^{(k+1)\tau}\|A_t\|^2_{V_{t-1}^{-1}} \leq \sum_{k=0}^{\lceil T/\tau\rceil-1}\sum_{t=k\tau+1}^{(k+1)\tau}\|A_t\|^2_{(W_{t-1}^{(k)})^{-1}} \leq 2\max(1,1/\lambda)\sum_{k=0}^{\lceil T/\tau\rceil-1}\log\left(\frac{\det\left(W_{(k+1)\tau}^{(k)}\right)}{\lambda^d}\right) .$$

Then, we upper bound $\det(W_{(k+1)\tau}^{(k)})$ using similar arguments than for Corollary 5.29,

$$\det(W_{(k+1)\tau}^{(k)}) \leq \left(\lambda+\frac{\tau}{d}\right)^d .$$

Applying the logarithm function on both sides concludes the proof. □

5.D.4 Link Between Δ_t and the Instantaneous Regret

For any optimistic algorithm, even in a non-stationary environment the instantaneous regret can be directly related to $\Delta_t(a,\theta)$ defined as

$$\Delta_t(a,\theta) = |\mu(a^\top\theta) - \mu(a^\top\theta_t^\star)| .$$

Proposition 5.32 (Based on Lemma 14 in [Faury et al., 2020]). *Consider any optimistic algorithm in a possibly non-stationary environment such that the exploration bonus for action a at time t is defined by $\beta_t(a)$. Let θ_t be the estimator used at time t by the algorithm to compute the UCB, i.e. $UCB_t(a) = \mu(a^\top\theta_t)+\beta_t(a)$. Under the assumption $\Delta_t(a,\theta_t)\leq\beta_t(a)$, the following inequality holds*

$$r_t \leq 2\beta_t(a_t) .$$

Proof. Let $A_{t,\star} = \text{argmax}_{a\in\mathcal{A}_t}\mu(a^\top\theta_t^\star)$

$$\begin{aligned} r_t &= \mu(A_{t,\star}^\top\theta_t^\star) - \mu(A_t^\top\theta_t^\star) \leq |\mu(A_{t,\star}^\top\theta_t^\star) - \mu(A_{t,\star}^\top\theta_t)| + \mu(A_{t,\star}^\top\theta_t) - \mu(A_t^\top\theta_t) + |\mu(A_t^\top\theta_t) - \mu(A_t^\top\theta_t^\star)| \\ &= \Delta_t(A_t,\theta_t) + \Delta_t(A_{t,\star},\theta_t) + \mu(A_{t,\star}^\top\theta_t) - \mu(A_t^\top\theta_t) \\ &= \Delta_t(A_t,\theta_t) + \Delta_t(A_{t,\star},\theta_t) + \mu(A_{t,\star}^\top\theta_t) + \beta_t(A_t^\star) - \mu(A_t^\top\theta_t) - \beta_t(A_t) + \beta_t(A_t) - \beta_t(A_t^\star) . \end{aligned}$$

For any optimistic algorithm with an exploration bonus of $\beta_t(.)$ and such that the upper confidence bound of the action a at time t is given by $\mu(a^\top\theta_t)+\beta_t(a)$, by definition for all $a\in\mathcal{A}_t$

$$\mu(a^\top\theta_t)+\beta_t(a) \leq \mu(A_t^\top\theta_t)+\beta_t(A_t) .$$

In particular, this is also true for the action $A_{t,\star}$. Therefore, plugging this inequality in the expression of the instantaneous regret gives

$$r_t \leq \Delta_t(A_t,\theta_t) + \Delta_t(A_{t,\star},\theta_t) + \beta(A_t) - \beta(A_t^\star) .$$

Under the additional assumption that $\Delta_t(a,\theta)\leq\beta_t(a)$, we obtain the announced result. □

This proposition shows that any improvement in an upper bound of $\Delta_t(a,\theta_t)$ will result in an improvement of the regret, as long as the exploration bonus satisfies the assumption stated in the proposition.

Appendix 5.E On the Worst Case Regret in the K-arm Setting

In this section, we build upon the analysis from [Garivier and Moulines, 2011] to provide a worst case regret bound for the sliding window policy in the K-arm setting. Even if a proper lower bound is missing, the results we provide here suggest that in some cases sliding window policies can suffer a regret of order $\mathcal{O}(\Gamma_T^{1/3}T^{2/3})$ in the simpler multi-armed bandit setting. In particular, this would mean that the $T^{2/3}$ dependency is not a sub-optimality from our setting but can already be seen for forgetting policies in the non-contextual setting. Worst-case regret bounds (i.e. gap-independent) for forgetting policies in non-stationary environments have seen little treatment in the literature.

Setting. The setting considered in this section is the one from [Garivier and Moulines, 2011]. At each time t, the player chooses an arm $I_t \in \{1, ..., K\}$ based on the previous rewards and actions. Upon selecting I_t a reward $X_t(I_t)$ is observed. We consider abruptly changing environments as in other sections, where the distribution of the rewards remains constant during phases and changes at unknown time instants. At time t, the arm i has a mean reward $\mu_t(i)$. As before, Γ_T denote the number of abrupt changes in the reward distributions before time T. Following the notation from [Trovo et al., 2020], we denote the Γ_T breakpoints $\mathcal{B} = \{b_1, ..., b_{\Gamma_T}\}$. We can associate Γ_T stationary phases $\{\phi_1, ..., \phi_{\Gamma_T}\}$ with these breakpoints, where $\phi_i = \{t \in \{1, ..., T\} \text{ s.t } b_{i-1} \leq t < b_i\}$ and $b_0 = 1$. It is further assumed that for all arms and all time instants the means of the reward distributions lie in $[0, B]$. In this section the focus is on the forgetting policy using a sliding window but the same arguments can be used with exponentially increasing weights.

Improving the problem-dependent bound. In [Garivier and Moulines, 2011, Theorem 2], the number of times the arm i is played before time T while being sub-optimal is upper bounded in expectation as

$$\mathbb{E}\left[N_T(i)\right] \leq \frac{C(\tau)}{(\Delta\mu_T(i))^2}\frac{T\log(\tau)}{\tau} + \tau\Gamma_T + \log^2(\tau)\ , \tag{5.46}$$

where

$$\Delta\mu_T(i) = \min\{\mu_t(i_t^\star) - \mu_t(i) : t \in \{1, ..., T\}, \mu_t(i) < \mu_t(i_t^\star)\}\ .$$

This result has a worst case flavor in the sense that $\Delta\mu_T(i)$ is the minimum distance between the mean of the optimal arm and the mean of the i-th arm when i is sub-optimal over the entire time horizon. We obtain a less pessimistic bound by decomposing the regret into the Γ_T different stationary phases and upper-bounding the number of times a sub-optimal arm is drawn in each of these phases ϕ. The upper-bound naturally depends on Δ_i^ϕ, the difference between the mean of the optimal arm and the i-th arm in the ϕ-th stationary phase rather than $\Delta\mu_T(i)$. This is of utmost importance as for some phases Δ_i^ϕ can be significantly larger than $\Delta\mu_T(i)$.

During the ϕ-th stationary phase, let μ_i^ϕ denote the mean of the i-th arm and N_i^ϕ denote the number of times the arm i is selected. The regret can be decomposed as follows:

$$\mathbb{E}\left[R(T)\right] = \sum_{t=1}^{T}(\mu_t^\star - \mu_t(i_t)) = \sum_{i=1}^{K}\sum_{\phi=1}^{\Gamma_T}\Delta_i^\phi\mathbb{E}[N_i^\phi]\ . \tag{5.47}$$

A worst-case bound. The bound from Equation (5.46) is problem-dependent and depends explicitly on the minimum gap. It is interesting to study the worst case regret. In particular when $\Delta\mu_T(i)$ goes to 0 the upper bound from Equation (5.46) becomes uninformative. At the same time, with a small gap Δ_i^ϕ the cost of selecting the i-th arm rather than the optimal one diminishes. The trade-off between these two opposite effects is made explicit in the following result.

Theorem 5.33. *The following upper bound holds for the worst case regret of the sliding window policy from [Garivier and Moulines, 2011]*

$$\mathbb{E}[R(T)] \leq C_1\sqrt{K}\frac{T}{\sqrt{\tau}} + C_2\sqrt{K}\tau\Gamma_T + C_3K\frac{T}{\tau} ,$$

with C_1, C_2 and C_3 universal constants that depends only on the logarithm of τ. In particular, setting $\tau = \frac{T^{2/3}}{K^{1/3}\Gamma_T^{2/3}}$ yields:

$$\mathbb{E}[R_T] = \widetilde{\mathcal{O}}(K^{2/3}\Gamma_T^{1/3}T^{2/3}) .$$

Proof.

$$\mathbb{E}[R(T)] = \sum_{i=1}^{K}\sum_{\phi=1}^{\Gamma_T}\Delta_i^\phi\mathbb{E}[N_i^\phi] = \sum_{i,\phi:\Delta_i^\phi>\Delta}\Delta_i^\phi\mathbb{E}[N_i^\phi] + \sum_{i,\phi:\Delta_i^\phi\leq\Delta}\Delta_i^\phi\mathbb{E}[N_i^\phi]$$
$$\leq \sum_{i,\phi:\Delta_i^\phi>\Delta}\Delta_i^\phi\mathbb{E}[N_i^\phi] + \Delta\sum_{i=1}^{K}\sum_{\phi=1}^{\Gamma_T}\mathbb{E}[N_i^\phi] \leq \sum_{i,\phi:\Delta_i^\phi>\Delta}\Delta_i^\phi\mathbb{E}[N_i^\phi] + \Delta T .$$

The next step consists in upper bounding the expected number of times the arm i is selected in the ϕ-th phase. We recall that N_i^ϕ is defined as

$$N_i^\phi = \sum_{t\in\phi}\mathbb{1}(I_t = i \neq i_t^\star) = \sum_{t=b_{\phi-1}}^{b_\phi}\mathbb{1}(I_t = i \neq i_t^\star) .$$

We introduce $N_t(\tau, i) = \sum_{s=t-\tau+1}^{t}\mathbb{1}(I_s = i)$, the number of times the arm i was selected in the τ steps preceding t. We have the following:

$$N_i^\phi = \sum_{t=b_{\phi-1}}^{b_{\phi-1}+\tau-1}\mathbb{1}(I_t = i \neq i_t^\star) + \sum_{t=b_{\phi-1}+\tau}^{b_\phi}\mathbb{1}(I_t = i \neq i_t^\star) \leq \tau + \sum_{t=b_{\phi-1}+\tau}^{b_\phi}\mathbb{1}(I_t = i \neq i_t^\star)$$
$$\leq \tau + \sum_{t=b_{\phi-1}+\tau}^{b_\phi}\mathbb{1}(I_t = i \neq i_t^\star, N_t(\tau,i) \leq A_i^\phi) + \sum_{t=b_{\phi-1}+\tau}^{b_\phi}\mathbb{1}(I_t = i \neq i_t^\star, N_t(\tau,i) > A_i^\phi) .$$

The first term can be bounded using [Garivier and Moulines, 2011, Lemma 1] that is restated here.

Lemma 5.34 (Lemma 1 in [Garivier and Moulines, 2011])**.** *Let $i \in \{1, ..., K\}$. For any positive integer τ and any positive m,*

$$\sum_{t=K+1}^{T}\mathbb{1}(I_t = i, N_t(\tau,i) \leq m) \leq \lceil T/\tau\rceil m .$$

Lemma 5.34 can be adapted to our setting and by introducing T^ϕ the length of the ϕ-th stationary phase, one has:

$$\sum_{t=b_{\phi-1}+\tau}^{b_\phi} \mathbb{1}(I_t = i \neq i_t^\star, N_t(\tau, i) \leq A_i^\phi) \leq \lceil T^\phi/\tau \rceil A_i^\phi \ .$$

This in turn gives,

$$N_i^\phi \leq \tau + \lceil T^\phi/\tau \rceil A_i^\phi + \sum_{t=b_{\phi-1}+\tau}^{b_\phi} \mathbb{1}(I_t = i \neq i_t^\star, N_t(\tau, i) > A_i^\phi) \ .$$

We recall that the upper confidence bound for the sliding-window strategy has the following form in the K arm setting [Garivier and Moulines, 2011]:

$$UCB_i(t) = \bar{X}_t(\tau, i) + c_t(\tau, i) \ ,$$

with

$$\bar{X}_t(\tau, i) = \frac{1}{N_t(\tau, i)} \sum_{s=t-\tau+1}^{t} X_s(i)\mathbb{1}(I_s = i) \quad \text{and} \quad c_t(\tau, i) = B\sqrt{\frac{\xi \log(\min(t, \tau))}{N_t(\tau, i)}} \ .$$

Following the same arguments than [Garivier and Moulines, 2011] when the event $\{I_t = i \neq i_t^\star, N_t(\tau, i) > A_i^\phi\}$ holds, at least one of the three following events E_1, E_2, E_3 must be true where:

$$E_1 = \{\bar{X}_t(\tau, i) > \mu_t(i) + c_t(\tau, i)\} \quad \text{the case where } \mu_t(i) \text{ is over-estimated.}$$

$E_2 = \{\bar{X}_t(\tau, i_t^\star) < \mu_t^\star - c_t(\tau, i_t^\star)\}$ the case where the best arm at time t is under-estimated.

$E_3 = \{\mu_t^\star - \mu_t(i) \leq 2c_t(\tau, i), N_t(\tau, i) > A_i^\phi\}$ the case where the means are too close to each others.

From now on, we set

$$A_i^\phi = \frac{4B^2 \xi \log(\tau)}{(\Delta_i^\phi)^2} \ .$$

In doing so, on the event E_3 the following holds:

$$c_t(\tau, i) = B\sqrt{\frac{\xi \log(\min(t, \tau))}{N_t(\tau, i)}} < B\sqrt{\frac{\xi \log(\min(t, \tau))}{A_i^\phi}} < \frac{\Delta_i^\phi}{2}\sqrt{\frac{\log(\min(t, \tau))}{\log(\tau)}} < \frac{\Delta_i^\phi}{2} \ .$$

Therefore, this choice of A_i^ϕ ensures that the event E_3 never occurs. Bounding the probability of the events E_1 and E_2 can be done with the concentration inequality established in [Garivier and Moulines, 2011]. For any $\eta > 0$, by selecting a specific value of ξ one can obtain,

$$\mathbb{P}(E_1) \leq \frac{\left\lceil \frac{\log(\min(t,\tau))}{\log(1+\eta)} \right\rceil}{\min(t, \tau)} \quad \text{and} \quad \mathbb{P}(E_2) \leq \frac{\left\lceil \frac{\log(\min(t,\tau))}{\log(1+\eta)} \right\rceil}{\min(t, \tau)} \ .$$

Consequently we have,

$$\mathbb{E}[N_i^\phi] \leq \tau + \lceil T^\phi/\tau \rceil \frac{4B^2 \xi \log(\tau)}{(\Delta_i^\phi)^2} + 2 \sum_{t=b_{\phi-1}+\tau}^{b_\phi} \frac{\left\lceil \frac{\log(\min(t,\tau))}{\log(1+\eta)} \right\rceil}{\min(t, \tau)} \ .$$

Plugging this in the regret's upper bound gives:

$$\begin{aligned}\mathbb{E}[R(T)] &\leq \sum_{i,\phi:\Delta_i^\phi>\Delta} \Delta_i^\phi \left(\tau + \lceil T^\phi/\tau \rceil \frac{4B^2\xi\log(\tau)}{(\Delta_i^\phi)^2} + 2\sum_{t=b_{\phi-1}+\tau}^{b_\phi} \frac{\left\lceil \frac{\log(\min(t,\tau))}{\log(1+\eta)} \right\rceil}{\min(t,\tau)} \right) + \Delta T \\ &\leq \sum_{i,\phi:\Delta_i^\phi>\Delta} \frac{4B^2\xi\log(\tau)}{\Delta_i^\phi} \lceil T^\phi/\tau \rceil + \sum_{i,\phi:\Delta_i^\phi>\Delta} \Delta_i^\phi \left(\tau + 2\sum_{t=b_{\phi-1}+\tau}^{b_\phi} \frac{\left\lceil \frac{\log(\min(t,\tau))}{\log(1+\eta)} \right\rceil}{\min(t,\tau)} \right) + \Delta T \\ &\leq \frac{4B^2\xi\log(\tau)K}{\Delta}\frac{T}{\tau} + \tau K\Gamma_T B + 2KB\sum_{\phi=1}^{\Gamma_T}\sum_{t=b_{\phi-1}+\tau}^{b_\phi} \frac{\left\lceil \frac{\log(\min(t,\tau))}{\log(1+\eta)} \right\rceil}{\min(t,\tau)} + \Delta T\,.\end{aligned}$$

In the last inequality we have used $\Delta_i^\phi \leq B$ coming from $\mu_i(t) \in [0, B]$ for all i and all $t \leq T$. Furthermore,

$$\sum_{\phi=1}^{\Gamma_T}\sum_{t=b_{\phi-1}+\tau}^{b_\phi} \frac{\left\lceil \frac{\log(\min(t,\tau))}{\log(1+\eta)} \right\rceil}{\min(t,\tau)} \leq \sum_{t=\tau}^{T} \frac{\frac{\log(\min(t,\tau))}{\log(1+\eta)}+1}{\min(t,\tau)} = \frac{T}{\tau}\left(\frac{\log(\tau)}{\log(1+\eta)}+1\right)\,.$$

Hence,

$$\mathbb{E}[R(T)] \leq \frac{4B^2\xi\log(\tau)K}{\Delta}\frac{T}{\tau} + \Delta T + \tau K\Gamma_T B + 2KB\left(\frac{\log(\tau)}{\log(1+\eta)}+1\right)\frac{T}{\tau}\,.$$

By differentiating with respect to Δ, the right hand side is maximized when setting $\Delta = 2B\sqrt{\frac{\xi\log(\tau)K}{\tau}}$. With this value of Δ,

$$\mathbb{E}[R(T)] \leq 4B\sqrt{\xi\log(\tau)}\sqrt{K}\frac{T}{\sqrt{\tau}} + BK\tau\Gamma_T + 2BK\log(\tau)\frac{T}{\tau}\,.$$

Now by selecting $\tau = \frac{T^{2/3}}{K^{1/3}\Gamma_T^{2/3}}$, we obtain the announced scaling. □

Remark 5.7. *The term $T/\sqrt{\tau}$ that can be seen in the worst case bound proposed in Theorem 5.33 also appears in the gap independent bound of* SC-SW-GLUCB *(Theorem 5.20). When focusing on gap dependent bounds, there is also a strong similarity. In the K-arm setting, Equation (5.46) has a T/τ dependency. This term can also be seen in the GLB setting in Theorem 5.22 using an analogous assumption on the gap. This analogy explains why the upper-bounds have the same scaling in the K-arm and in the GLB setting. Going from $T/\sqrt{\tau}$ to T/τ when adding the assumption on the gaps is the key step allowing a scaling of the regret of order $\widetilde{\mathcal{O}}(\sqrt{T\Gamma_T})$.*

6 | Generalized Linear Bandits under Parameter Drift

In this chapter, we still consider GLBs in non-stationary environments, but non-stationarity is now characterized by the general metric known as the variation-budget or *parameter-drift*, denoted B_T. While previous attempts have been made to extend linear bandit algorithms to this setting, they overlook a salient feature of GLBs which flaws their results. In this work, we introduce a new algorithm that addresses this difficulty. We prove that under a geometric assumption on the action sets, our approach enjoys a $\widetilde{\mathcal{O}}(B_T^{1/3}T^{2/3})$ regret bound. In the general case, we show that it suffers at most a $\widetilde{\mathcal{O}}(B_T^{1/5}T^{4/5})$ regret. At the core of our contribution is a generalization of the projection step introduced in [Filippi et al., 2010], adapted to the non-stationary nature of the problem. Our analysis sheds light on central mechanisms inherited from the setting by explicitly splitting the treatment of the learning and tracking aspects of the problem. The results from this chapter are based on [Faury et al., 2021b] and [Faury et al., 2021a].

Outline

6.1 Introduction

Given that this chapter shares strong similarity with Chapter 5, we refer the reader to Chapter 5 for a more detailed introduction and a discussion on existing works. We will focus here on the works that consider exactly generalized linear bandits under parameter drift. At first glance, as the analysis of generalized linear bandits mainly relies on tools from the linear bandit literature, one could expect this demonstration to be straight-forward. As a matter of fact, the treatment of GLBs in non-stationary environments was already proposed as a direct extension of non-stationary linear bandit algorithms ([Cheung et al., 2021, Section 8.3] and [Zhao et al., 2020, Section 5.2]). However, as recently pointed out by [Russac et al., 2020], some crucial subtleties of the generalized linear bandits flaw the analysis and negates the validity of such extensions. An answer to this issue was brought by [Russac et al., 2020] and the analysis we proposed in Chapter 5, where a valid analysis for generalized linear bandits in non-stationary environments is obtained. However, those analysis are restricted to a specific kind of non-stationarity known as *abrupt changes*, leaving the treatment of the superior *parameter-drift* case for future work. To the best of our knowledge, a correct derivation of generalized linear bandits' behavior under this more general description of non-stationarity is still missing.

Scope and contributions. We focus in this chapter on closing this gap. Our main contribution is **(1)** the design of BVD-GLM-UCB (Algorithm 13), the first generalized linear bandit algorithm resilient to parameter-drift and matching the known minimax rates - though only for some action sets (Theorem 6.1). For more general configurations, we still provide a sub-linear regret bound, slightly lagging behind the known rates for non-stationary LBs. Our result relies on **(2)** a generalization of the projection step of [Filippi et al., 2010] to non-stationary environments, of similar complexity than its stationary counterpart (Proposition 6.2). Our analysis **(3)** sheds light on some salient mechanisms of non-stationary bandits.

6.2 Preliminaries

We consider in this work the stochastic contextual bandit setting under parameter-drift. The environment starts by picking a sequence of parameters $\{\theta_t^\star\}_{t=1}^\infty$. A repeated game then begins between the environment and an agent. At each round t, the environment presents the agent with a set of actions $\mathcal{A}_t$ (potentially contextual, large or even infinite). The agent selects an action $A_t \in \mathcal{A}_t$ and receives a (stochastic) reward X_t. The reward model we consider is the same than in Chapter 5 and with $\mathcal{F}_t = \sigma(\mathcal{A}_1, A_1, X_1, \dots, \mathcal{A}_t, A_t, X_t, A_{t+1}, \mathcal{A}_{t+1})$ the σ-field from the previous chapter, we assume

$$\mathbb{E}\left[X_t \,|\, \mathcal{F}_{t-1}\right] = \mu\left(A_t^\top \theta_t^\star\right). \tag{6.1}$$

μ is a strictly increasing, continuously differentiable real-valued function most often referred to as the inverse link function. Notable instances of such a problem include the logistic bandit and the Poisson bandit. The goal of the agent is to minimize the cumulative pseudo-regret:

$$R(T) \coloneqq \sum_{t=1}^T \mu(A_{t,\star}^\top \theta_t^\star) - \sum_{t=1}^T \mu(A_t^\top \theta_t^\star) \;\text{ where } A_{t,\star} = \operatorname{argmax}_{a\in\mathcal{A}_t} \mu(a^\top \theta_t^\star)\,.$$

We recall the assumptions that are made for this chapter.

Assumption 6.1 (Bounded decision set)**.** *For all $t \geq 1$, the following holds true: $\|\theta_t^\star\|_2 \leq S$. Further, the actions have bounded norms: $\|a\|_2 \leq L$ for all $a \in \mathcal{A}_t$.*

Assumption 6.2 (Bounded reward)**.** *There exists $m > 0$ s.t $0 \leq X_t \leq m$ holds almost surely.*

We will denote $\Theta = \{\theta, \|\theta\|_2 \leq S\}$ the set of admissible parameters and $\mathcal{A} = \{a, \|a\|_2 \leq L\}$. We assume that the quantities L, S and m are known to the agent. The true parameters $\{\theta_t^\star\}_{t=1}^\infty$ are unknown, and their drift is quantified by the variation *variation-budget*, which characterizes the magnitude of the non-stationarity in the environment:

$$\mathcal{B}_T \coloneqq \sum_{t=1}^{T-1} \|\theta_{t+1}^\star - \theta_t^\star\|_2 .$$

Naturally $\mathcal{B}_T$ is unknown. For the sake of simplicity and to isolate the main contribution of this chapter (*i.e* minimax-optimality in non-stationary GLBs), we will make the following assumption.

Assumption 6.3 (Variation-budget upper-bound)**.** *B_T is a* known *quantity such that $B_T \geq \mathcal{B}_T$.*

This assumption is common in non-stationary bandits [Besbes et al., 2014, Cheung et al., 2021, Zhao et al., 2020]. We will show in Section 6.4.4 how to bypass it with little to no impact on the regret. For a given inverse link function μ, we will follow the notation from [Filippi et al., 2010] and denote:

$$k_\mu \coloneqq \sup_{a\in\mathcal{A},\theta\in\Theta} \dot{\mu}(a^\top\theta), \qquad c_\mu \coloneqq \inf_{a\in\mathcal{A},\theta\in\Theta} \dot{\mu}(a^\top\theta), \qquad R_\mu \coloneqq \frac{k_\mu}{c_\mu} .$$

As in the stationary setting, learning can be performed through the *quasi-maximum likelihood* principle, albeit with adequate modifications. Let b be a primitive of μ. Thanks to the strict increasing nature of the latter, b is a strictly convex function. Let $\lambda > 0$ and for $\gamma \in (0, 1)$ define[1]:

$$\hat{\theta}_t = \text{argmin}_{\theta\in\mathbb{R}^d} \sum_{s=1}^{t} \gamma^{t-s} \left[b(A_s^\top\theta) - X_s A_s^\top\theta\right] + \frac{\lambda c_\mu}{2} \|\theta\|_2^2 , \tag{6.2}$$

which is well-defined and unique as the minimizer of a strictly convex and coercive function. Further:

$$g_t(\theta) \coloneqq \sum_{s=1}^{t} \gamma^{t-s} \mu(A_s^\top\theta) A_s + \lambda c_\mu \theta .$$

Finally, we will use

$$V_t \coloneqq \sum_{s=1}^{t} \gamma^{t-s} A_s A_s^\top + \lambda I_d \quad \text{and} \quad \widetilde{V}_t \coloneqq \sum_{s=1}^{t} \gamma^{2(t-s)} A_s A_s^\top + \lambda I_d .$$

Some of our results require the following assumption on the arm-sets $\mathcal{A}_t$. We will discuss the reasons behind this hypothesis, as well as its main implications in the following section.

Assumption 6.4 (Orthogonal arm-set)**.** *Let $\{e_i\}_{i=1}^d$ an orthonormal basis of $\mathbb{R}^d$. We call a collection of arm-sets $\{\mathcal{A}_t\}_t$* orthogonal *if for all $t \geq 1$ and any $a \in \mathcal{A}_t$, there exists α and i such that $a = \alpha e_i$.*

[1] We follow Chapter 4 and use an exponential moving-average strategy. Our contribution is not specific to this approach and can easily be extended to other alternatives, e.g the sliding window.

6.3 Related Work: Limitations and Challenges

6.3.1 Generalized Linear Bandits

Generalized linear bandits were first introduced by [Filippi et al., 2010] who studied optimistic algorithms which enjoy a $\tilde{\mathcal{O}}(R_\mu d\sqrt{T})$ regret upper-bound. This bound was later refined for K-arms problem to $\tilde{\mathcal{O}}(R_\mu\sqrt{d\log(K)T})$ [Li et al., 2017]. These findings were extended to randomized algorithms, both in the frequentist [Abeille and Lazaric, 2017] and Bayesian setting [Russo and Van Roy, 2014, Dong and Van Roy, 2018]. Generalized linear bandits also received an increasing attention targeted at improving their practical implementations [Jun et al., 2017, Dumitrascu et al., 2018]. In this chapter, we focus on generalized linear bandits in smoothly drifting environments.

6.3.2 Toward Non-Stationary GLBs: Limitations

On the limits of piece-wise stationarity. To the best of our knowledge, the first valid analysis of non-stationary GLBs was conducted by [Russac et al., 2020, Russac et al., 2021a]. However, their work is restricted to piece-wise stationary environments, characterized by the number Γ_T of switches of the reward signal. On the practical side, this drastically narrows down the non-stationary scenarios that can be efficiently addressed, as the measure Γ_T can grossly overestimate the importance of the non-stationarity. In such case, any algorithm based on this measure will be sub-optimal and discard too fast previous data, quickly judged uninformative since the level of non-stationarity is expected to be high. This is typically the case in environments with many switches of small amplitude, characteristic of smooth drifts (e.g user-fatigue in recommender systems). On the theoretical side, this approach tells us little about the difficulties and challenges brought by the non-stationarity, as it relies on the fact that far enough from a switch, the environment is stationary. On the contrary, the variation-budget metric B_T introduced and discussed in [Besbes et al., 2014, Section 2], allows for much finer considerations. It stands as a powerful characterization of the non-stationarity, measuring the number of switches and their amplitude *jointly*. As a result, it can efficiently cover different scenarios, from drifting to piece-wise stationary environments. An adequate treatment of generalized linear bandits under this superior metric is therefore a crucial missing piece, and requires a sensibly different analysis and an appropriate algorithmic design.

Parameter-drift and GLBs: flaws of previous approaches. Most of the existing non-stationary linear bandit algorithms address the parameter-drift setting, and their extension to generalized linear bandits was at first considered as relatively straight-forward [Cheung et al., 2021, Zhao et al., 2020]. Unfortunately, existing analyses suffer from important caveats because they overlook a crucial feature of generalized linear bandits. Following [Filippi et al., 2010], they rely on a linearization of the reward function around $\hat{\theta}_t$. Naturally, the linear approximation must accurately describe the *effective* behavior of the reward signal (characterized by the ground-truth $\theta_t^\star$). From Assumption 6.2, this translates in the constraint $\hat{\theta}_t \in \Theta$, which is implicitly assumed to hold in previous attempts. Unfortunately, there exists no proof guaranteeing that $\hat{\theta}_t \in \Theta$ could hold. Even worse, existing deviation bounds [Abbasi-Yadkori et al., 2011, Theorem 1] rather suggest that in some directions, *even in the stationary case*, $\hat{\theta}_t$ can grow to be $\sqrt{\log(t)}$ far from Θ! The situation is even worse under non-stationarity since, as we shall see, $\hat{\theta}_t$ can be B_t far from Θ. This flaw in the analysis is critical and cannot be easily fixed without severely

degrading the regret guarantee. When $\hat{\theta}_t \notin \Theta$, this impacts the ratio R_μ which captures the degree of non-linearity of the inverse link function. For the highly non-linear logistic function, easy computations show that $R_\mu \geq e^{SL}$. If we were to inflate the radius of the admissible set Θ from S to $S + \delta_S$ (so that it contains $\hat{\theta}_t$), the estimated non-linearity of the reward function would be even stronger and R_μ would be multiplied by a factor $e^{L\delta_S}$! Because the regret bound scales linearly with R_μ, this exponential growth would lead to prohibitively deficient performance guarantees.

Remark 6.1. *The fact that $\hat{\theta}_t$ can leave the admissible set Θ is not merely a theoretical construction inherited from potentially loose deviation bounds. As highlighted in Figure 6.2b, we can see in our numerical simulations that this often happens in practice when the environment is non-stationary.*

6.3.3 Non-stationary GLBs: Challenges

In their seminal work, [Filippi et al., 2010] countered the aforementioned difficulty by introducing a *projection* step, mapping $\hat{\theta}_t$ back to an admissible parameter $\tilde{\theta}_t \in \Theta$. Formally, they compute:

$$\tilde{\theta}_t = \text{argmin}_{\theta \in \Theta} \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{V_t^{-1}} \tag{P0}$$

and use $\tilde{\theta}_t$ to predict the performance of the available actions. The projection step Equation (**P0**) essentially incorporates the prior knowledge $\theta_\star \in \Theta$ (Assumption 6.2) without degrading the learning guarantees of the maximum likelihood estimator. This strategy was also leveraged by [Russac et al., 2020], which was made possible thanks to their piece-wise stationarity assumption.

The situation is different in our setting, as the parameter-drift framework allows the sequence $\{\theta_t^\star\}$ to change *at every round*. This introduces **(1)** the need to characterize two phenomenons of different nature that we will designate as *learning* and *tracking*. The former (learning) is linked to the deviation of the maximum-likelihood estimator $\hat{\theta}_t$ from its noiseless counterpart $\bar{\theta}_t$ (the estimator that one would have obtained if one could have averaged an infinite number of realization of the trajectory). The later (tracking) measures the deviation of $\bar{\theta}_t$ from the current $\theta_{t+1}^\star$, due to an incompressible error inherited from the drifting nature of the sequence $\{\theta_s^\star\}_{s=1}^t$. The learning and tracking mechanisms are both sources of deviation of $\hat{\theta}_t$ away from Θ, each under a different metric. This leads to **(2)** a tension in the design of the projection as this requires to incorporate the knowledge $\{\theta_t^\star\} \in \Theta$, without degrading neither the learning nor the tracking guarantees. This rules out the projection step Equation (**P0**), oblivious to the tracking aspect of the problem and which needs to be generalized to adapt to the two sources of deviation (i.e learning and tracking).

6.4 Algorithm and Regret Bound

6.4.1 Algorithm

This section is dedicated to the description of the design of our new algorithm BVD-GLM-UCB. It operates in two steps: **(Step 1)** the computation of an appropriate admissible parameter $\tilde{\theta}_t \in \Theta$ (to be used for predicting the rewards associated with the actions $a \in \mathcal{A}_t$ available

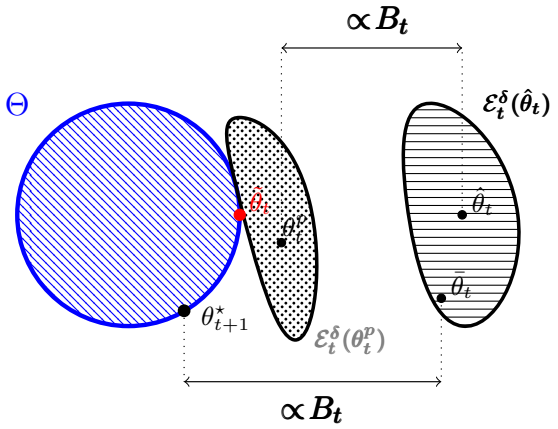


Figure 6.1: Illustration of the different parameters of interest. As stated by Lemma 6.5 and Lemma 6.7, the deviations $(\theta_t^p \leftrightarrow \hat{\theta}_t)$ and $(\bar{\theta}_t \leftrightarrow \theta_{t+1}^\star)$ are linked to the parameter-drift B_t. On the other hand, the deviations $(\hat{\theta}_t \leftrightarrow \bar{\theta}_t)$ and $(\tilde{\theta}_t \leftrightarrow \theta_t^p)$ are characterized by the stochastic nature of the problem.

at round t) and **(Step 2)** the construction of a suitable exploration bonus to compensate for prediction errors.

The first step builds on the following set, linked to the deviation incurred through the learning process:

$$\mathcal{E}_t^\delta(\theta) := \left\{ \theta' \in \mathbb{R}^d \text{ s.t } \left\| g_t(\theta') - g_t(\theta) \right\|_{\widetilde{V}_t^{-1}} \leq \beta_t(\delta) \right\}, \tag{6.3}$$

where $\beta_t(\delta)$ is a slowly-increasing function of time (to be defined later) and $\delta \in (0, 1]$.

Step 1. We start by identifying an intermediary parameter θ_t^p, solution of the following constrained optimization program (ties can be broken arbitrarily):

$$\theta_t^p \in \operatorname{argmin}_{\theta \in \mathbb{R}^d} \left\{ \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{V_t^{-2}} \text{ s.t } \Theta \cap \mathcal{E}_t^\delta(\theta) \neq \emptyset \right\}. \tag{P1}$$

The optimization program Equation (**P1**) is well-posed as it consists in minimizing a smooth function over a non-empty compact set[2]. Once θ_t^p is computed, the algorithm simply chooses any parameter $\tilde{\theta}_t \in \Theta \cap \mathcal{E}_t^\delta(\theta_t^p)$. An efficient procedure to find such a parameter is detailed in Section 6.4.3. The different parameters of interest for BVD-GLM-UCB are illustrated in Figure 6.1.

Remark 6.2. *Notice the difference with the projection step used in the stationary case. In our case it is possible that $\mathcal{E}_t^\delta(\hat{\theta}_t)$ (which is the confidence set centered at $\hat{\theta}_t$) does not intersect the admissible set Θ. Our strategy for finding $\tilde{\theta}_t$ is then to compute an appropriate **vibration** $\mathcal{E}_t^\delta(\theta_t^p)$ of $\mathcal{E}_t^\delta(\hat{\theta}_t)$ which does intersect Θ, while minimizing the deviation between θ_t^p and $\hat{\theta}_t$ according to a metric related to the tracking error (through the map g_t and the squared inverse of the design matrix).*

[2]Notice that $\{\theta \text{ s.t } \Theta \cap \mathcal{E}_t^\delta(\theta) \neq \emptyset\}$ always contains 0_d, while the compactness is inherited from Θ.

Step 2. The exploration bonus at round t for a given arm $a \in \mathcal{A}_t$ is defined as $b_{t-1}(a) \coloneqq 2R_\mu \beta_{t-1}(\delta)\|a\|_{V_{t-1}^{-1}}$, where $\delta \in (0,1]$ and:

$$\beta_t(\delta) \coloneqq \sqrt{\lambda} c_\mu S + \frac{m}{2}\sqrt{2\log(1/\delta) + d\log\left(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)}\right)}\,.$$

BVD-GLM-UCB then follows an optimistic strategy, boosting the predicted reward associated with $\tilde{\theta}_{t-1}$ by b_{t-1} and plays $A_t \in \operatorname{argmax}_{a\in\mathcal{A}_t} \mu(a^\top \tilde{\theta}_{t-1}) + b_{t-1}(a)$. The pseudo-code is summarized in Algorithm 13.

Input: Regularization λ, confidence δ, inverse link function μ, discount factor γ, constants S, L and m.
Initialization: Compute R_μ, let $V_0 \leftarrow \lambda I_d$ and $\hat{\theta}_0 \leftarrow 0_d$.
for $t \geq 1$ **do**
  Find θ_{t-1}^p by solving Equation (**P1**) and select $\tilde{\theta}_{t-1} \in \Theta \cap \mathcal{E}_{t-1}^\delta(\theta_{t-1}^p)$.
  Play $A_t \leftarrow \operatorname{argmax}_{a\in\mathcal{A}_t} \mu(a^\top \tilde{\theta}_{t-1}) + 2R_\mu \beta_{t-1}(\delta)\|a\|_{V_{t-1}^{-1}}$.
  Observe reward X_t, update $\hat{\theta}_t$ by solving Equation (6.2).
  Update design matrix: $V_t \leftarrow \gamma V_{t-1} + A_t A_t^\top + (1-\gamma)\lambda I_d$.

Algorithm 13: BVD-GLM-UCB

6.4.2 Regret bound

We provide in Theorem 6.1 a high-probability bound on the regret of BVD-GLM-UCB.

Theorem 6.1. *Under Assumptions 6.1-6.2-6.3 and 6.4, setting* $\gamma = 1 - (B_T/(dT))^{2/3}$ *ensures that the regret of BVD-GLM-UCB satisfies:*

$$R(T) = \tilde{\mathcal{O}}\left(R_\mu d^{2/3} B_T^{1/3} T^{2/3}\right) \qquad w.h.p$$

Under general arm-set geometry and Assumptions 6.1-6.2-6.3, setting $\gamma = 1-(B_T/(\sqrt{d}T))^{2/5}$ *ensures that the regret of BVD-GLM-UCB satisfies:*

$$R(T) = \tilde{\mathcal{O}}\left(R_\mu d^{9/10} B_T^{1/5} T^{4/5}\right) \qquad w.h.p$$

A few comments are in order. First, we note that as in the linear case, under Assumption 6.4 the upper-bound on $R(T)$ matches the asymptotic rates of the linear bandit lower-bound under parameter drift [Cheung et al., 2021, Theorem 1]. Without this assumption, the upper-bound suffers a small lag behind the linear bandit rates, from $T^{3/4}$ to $T^{4/5}$ (Corollary 4.4). Second, one can notice the presence in the bound of the ratio R_μ, typical of the linearization approach performed to analyze generalized linear bandits. The bounds presented in Theorem 6.1 are therefore quite natural and extends the work of [Filippi et al., 2010] to non-stationary worlds. We emphasize that if the result seems unsurprising, it required a substantially different machinery, both for the design of the algorithm and its analysis. We highlight this last point in Section 6.5, dedicated at providing a comprehensive sketch of proof for Theorem 6.1. The complete and detailed proof is deferred to Section 6.B in the supplementary material.

6.4.3 Solving the Projection Step

The optimization program Equation (**P1**) and the subsequent search of a valid parameter $\tilde{\theta}_t$ can raise some legitimate concerns regarding the ease of practical implementation. Indeed, the feasible set of Equation (**P1**) is given by $\{\theta \text{ s.t. } \Theta \cap \mathcal{E}_t^\delta(\theta) \neq \emptyset\}$, where $\mathcal{E}_t^\delta(\theta)$ is defined in Equation (6.3). Hence, the associated constraint is *implicit* as it involves an additional *non-convex* minimization program. As a result, it makes the constraint uneasy to manipulate and even hard to check. The same difficulty arises when searching for $\tilde{\theta}_t \in \Theta \cap \mathcal{E}_t^\delta(\theta_t^p)$ where θ_t^p is a solution of Equation (**P1**), due to the non-convexity of the set $\mathcal{E}_t^\delta(\theta_t^p)$. The following proposition provides an alternative that avoids those difficulties.

Proposition 6.2. *Let $\tilde{\theta}_t$ be such that:*

$$\begin{pmatrix} \tilde{\theta}_t \\ \eta_t^p \end{pmatrix} \in \operatorname{argmin}_{\theta' \in \mathbb{R}^d, \eta \in \mathbb{R}^d} \left\{ \left\| g_t(\theta') + \beta_t(\delta) \widetilde{V}_t^{1/2} \eta - g_t(\hat{\theta}_t) \right\|_{V_t^{-2}} \; s.t \; \|\theta'\|_2 \leq S, \|\eta\|_2 \leq 1 \right\}. \tag{P2}$$

*It exists θ_t^p solution of Equation (**P1**) such that $\tilde{\theta}_t \in \Theta \cap \mathcal{E}_t^\delta(\theta_t^p)$.*

Proposition 6.2 shows that a valid $\tilde{\theta}_t$ can be found by solving Equation (**P2**), bypassing the need to compute θ_t^p. Essentially, the initial two-steps procedure to find $\tilde{\theta}_t$ (through the intermediary program Equation (**P1**)) is replaced by a single minimization program augmented with a slack variable η. The attentive reader may notice that Equation (**P2**) is now similar to Equation (**P0**), the projection step employed in [Filippi et al., 2010]. As a result, BVD-GLM-UCB is comparable to the original algorithm GLM-UCB in terms of computational burden. The proof of Proposition 6.2 is given in Section 6.C in the appendix.

6.4.4 Online Estimation of the Variation Budget

Motivation. The attentive reader may notice that the minimax-optimality of BVD-GLM-UCB is conditioned on the knowledge of an upper-bound B_T for the true parameter-drift $\mathcal{B}_T$. Naturally, the tighter this upper-bound, the better the performance. Yet, whether such a knowledge is available in real-life problems is, to say the least, questionable. This issue is not specific to our approach but is shared with all non-stationary parametric bandit methods - see for instance [Cheung et al., 2019, Zhao et al., 2020]. For linear bandits, previous approaches circumvented this drawback with a Bandit-over-Bandit strategy [Cheung et al., 2021, Section 7], where $\mathcal{B}_T$ is learned online by a *master* algorithm. This guarantees sub-linear regret without having the knowledge of $\mathcal{B}_T$. We however note that this technique was specialized for linear bandits and for the sliding-window strategy. One could easily design a sliding-window approach of BVD-GLM-UCB (using very similar arguments as the ones displayed in this chapter) and extend the Bandit-over-Bandit of [Cheung et al., 2021] to the GLB framework. Here, we follow a different path and introduce an equivalent method for the exponential-weighting strategy. To the best of our knowledge, this technique was missing in the non-stationary parametric bandit literature. It notably proves that the online learning of $\mathcal{B}_T$ can be efficiently performed under discounted strategies.

Bandit-over-Bandit for discounted strategies. For the sake of simplicity, we describe the Bandit-over-bandit approach adopted when Assumption 6.4 holds. A similar reasoning holds in general but naturally yields different rates. Notice that naive bounding gives $\mathcal{B}_T \in (0, 2ST]$. The main idea for learning $\mathcal{B}_T$ online is to grid on a log-scale the interval $(0, 2ST]$ with N values $\{\mathcal{B}_j\}_{j=1}^N$. We then create N instances of BVD-GLM-UCB, each set with a different discount factor:

$$\gamma_j = 1 - \left(\frac{\mathcal{B}_j}{dT}\right)^{2/3} = 1 - \frac{2^{j-1}}{2^{5/3} d^{2/3} T S^{2/3}} \, .$$

These instances will be our *experts*. We then deploy a *master* algorithm - a version of EXP3 [Auer et al., 2002b], which acts repeatedly as follows: **1.** it chooses an expert j (*i.e* a new instance of BVD-GLM-UCB with parameter γ_j) to interact with the environment during a time frame of length H (H is a positive integer). **2.** The master algorithm then observes the cumulative reward (aggregated on the time frame) of the expert j. We give the pseudo-algorithm of this procedure in Algorithm 14.

Input: Length H, time horizon T, regularization λ, confidence δ, inverse link function μ, constants S, L and m.
Initialization: Let $N \leftarrow \lceil 2\log_2(2ST^{3/2}) \rceil$ and $\mathcal{H} \leftarrow \{\gamma_j = 1 - \frac{2^{j-1}}{2^{5/3} d^{2/3} T S^{2/3}}\}_{j=1}^N$, initialize EXP3 with action set indexed by $\mathcal{H}$.
for $i = 1, \dots, \lceil T/H \rceil$ **do**
  $j \leftarrow$ action selected by EXP3
  Initialize a sub-routine BVD-GLM-UCB with parameter γ_j.
  for $t = 1, \dots, H$ **do**
    Play with BVD-GLM-UCB with parameter γ_j, observe reward X_t.
  Update EXP3 with reward $\sum_{t=1}^H X_t$.

Algorithm 14: BOB-BVD-GLM-UCB (a more detailed version is deferred to Appendix 6.D.2).

Informally, the idea is that EXP3 will learn to select the best performing γ_j associated with the best estimate $\mathcal{B}_j$ of $\mathcal{B}_T$. Intuitively, this should guarantee small regret as EXP3 will mostly play instances of BVD-GLM-UCB which nearly capture the true magnitude of the non-stationarity. This intuition is made rigorous in Theorem 6.3, whose proof is deferred to Section 6.D in the appendix.

Theorem 6.3. *Under Assumptions 6.1-6.2 and 6.4, the expected regret of* BOB-BVD-GLM-UCB *when setting* $H = \lfloor d\sqrt{T} \rfloor$ *satisfies:*

$$\mathcal{R}(T) = \tilde{\mathcal{O}}\left(R_\mu d^{2/3} T^{2/3} \max\left(\mathcal{B}_T, d^{-1/2} T^{1/4}\right)^{1/3}\right) \, .$$

Under the orthogonal arm-set assumption, we obtain a regret bound which is identical to the ones of the Bandit-over-Bandit algorithms of [Cheung et al., 2021] and [Zhao et al., 2020]. The conclusions are therefore of similar nature: namely, when $\mathcal{B}_T \geq d^{-1/2} T^{1/4}$ we obtain a minimax rate, *without* knowing $\mathcal{B}_T$. Again, note here the presence of the problem-dependant constant R_μ, inherited from the non-linear reward structure imposed in GLBs.

6.5 Proof Sketch

In this section, we detail the key steps of the proof of Theorem 6.1. In particular, we shed light on the tension between the learning and tracking aspects of the problem and their role in the choice of the estimator $\tilde{\theta}_t$, through the use of an appropriate projection step. For simplicity we assume that Assumption 6.4 holds, although the spirit of the proof is almost identical in the general case.

Learning versus tracking. A crucial feature of non-stationary GLBs lies in the singular nature of the deviation of $\hat{\theta}_t$ from $\theta^\star_{t+1}$. This arises from two fundamentally different mechanisms: learning and tracking. We introduce the following estimator, which allows for a clean-cut distinction between the two phenomenons:

$$\bar{\theta}_t \coloneqq \operatorname{argmin}_{\theta\in\mathbb{R}^d} \left\{ \sum_{s=1}^{t} \gamma^{t-s} \left[b(A_s^\top \theta) - \mu(A_s^\top \theta_s^\star) A_s^\top \theta \right] + \frac{\lambda c_\mu}{2} \left\| \theta - \theta^\star_{t+1} \right\|_2^2 \right\}. \tag{6.4}$$

The parameter $\bar{\theta}_t$ is the minimizer of a strictly convex and coercive function, thus is well-defined and unique. Intuitively, $\bar{\theta}_t$ would be the estimator obtained under a perfect (e.g noiseless) observation of the reward[3]. As a result, the deviation between $\hat{\theta}_t$ and $\bar{\theta}_t$ is solely due to the stochastic nature of the problem (*learning*). On the other hand, the deviation between $\bar{\theta}_t$ and $\theta^\star_{t+1}$ is a consequence of the unpredictable changes of the sequence $\{\theta^\star_s\}_s$ (*tracking*). The introduction of the reference point $\bar{\theta}_t$ allows us to characterize both deviations separately in Lemma 6.4 and Lemma 6.5.

Lemma 6.4. *[Learning] Let $\delta \in (0, 1]$. With probability at least $1 - \delta$:*

$$\text{for all } t \geq 1, \quad \bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t) = \left\{ \theta \in \mathbb{R}^d \ s.t \ \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\widetilde{V}_t^{-1}} \leq \beta_t(\delta) \right\}.$$

Lemma 6.4 ensures that with high probability the set $\mathcal{E}_t^\delta(\hat{\theta}_t)$ is a *confidence set* for $\bar{\theta}_t$. A complete proof of this result is deferred to Section 6.A.1 in the supplementary material.

Lemma 6.5. *[Tracking with orthogonal action sets] Let $D \in \mathbb{N}^*$. The following holds:*

$$\left\| g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1}) \right\|_{V_t^{-2}} \leq \frac{2k_\mu L^2 S}{\lambda} \frac{\gamma^D}{1-\gamma} + k_\mu \sum_{s=t-D+1}^{t} \left\| \theta^\star_s - \theta^\star_{s+1} \right\|_2.$$

Lemma 6.5 effectively links the deviation of $\bar{\theta}_t$ from $\theta^\star_t$ to the variation-budget $\mathcal{B}_T$ through the drift $\sum_{s=t-D+1}^{t} \left\| \theta^\star_s - \theta^\star_{s+1} \right\|_2$. The proof of this result borrows tools from [Russac et al., 2019] and is deferred to Section 6.A.5 in appendix. The integer D appearing in Lemma 6.5 is introduced for the sake of the analysis only. It allows to treat separately old and recent observations. We provide its optimal value later in this section.

[3]Note the difference between $\hat{\theta}_t$ and $\bar{\theta}_t$, where the rewards X_t are replaced by their conditional expected values $\mu(A_s^\top \theta_s^\star)$

Remark 6.3. *Behind the statement of Lemma 6.4 and Lemma 6.5 hides the main reason why the projection step of [Filippi et al., 2010] needs to be generalized. Indeed, it appears that the deviations $(\hat{\theta}_t \leftrightarrow \bar{\theta}_t)$ and $(\bar{\theta}_t \leftrightarrow \theta^\star_{t+1})$ are controlled through different metrics ($\widetilde{V}_t^{-1}$ and V_t^{-2}, respectively). Projecting according to the first metric would corrupt the control of the second deviation, and conversely.*

Regret decomposition and prediction error. To bound the instantaneous regret at round t, we rely on the prediction error Δ_t defined as follows for any arm $a \in \mathcal{A}_t$:

$$\Delta_t(a) \coloneqq \left|\mu(a^\top \tilde{\theta}_t) - \mu(a^\top \theta^\star_{t+1})\right|.$$

The next Lemma ties the cumulative regret to the sum of prediction errors. This derivation is classical and the proof is deferred to Section 6.B.1 in the supplementary material.

Lemma 6.6. *The following holds:*

$$R(T) \le 2R_\mu \sum_{t=1}^{T} \beta_{t-1}(\delta)\left[\|A_t\|_{V_{t-1}^{-1}} - \|A_{t,\star}\|_{V_{t-1}^{-1}}\right] + \sum_{t=1}^{T}\left[\Delta_{t-1}(A_t) + \Delta_{t-1}(A_{t,\star})\right].$$

Thanks to Lemma 6.6 we are left to characterize the prediction error $\Delta_t(a)$ for any $a \in \mathcal{A}_{t+1}$. Following [Filippi et al., 2010], we rely on the mean-value theorem to ensure that it exists $\mathring{\theta}_t \in [\tilde{\theta}_t, \theta^\star_t]$ such that[4]:

$$\Delta_t(a) \le k_\mu \left\langle a, H_t(\mathring{\theta}_t)\left(g_t(\tilde{\theta}_t) - g_t(\theta^\star_t)\right)\right\rangle, \tag{6.5}$$

where $H_t(\theta) \coloneqq \sum_{s=1}^{t} \dot{\mu}(A_s^\top \theta) A_s A_s^\top + \lambda c_\mu I_d$. Since $\tilde{\theta}_t, \theta^\star_t \in \Theta$, we obtain by convexity that $\mathring{\theta}_t \in \Theta$ and we can use the lower bound $H_t(\mathring{\theta}_t) \succeq c_\mu V_t$.

Remark 6.4. *In this last inequality resides the mistake that was made in previous extension of [Filippi et al., 2010] to the non-stationary setting [Cheung et al., 2021, Zhao et al., 2020]. Indeed, if the prediction error is measured at $\hat{\theta}_t$, we are left with $\mathring{\theta}_t \in [\theta^\star_t, \hat{\theta}_t]$, and $\mathring{\theta}_t$ can lie outside of the admissible set Θ (since $\hat{\theta}_t$ can). The lower-bound linking $H_t(\mathring{\theta}_t)$ and V_t would therefore not hold. More precisely, and as detailed in Section 6.3.2, when $\mathring{\theta}_t \in [\theta^\star_t, \hat{\theta}_t]$ not much can be said on the link between $H_t(\mathring{\theta}_t)$ and V_t without severely degrading the final regret guarantees.*

Adding and removing $g_t(\hat{\theta}_t) + g_t(\theta^p_t) + g_t(\bar{\theta}_t)$ inside the inner-product in Equation (6.5), followed by easy manipulations yields:

$$\begin{aligned}\Delta_t(a) \le &\underbrace{R_\mu \|a\|_{V_t^{-1}}\left(\left\|g_t(\tilde{\theta}_t) - g_t(\theta^p_t)\right\|_{\widetilde{V}_t^{-1}} + \left\|g_t(\bar{\theta}_t) - g_t(\hat{\theta}_t)\right\|_{\widetilde{V}_t^{-1}}\right)}_{\coloneqq \Delta_t^{\mathrm{learn}}(a)} \\ &+ \underbrace{R_\mu \|a\|_2\left(\left\|g_t(\theta^p_t) - g_t(\hat{\theta}_t)\right\|_{V_t^{-2}} + \left\|g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1})\right\|_{V_t^{-2}}\right)}_{\coloneqq \Delta_t^{\mathrm{track}}(a)}.\end{aligned}$$

[4]Formally, $\mathring{\theta}_t \in [\tilde{\theta}_t, \theta^\star_t]$ means that there exists $v \in [0,1]$ such that $\mathring{\theta}_t = v\tilde{\theta}_t + (1-v)\theta^\star_t$.

Leveraging the projection step. We can now bound the terms $\Delta_t^{\text{learn}}(a)$ and $\Delta_t^{\text{track}}(a)$ separately. Lemma 6.4 along with the design $\tilde{\theta}_t \in \mathcal{E}_t^{\delta}(\theta_t^p)$ leads to:

$$\Delta_t^{\text{learn}}(a) \leq 2R_\mu \|a\|_{V_t^{-1}} \beta_t(\delta) \quad \text{w.h.p} \tag{6.6}$$

The first term in $\Delta_t^{\text{track}}(a)$ is kept under control by the specific design of the projection step Equation (**P1**). This is formalized in the following Lemma, whose proof is deferred to Section 6.A.4 in the appendix.

Lemma 6.7. *Under the event $\{\bar{\theta}_t \in \mathcal{E}_t^{\delta}(\hat{\theta}_t)\}$ the following holds:*

$$\|g_t(\theta_t^p) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \leq \|g_t(\bar{\theta}_t) - g_t(\theta_{t+1}^\star)\|_{V_t^{-2}} .$$

As a result, bounding $\Delta_t^{\text{track}}(a)$ reduces to bounding $\|g_t(\bar{\theta}_t) - g_t(\theta_{t+1}^\star)\|_{V_t^{-2}}$. Combined with Lemma 6.5, this result states that the deviation between θ_t^p and $\hat{\theta}_t$ is characterized by $\mathcal{B}_t$, the parameter-drift up to round t, as illustrated in Figure 6.1. This leads to:

$$\Delta_t^{\text{track}}(a) \leq 2R_\mu \|a\|_2 \left(\frac{2k_\mu L^2 S}{\lambda} \frac{\gamma^D}{1-\gamma} + k_\mu \sum_{s=t-D+1}^{t} \left\|\theta_s^\star - \theta_{s+1}^\star\right\|_2 \right) \quad \text{w.h.p} \tag{6.7}$$

Putting everything together. Combining Equation (6.6) and Equation (6.7) with Lemma 6.6 and the Elliptical Lemma (Lemma 5.30 from Chapter 5) yields:

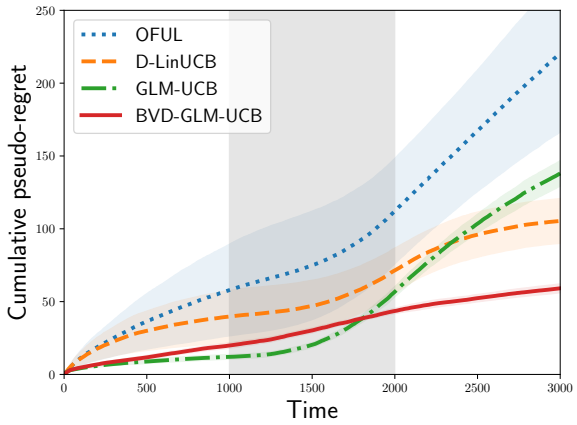
$$R(T) \leq C_1 R_\mu dT \log(1/\gamma) + C_2 R_\mu \gamma^D T/(1-\gamma) + C_3 R_\mu D B_T \qquad \text{w.h.p}$$

where the constants C_1, C_2 and C_3 hide $\log(T)$ multiplicative dependencies. A detailed proof of this result is deferred to Section 6.B.2 in the supplementary material. Setting the hyper-parameters $D = \log(T)/(1-\gamma)$ and $\gamma = 1 - (\frac{B_T}{dT})^{2/3}$ concludes the proof of Theorem 6.1.

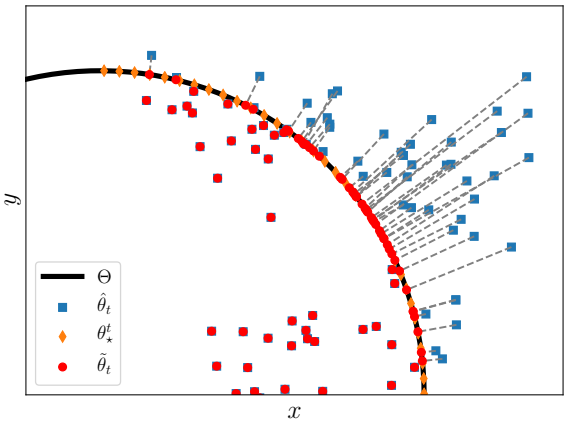
6.6 Experiments

We illustrate in Figure 6.2 the behavior and performance of BVD-GLM-UCB with numerical simulations in a two-dimensional non-stationary logistic environment. Formally, we let $X_t \sim \text{Bernoulli}(\mu(A_t^\top \theta_t^\star))$ where $\mu(z) = (1+e^{-z})^{-1}$ is the logistic function. The sequence $\{\theta_t^\star\}_{t\geq 1}$ evolves as follows: we let $\theta_t^\star = (0,1)$ for $t \in [1, T/3]$. Between $t = T/3$ and $t = 2T/3$ we smoothly rotate $\theta_t^\star$ from $(0,1)$ to $(1,0)$. Finally we let $\theta_t^\star = (0,1)$ for $t \in [2T/3, T]$. A thorough description of the experimental setting can be found in Appendix 6.E. We compare in Figure 6.2a the four following algorithms: OFUL [Abbasi-Yadkori et al., 2011] (stationary, here mispecified), GLM-UCB [Filippi et al., 2010] (stationary, here well-specified), D-LinUCB [Russac et al., 2019] (an exponentially weighted LB algorithm, non-stationary but here mispecified) and BVD-GLM-UCB (non-stationary, well-specified). For D-LinUCB and BVD-GLM-UCB we use the value of γ recommended by the asymptotic analysis. This figure highlights the necessity to employ algorithms that are well-specified; both GLM-UCB and BVD-GLM-UCB outperform their linear counterparts (OFUL and D-LinUCB, respectively). Note that an appropriate treatment of non-stationarity is also crucial to obtain small regret as for the considered horizon the two best performing algorithms are D-LinUCB and BVD-GLM-UCB. The latter being well-specified

and resilient to non-stationary, it naturally performs best. In Figure 6.2b we highlight the fact that the projection step is necessary as, in this non-stationary setting, $\hat{\theta}_t$ regularly leaves the admissible set Θ.



(a) Regret bounds of different stochastic bandit algorithms under parameter-drift. The grey region indicates a smooth drift of $\theta_t^\star$.



(b) Evolution of the parameters of interest $(\theta_t^\star, \hat{\theta}_t, \tilde{\theta}_t)$ for BVD-GLM-UCB. Note that in this non-stationary setting $\hat{\theta}_t \notin \Theta$ is frequent.

Figure 6.2: Numerical simulations in a non-stationary logistic setting. For the first figure, results are averaged over 50 independent runs and shaded areas represent one standard-deviation variation.

6.7 Conclusion

We highlighted in this chapter a central difficulty in the theoretical treatment of non-stationary GLBs, overlooked in existing approaches and intimately linked to the non-linear nature of the reward function. To overcome this difficulty, we introduced a generalization of the projection step from [Filippi et al., 2010], which allows to simultaneously *track* the non-stationary ground-truth while preserving the *learning* guarantees of weighted maximum-likelihood strategies. This novel algorithmic design along with a careful analysis proves that an order-optimal (w.r.t d, T and B_T) regret-bound can be achieved for GLBs under parameter-drift, although up to a rather restrictive assumption on the arm set's geometry. The nature of the minimax-rates in the general case is open, as in both LB (c.f. [Touati and Vincent, 2020]) and GLB setting (this chapter) we observe a mismatch between existing upper-bounds and the lower-bound of [Cheung et al., 2019].

We underlined in Section 6.3.2 the problematic scaling of the problem-dependent constant R_μ. Consequent research efforts have recently been deployed to reduce its impact on regret-bounds, both in the stationary [Faury et al., 2020, Jun et al., 2021] and piece-wise stationary settings (Chapter 5). What is the optimal dependency w.r.t R_μ in the more general parameter-drift setting, and how it can be achieved are exciting open questions that we here leave for future work.

Appendix

The appendix is organized as follows:

- In Section 6.A we provide some concentration results, along with a bound on the prediction error Δ_t inherited from the design of the projection step.
- In Section 6.B we link the prediction error Δ_t to the regret $R(T)$ of BVD-GLM-UCB. We then proceed to prove the bound on $R(T)$ announced in Theorem 6.1.
- In Section 6.C we provide a proof for the equivalence of the optimization programs Equation (**P1**) (along with the computation of $\tilde{\theta}_t$) and Equation (**P2**).
- In Section 6.D we provide a proof for the regret upper-bound of BOB-BVD-GLM-UCB claimed in Theorem 6.3.
- Finally, in Section 6.E we provide some details on our numerical simulations.

Appendix 6.A Concentration and Predictions Bound

6.A.1 Confidence sets

Lemma 6.4. *[Learning] Let $\delta \in (0,1]$. With probability at least $1-\delta$:*

$$\text{for all } t \geq 1, \quad \bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t) = \left\{ \theta \in \mathbb{R}^d \ s.t \ \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\widetilde{V}_t^{-1}} \leq \beta_t(\delta) \right\}.$$

Proof. Recall that:

$$\mathcal{E}_t^\delta(\hat{\theta}_t) = \left\{ \theta \in \mathbb{R}^d \text{ s.t } \|g_t(\theta) - g_t(\hat{\theta}_t)\|_{\widetilde{V}_t^{-1}} \leq \beta_t(\delta) \right\} ,$$

where

$$\beta_t(\delta) \coloneqq \sqrt{\lambda} c_\mu S + \frac{m}{2} \sqrt{2\log(1/\delta) + d\log\left(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)}\right)} .$$

Also, from the definition of $\bar{\theta}_t$ in Equation (6.4), by setting to 0 the differential of the convex objective minimized by $\bar{\theta}_t$ we obtain that:

$$g_t(\bar{\theta}_t) = \sum_{s=1}^{t} \gamma^{t-s} \mu(A_s^\top \theta_s^\star) A_s + \lambda c_\mu \theta_{t+1}^\star . \tag{6.8}$$

Further, for all $s \geq 1$, define

$$\epsilon_s = X_s - \mu(A_s^\top \theta_s^\star) . \tag{6.9}$$

Note that:

$$\begin{cases} \mathbb{E}\left[\epsilon_s | \mathcal{F}_{s-1}\right] = 0 & \text{(Equation (6.1))} \\ -\mu(A_s^\top \theta_s^\star) \leq \epsilon_s \leq m + \mu(A_s^\top \theta_s^\star) \quad \text{a.s.} & \text{(Assumption 6.2)} \end{cases}$$

Therefore ϵ_s is $m/2$-subGaussian conditionally on $\mathcal{F}_{s-1}$. Furthermore, by optimality of $\hat{\theta}_t$,

differentiating the objective function in Equation (6.2) yields:

$$\begin{aligned}
&\sum_{s=1}^{t} \gamma^{t-s} \left[\mu(A_s^\top \hat{\theta}_t) - X_s\right] A_s + \lambda c_\mu \hat{\theta}_t = 0 \\
&\Leftrightarrow g_t(\hat{\theta}_t) = \sum_{s=1}^{t} \gamma^{t-s} \mu(A_s^\top \theta_s^\star) A_s + \sum_{s=1}^{t} \gamma^{t-s} \epsilon_s A_s \qquad \text{(Equation (6.9))} \\
&\Leftrightarrow g_t(\hat{\theta}_t) = g_t(\bar{\theta}_t) + \sum_{s=1}^{t} \gamma^{t-s} \epsilon_s A_s - \lambda c_\mu \theta_{t+1}^\star \qquad \text{(Equation (6.8))} \qquad (6.10) \\
&\Leftrightarrow \|g_t(\bar{\theta}_t) - g_t(\hat{\theta}_t)\|_{\widetilde{V}_t^{-1}} = \left\| \sum_{s=1}^{t} \gamma^{t-s} \epsilon_s A_s - \lambda c_\mu \theta_{t+1}^\star \right\|_{\widetilde{V}_t^{-1}} .
\end{aligned}$$

Therefore since $\theta_{t+1}^\star \in \Theta$ and $\widetilde{V}_t \succeq \lambda I_d$ we obtain:

$$\|g_t(\bar{\theta}_t) - g_t(\hat{\theta}_t)\|_{\widetilde{V}_t^{-1}} \leq \sqrt{\lambda} c_\mu S + \left\| \sum_{s=1}^{t} \gamma^{t-s} \epsilon_s A_s \right\|_{\widetilde{V}_t^{-1}} .$$

Simplifying the factors γ^t in the most right term and applying Proposition 4.8 from Chapter 4 proves that with probability at least $1-\delta$, for all $t \geq 1$:

$$\|g_t(\bar{\theta}_t) - g_t(\hat{\theta}_t)\|_{\widetilde{V}_t^{-1}} \leq \sqrt{\lambda} c_\mu S + \frac{m}{2} \sqrt{2\log(1/\delta) + d\log\left(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)}\right)} = \beta_t(\delta) ,$$

hence proving the desired result. □

6.A.2 Bounding the prediction error

Lemma 6.8. *Let $\delta \in (0,1]$ and $D \in \mathbb{N}^*$. With probability at least $1-\delta$: for all $t \geq 1$, for all $a \in \mathcal{A}_t$, under Assumption 6.4 the following holds.*

$$\Delta_t(a) \leq \frac{2k_\mu}{c_\mu} \beta_t(\delta) \|a\|_{V_t^{-1}} + \frac{4k_\mu^2 L^3 S}{c_\mu \lambda} \frac{\gamma^D}{(1-\gamma)} + \frac{2k_\mu^2 L}{c_\mu} \sum_{s=t-D+1}^{t} \|\theta_s^\star - \theta_{s+1}^\star\|_2 .$$

Without Assumption 6.4, under general arm-set geometry, the following holds.

$$\begin{aligned}
\Delta_t(a) \leq{}& \frac{2k_\mu}{c_\mu} \beta_t(\delta) \|a\|_{V_t^{-1}} \\
&+ \frac{2k_\mu L}{c_\mu} \sqrt{1 + \frac{L^2}{\lambda(1-\gamma)}} \left(\frac{2k_\mu S L^2}{\lambda} \frac{\gamma^D}{1-\gamma} + k_\mu \sqrt{\frac{d}{\lambda(1-\gamma)}} \sum_{s=t-D+1}^{t} \|\theta_s^\star - \theta_{s+1}^\star\|_2 \right) .
\end{aligned}$$

Proof. In the following, we assume that the event $E_\delta = \{\bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t) \text{ for all } t \geq 1\}$ holds, which happens with probability at least $1-\delta$ (Lemma 6.4). From the definition of the

prediction error:

$$\begin{aligned}\Delta_t(a) &= \left|\mu(a^\top \tilde{\theta}_t) - \mu(a^\top \theta^\star_{t+1})\right| \\ &\leq \left(\sup_{a\in\mathcal{A},\theta\in\Theta} \dot{\mu}(a^\top\theta)\right)\left|a^\top(\tilde{\theta}_t - \theta^\star_{t+1})\right| \qquad (a\in\mathcal{A}_t, \theta^\star_{t+1}\in\Theta, \tilde{\theta}_t\in\Theta) \\ &\leq k_\mu \left|\langle a, \tilde{\theta}_t - \theta^\star_{t+1}\rangle\right| . \qquad \text{(by definition of } k_\mu) \end{aligned} \tag{6.11}$$

Further, thanks to the mean value theorem:

$$\begin{aligned} g_t(\tilde{\theta}_t) - g_t(\theta^\star_{t+1}) &= \sum_{s=1}^{t} \gamma^{t-s}\left[\mu(A_s^\top\tilde{\theta}_t) - \mu(A_s^\top\theta^\star_{t+1})\right] + \lambda c_\mu(\tilde{\theta}_t - \theta^\star_{t+1}) \\ &= G_t\cdot(\tilde{\theta}_t - \theta^\star_{t+1}) , \end{aligned} \tag{6.12}$$

where:

$$G_t \coloneqq \sum_{s=1}^{t}\gamma^{t-s}\left[\int_{v=0}^{1}\dot{\mu}\left(\langle A_s, (1-v)\theta^\star_{t+1} + v\tilde{\theta}_t\rangle\right)\,dv\right]A_sA_s^\top + \lambda c_\mu I_d \succeq c_\mu V_t .$$

Note that because $A_s\in\mathcal{A}$ for all $s\in[t-1]$ and $\tilde{\theta}_t, \theta^\star_{t+1}\in\Theta$ we have $G_t\geq c_\mu V_t$. Assembling together Equation (6.11) and Equation (6.12) we get:

$$\begin{aligned}\Delta_t(a) &\leq k_\mu\left|\left\langle a, G_t^{-1}(g_t(\tilde{\theta}_t) - g_t(\theta^\star_{t+1}))\right\rangle\right| \\ &\leq k_\mu\left|\left\langle a, G_t^{-1}(g_t(\tilde{\theta}_t) - g_t(\theta^p_t) + g_t(\theta^p_t) - g_t(\hat{\theta}_t) + g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1}))\right\rangle\right| \\ &\leq \underbrace{k_\mu\left|\left\langle a, G_t^{-1}(g_t(\tilde{\theta}_t) - g_t(\theta^p_t) + g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t))\right\rangle\right|}_{\coloneqq\Delta_t^{\text{learn}}(a)} \\ &\quad + \underbrace{k_\mu\left|\left\langle a, G_t^{-1}(g_t(\theta^p_t) - g_t(\hat{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1}))\right\rangle\right|}_{\coloneqq\Delta_t^{\text{track}}(a)} \\ &\leq \Delta_t^{\text{learn}}(a) + \Delta_t^{\text{track}}(a) . \end{aligned} \tag{6.13}$$

This decomposition brings out the contribution of two different phenomenons (*learning* and *tracking*) which will be handled separately. Starting with the learning:

$$\begin{aligned}\Delta_t^{\text{learn}}(a) &= k_\mu\left|\left\langle a, G_t^{-1}(g_t(\tilde{\theta}_t) - g_t(\theta^p_t) + g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t))\right\rangle\right| \\ &= k_\mu\left|\left\langle \widetilde{V}_t^{1/2}G_t^{-1}a, \widetilde{V}_t^{-1/2}(g_t(\tilde{\theta}_t) - g_t(\theta^p_t) + g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t))\right\rangle\right| \\ &\leq k_\mu\|a\|_{G_t^{-1}\widetilde{V}_tG_t^{-1}}\left(\|g_t(\tilde{\theta}_t) - g_t(\theta^p_t)\|_{\widetilde{V}_t^{-1}} + \|g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t)\|_{\widetilde{V}_t^{-1}}\right) \quad \text{(Cauchy-Schwarz)} \\ &\leq k_\mu\|a\|_{G_t^{-1}V_tG_t^{-1}}\left(\|g_t(\tilde{\theta}_t) - g_t(\theta^p_t)\|_{\widetilde{V}_t^{-1}} + \|g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t)\|_{\widetilde{V}_t^{-1}}\right) \qquad (\widetilde{V}_t\leq V_t) \\ &\leq \frac{k_\mu}{\sqrt{c_\mu}}\|a\|_{G_t^{-1}}\left(\|g_t(\tilde{\theta}_t) - g_t(\theta^p_t)\|_{\widetilde{V}_t^{-1}} + \|g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t)\|_{\widetilde{V}_t^{-1}}\right) \qquad (V_t\leq c_\mu^{-1}G_t) \\ &\leq \frac{k_\mu}{c_\mu}\|a\|_{V_t^{-1}}\left(\|g_t(\tilde{\theta}_t) - g_t(\theta^p_t)\|_{\widetilde{V}_t^{-1}} + \|g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t)\|_{\widetilde{V}_t^{-1}}\right) \qquad (G_t^{-1}\leq c_\mu^{-1}V_t^{-1}) \\ &\leq \frac{k_\mu}{c_\mu}\|a\|_{V_t^{-1}}\left(\beta_t(\delta) + \|g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t)\|_{\widetilde{V}_t^{-1}}\right) \qquad (\tilde{\theta}_t\in\mathcal{E}_t^\delta(\theta^p_t)) \\ &\leq \frac{k_\mu}{c_\mu}\|a\|_{V_t^{-1}}\left(\beta_t(\delta) + \beta_t(\delta)\right) . \qquad (E_\delta \text{ holds}) \end{aligned}$$

We used $\widetilde{V}_t \leq V_t$ which is a consequence of $\gamma \in (0,1)$. As a result:

$$\Delta_t^{\text{learn}}(a) \leq \frac{2k_\mu}{c_\mu}\beta_t(\delta)\|a\|_{V_t^{-1}} \,. \tag{6.14}$$

The tracking term is bounded differently when the action set satisfies Assumption 6.4 or for general arm-set geometry. The bound on the tracking term is reported in Lemma 6.9 and its proof is reported in Section 6.A.3

Lemma 6.9. *Let $D \in \mathbb{N}^*$. When Assumption 6.4 holds, we have the following:*

$$\Delta_t^{track}(a) \leq \frac{4k_\mu^2 L^3 S}{c_\mu \lambda}\frac{\gamma^D}{(1-\gamma)} + \frac{2k_\mu^2 L}{c_\mu}\sum_{s=t-D+1}^{t}\left\|\theta_s^\star - \theta_{s+1}^\star\right\|_2 \,. \tag{6.15}$$

For general arm-set geometry, we have the following

$$\Delta_t^{track}(a) \leq \frac{2k_\mu L}{c_\mu}\sqrt{1+\frac{L^2}{\lambda(1-\gamma)}}\left(\frac{2k_\mu S L^2}{\lambda}\frac{\gamma^D}{1-\gamma} + k_\mu\sqrt{\frac{d}{\lambda(1-\gamma)}}\sum_{s=t-D+1}^{t}\left\|\theta_s^\star - \theta_{s+1}^\star\right\|_2\right)$$

Assembling Equation (6.13), Equation (6.14) and the two different inequalities from Lemma 6.9 gives the two statements of the proof. □

6.A.3 Proof of Lemma 6.9

Proof. Throughout the proof, we will use the following lemma, proven in Section 6.A.4.

Lemma 6.7. *Under the event $\{\bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t)\}$ the following holds:*

$$\|g_t(\theta_t^p) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \leq \|g_t(\bar{\theta}_t) - g_t(\theta_{t+1}^\star)\|_{V_t^{-2}} \,.$$

With Assumption 4. In this first part of the proof, we assume that Assumption 6.4 holds. We have the following:

$$\begin{aligned}
\Delta_t^{\text{track}}(a) &= k_\mu\left|\left\langle a, G_t^{-1}(g_t(\theta_t^p) - g_t(\hat{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta_{t+1}^\star))\right\rangle\right| \\
&\leq k_\mu \|a\|_2 \left\|g_t(\theta_t^p) - g_t(\hat{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta_{t+1}^\star)\right\|_{G_t^{-2}} && \text{(Cauchy-Schwarz)} \\
&\leq \frac{k_\mu L}{c_\mu}\left\|g_t(\theta_t^p) - g_t(\hat{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta_{t+1}^\star)\right\|_{V_t^{-2}} && (\|x\|_2 \leq L, G_t^2 \succeq c_\mu^2 V_t^2) \\
&\leq \frac{k_\mu L}{c_\mu}\left(\left\|g_t(\theta_t^p) - g_t(\hat{\theta}_t)\right\|_{V_t^{-2}} + \left\|g_t(\bar{\theta}_t) - g_t(\theta_{t+1}^\star)\right\|_{V_t^{-2}}\right) && \text{(Triangle inequality)} \\
&\leq \frac{2k_\mu L}{c_\mu}\left\|g_t(\bar{\theta}_t) - g_t(\theta_{t+1}^\star)\right\|_{V_t^{-2}} && \text{(Lemma 6.7)}
\end{aligned}$$

where the third inequality can be obtained only because when Assumption 6.4 holds, G_t and V_t commute.
The final result is obtained using Lemma 6.5 reported here and established in Section 6.5

Lemma 6.5. *[Tracking with orthogonal action sets] Let $D \in \mathbb{N}^*$. The following holds:*

$$\|g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1})\|_{V_t^{-2}} \leq \frac{2k_\mu L^2 S}{\lambda}\frac{\gamma^D}{1-\gamma} + k_\mu \sum_{s=t-D+1}^{t} \|\theta^\star_s - \theta^\star_{s+1}\|_2 .$$

Without Assumption 4. We now explain how to extend the analysis with general arm-set geometry.

$$\begin{aligned}
\Delta_t^{\text{track}}(a) &= k_\mu \left|\left\langle a, G_t^{-1}(g_t(\theta^p_t) - g_t(\hat{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1}))\right\rangle\right| \\
&= k_\mu \left|\left\langle a, G_t^{-1}V_tV_t^{-1}(g_t(\theta^p_t) - g_t(\hat{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1}))\right\rangle\right| \\
&\leq k_\mu \|a\|_{G_t^{-1}V_t^2G_t^{-1}} \left\|g_t(\theta^p_t) - g_t(\hat{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1})\right\|_{V_t^{-2}} \quad \text{(Cauchy-Schwarz)} \\
&\leq k_\mu \sqrt{\lambda_{\max}(V_t)}\|a\|_{G_t^{-1}V_tG_t^{-1}} \left\|g_t(\theta^p_t) - g_t(\hat{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1})\right\|_{V_t^{-2}} \\
&\leq \frac{k_\mu}{\sqrt{c_\mu}}\sqrt{\lambda_{\max}(V_t)}\|a\|_{G_t^{-1}} \left\|g_t(\theta^p_t) - g_t(\hat{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1})\right\|_{V_t^{-2}} \quad (G_t \succeq c_\mu V_t) \\
&\leq \frac{k_\mu L}{\sqrt{\lambda c_\mu}}\sqrt{\lambda_{\max}(V_t)} \left\|g_t(\theta^p_t) - g_t(\hat{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1})\right\|_{V_t^{-2}} \quad (G_t \succeq \lambda c_\mu I_d) \\
&\leq \frac{k_\mu L}{\sqrt{\lambda c_\mu}}\sqrt{\lambda_{\max}(V_t)} \left(\left\|g_t(\theta^p_t) - g_t(\hat{\theta}_t)\right\|_{V_t^{-2}} + \left\|g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1})\right\|_{V_t^{-2}}\right) \\
&\leq \frac{2k_\mu L}{\sqrt{\lambda c_\mu}}\sqrt{\lambda_{\max}(V_t)} \left\|g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1})\right\|_{V_t^{-2}} \quad \text{(Lemma 6.7)}
\end{aligned}$$

We then use:

$$\lambda_{\max}(V_t) \leq \frac{L^2}{1-\gamma} + \lambda . \tag{6.16}$$

That can be obtained by computing the operator norm of the matrix V_t. Combining this with Lemma 6.10 reported here and proved in Section 6.A.6 achieves the proof.

Lemma 6.10. *[Tracking with general action sets] Let $D \in \mathbb{N}^*$. The following holds:*

$$\|g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1})\|_{V_t^{-2}} \leq \frac{2k_\mu L^2 S}{\lambda}\frac{\gamma^D}{1-\gamma} + \frac{k_\mu}{\sqrt{\lambda}}\frac{\sqrt{d}}{\sqrt{1-\gamma}} \sum_{s=t-D+1}^{t} \|\theta^\star_s - \theta^\star_{s+1}\|_2 .$$

□

6.A.4 Proof of Lemma 6.7

Lemma 6.7. *Under the event $\{\bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t)\}$ the following holds:*

$$\|g_t(\theta^p_t) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \leq \|g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1})\|_{V_t^{-2}} .$$

Proof. We prove this result by contradiction. Assume that:

$$\|g_t(\theta^p_t) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} > \|g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1})\|_{V_t^{-2}}, \tag{6.17}$$

For all $s \geq 1$ define:

$$\tilde{X}_s \coloneqq \mu(A_s^\top \theta^\star_{t+1}) + \epsilon_s \ , \tag{6.18}$$

where $\{\epsilon_s\}_s$ is defined in Equation (6.9). Further, let:

$$\theta_c \coloneqq \operatorname{argmin}_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t} \gamma^{t-s} \left[b(A_s^\top \theta) - \tilde{X}_s A_s^\top \theta \right] + \frac{\lambda c_\mu}{2} \|\theta\|_2^2 \ ,$$

which is well-defined as the minimizer of a strictly convex, coercive function. Upon differentiating we get:

$$\begin{aligned}
g_t(\theta_c) &= \sum_{s=1}^{t} \gamma^{t-s} \tilde{X}_s A_s \\
&= \sum_{s=1}^{t} \gamma^{t-s} \epsilon_s A_s + \sum_{s=1}^{t} \gamma^{t-s} \mu(A_s^\top \theta^\star_{t+1}) A_s && \text{(Equation (6.18))} \\
&= g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t) + \lambda c_\mu \theta^\star_{t+1} + \sum_{s=1}^{t} \gamma^{t-s} \mu(A_s^\top \theta^\star_{t+1}) A_s && \text{(Equation (6.10))} \\
&= g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t) + g_t(\theta^\star_{t+1}) \ . && (6.19)
\end{aligned}$$

Therefore:

$$\begin{aligned}
\|g_t(\theta_c) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} &= \|g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1})\|_{V_t^{-2}} \\
&< \|g_t(\theta_t^p) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \ . && \text{(Equation (6.17))}
\end{aligned}$$

Further from Equation (6.19) we get:

$$\begin{aligned}
\|g_t(\theta_c) - g_t(\theta^\star_{t+1})\|_{\tilde{V}_t^{-1}} &= \|g_t(\bar{\theta}_t) - g_t(\hat{\theta}_t)\|_{\tilde{V}_t^{-1}} \\
&\leq \beta_t(\delta) && (\bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t)) \\
&\Leftrightarrow \theta^\star_{t+1} \in \mathcal{E}_t^\delta(\theta_c) \ .
\end{aligned}$$

To sum-up, we have $\|g_t(\theta_c) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} < \|g_t(\theta_t^p) - g_t(\hat{\theta}_t)\|_{V_t^{-2}}$ and $\mathcal{E}_t^\delta(\theta_c) \cap \Theta \neq \emptyset$ since $\theta^\star_{t+1} \in \Theta \cap \mathcal{E}_t^\delta(\theta_c)$. This contradicts the definition of θ_t^p (in Equation (**P1**)) and therefore Equation (6.17) must be wrong, which proves the announced result. □

6.A.5 Proof of Lemma 6.5

Lemma 6.5. *[Tracking with orthogonal action sets] Let $D \in \mathbb{N}^*$. The following holds:*

$$\|g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1})\|_{V_t^{-2}} \leq \frac{2k_\mu L^2 S}{\lambda} \frac{\gamma^D}{1-\gamma} + k_\mu \sum_{s=t-D+1}^{t} \|\theta_s^\star - \theta^\star_{s+1}\|_2 \, .$$

Proof. Thanks to Equation (6.8) we have:

$$
\begin{aligned}
& g_t(\bar{\theta}_t) = \sum_{s=1}^{t} \gamma^{t-s} \mu(A_s^\top \theta_s^\star) A_s + \lambda c_\mu \theta_{t+1}^\star \\
\Leftrightarrow\ & g_t(\bar{\theta}_t) - g_t(\theta_{t+1}^\star) = \sum_{s=1}^{t} \gamma^{t-s} \left[\mu(A_s^\top \theta_s^\star) - \mu(A_s^\top \theta_{t+1}^\star) \right] A_s \\
\Leftrightarrow\ & g_t(\bar{\theta}_t) - g_t(\theta_{t+1}^\star) = \sum_{s=1}^{t} \gamma^{t-s} \left[\int_{v=0}^{1} \dot{\mu}\left(\langle A_s, v\theta_{t+1}^\star + (1-v)\theta_s^\star \rangle \right) dv \right] A_s A_s^\top (\theta_s^\star - \theta_{t+1}^\star) \\
\Leftrightarrow\ & g_t(\bar{\theta}_t) - g_t(\theta_{t+1}^\star) = \sum_{s=1}^{t} \gamma^{t-s} \alpha_s A_s A_s^\top (\theta_s^\star - \theta_{t+1}^\star) \ ,
\end{aligned}
$$

where we defined:

$$
\alpha_s \coloneqq \int_{v=0}^{1} \dot{\mu}\left(\langle A_s, v\theta_{t+1}^\star + (1-v)\theta_s^\star \rangle \right) dv \in [c_\mu, k_\mu] \ .
$$

Therefore:

$$
\| g_t(\bar{\theta}_t) - g_t(\theta_{t+1}^\star) \|_{V_t^{-2}} = \left\| \sum_{s=1}^{t} \gamma^{t-s} \alpha_s A_s A_s^\top (\theta_s^\star - \theta_{t+1}^\star) \right\|_{V_t^{-2}} \ . \tag{6.20}
$$

The rest of the proof follows the strategy from Section 4.3.2.3 to yield the announced result. Let $D \in \mathbb{N}^*$ and notice that:

$$
\begin{aligned}
\left\| \sum_{s=1}^{t} \gamma^{t-s} \alpha_s A_s A_s^\top (\theta_s^\star - \theta_{t+1}^\star) \right\|_{V_t^{-2}} \leq & \underbrace{\left\| \sum_{s=1}^{t-D} \gamma^{t-s} \alpha_s A_s A_s^\top (\theta_s^\star - \theta_{t+1}^\star) \right\|_{V_t^{-2}}}_{\coloneqq d_1} \\
& + \underbrace{\left\| \sum_{s=t-D+1}^{t} \gamma^{t-s} \alpha_s A_s A_s^\top (\theta_s^\star - \theta_{t+1}^\star) \right\|_{V_t^{-2}}}_{\coloneqq d_2} \ .
\end{aligned}
$$

Both terms are bounded separately; starting with d_1:

$$
\begin{aligned}
d_1 &\leq \lambda^{-1} \left\| \sum_{s=1}^{t-D} \gamma^{t-s} \alpha_s A_s A_s^\top (\theta_s^\star - \theta_{t+1}^\star) \right\|_2 && (V_t \geq \lambda I_d) \\
&\leq \lambda^{-1} \sum_{s=1}^{t-D} \gamma^{t-s} |\alpha_s| \left\| A_s A_s^\top (\theta_s^\star - \theta_{t+1}^\star) \right\|_2 && (\text{Triangle inequality}) \\
&\leq 2k_\mu \lambda^{-1} S L^2 \sum_{s=1}^{t-D} \gamma^{t-s} && (\|A_s\|_2 \leq L,\ \theta_s^\star, \theta_{t+1}^\star \in \Theta, |\alpha_s| \leq k_\mu) \\
&\leq 2k_\mu \lambda^{-1} S L^2 \gamma^D (1-\gamma)^{-1} \ .
\end{aligned}
$$

For d_2 a careful analysis is required.

$$\begin{aligned}
d_2 &= \left\| V_t^{-1} \sum_{s=t-D+1}^{t} \gamma^{t-s}\alpha_s A_s A_s^\top (\theta_s^\star - \theta_{t+1}^\star) \right\|_2 \\
&= \left\| \sum_{s=t-D+1}^{t} V_t^{-1}\gamma^{t-s}\alpha_s A_s A_s^\top (\theta_s^\star - \theta_{t+1}^\star) \right\|_2 \\
&= \left\| \sum_{s=t-D+1}^{t} V_t^{-1}\gamma^{t-s}\alpha_s A_s A_s^\top \sum_{p=s}^{t} \left(\theta_p^\star - \theta_{p+1}^\star\right) \right\|_2 && \text{(Telescopic sum)} \\
&\leq \left\| \sum_{p=t-D+1}^{t} V_t^{-1} \sum_{s=t-D+1}^{p} \gamma^{t-s}\alpha_s A_s A_s^\top \left(\theta_p^\star - \theta_{p+1}^\star\right) \right\|_2 && \text{(Re-arranging)} \\
&\leq \sum_{p=t-D+1}^{t} \left\| V_t^{-1} \sum_{s=t-D+1}^{p} \gamma^{t-s}\alpha_s A_s A_s^\top \left(\theta_p^\star - \theta_{p+1}^\star\right) \right\|_2 && \text{(Triangle inequality)}
\end{aligned}$$

At this point, Assumption 6.4 can be used to upper-bound the operator norm of the matrix $V_t^{-1}\sum_{s=t-D+1}^{p}\gamma^{t-s}\alpha_s A_s A_s^\top$. Under Assumption 6.4, the following holds:

$$V_t^{-1} \sum_{s=t-D+1}^{p} \gamma^{t-s}\alpha_s A_s A_s^\top = V_t^{-1/2} \sum_{s=t-D+1}^{p} \gamma^{t-s}\alpha_s A_s A_s^\top V_t^{-1/2} \coloneqq J_t \tag{6.21}$$

The advantage, now is that the matrix on the right-hand side of Equation (4.10) is symmetric and we can use the relation $\|Mx\| \leq \|M\|\,\|x\|_2$ that holds for all symmetric matrix M and where $\|M\|$ denotes the operator norm of M. The final step consists in upper-bounding the operator norm of J_t. Let x such that $\|x\|_2 \leq 1$, we have,

$$\begin{aligned}
x^\top J_t x &= x^\top V_t^{-1/2} \sum_{s=t-D+1}^{p} \gamma^{t-s}\alpha_s A_s A_s^\top V_t^{-1/2} x = \sum_{s=t-D+1}^{p} \alpha_s x^\top V_t^{-1/2} A_s A_s^\top V_t^{-1/2} x \\
&= \sum_{s=t-D+1}^{p} \gamma^{t-s}\alpha_s \left(A_s^\top V_t^{-1/2} x\right)^2 \leq k_\mu \sum_{s=t-D+1}^{p} \gamma^{t-s} \left(A_s^\top V_t^{-1/2} x\right)^2 \\
&\leq k_\mu x^\top V_t^{-1/2} \sum_{s=t-D+1}^{p} \gamma^{t-s} A_s A_s^\top V_t^{-1/2} x \,.
\end{aligned}$$

Furthermore, by adding some PSD matrices one has:

$$\begin{aligned}
\forall x, \|x\|_2 \leq 1, \quad x^\top V_t^{-1/2} \sum_{s=t-D+1}^{p} \gamma^{t-s} A_s A_s^\top V_t^{-1/2} x &\leq x^\top V_t^{-1/2} \left(\sum_{s=1}^{t} \gamma^{t-s} A_s A_s^\top\right) V_t^{-1/2} x \\
&\leq x^\top x \leq 1 \,.
\end{aligned}$$

Combining the two inequalities ensures that

$$\|J_t\| \leq k_\mu \tag{6.22}$$

Finally,

$$d_2 \leq k_\mu \sum_{p=t-D+1}^{t} \|\theta_p^\star - \theta_{p+1}^\star\|_2 \,.$$

□

6.A.6 Proof of Lemma 6.10

Lemma 6.10. *[Tracking with general action sets] Let $D \in \mathbb{N}^*$. The following holds:*

$$\|g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1})\|_{V_t^{-2}} \leq \frac{2k_\mu L^2 S}{\lambda}\frac{\gamma^D}{1-\gamma} + \frac{k_\mu}{\sqrt{\lambda}}\frac{\sqrt{d}}{\sqrt{1-\gamma}}\sum_{s=t-D+1}^{t}\left\|\theta^\star_s - \theta^\star_{s+1}\right\|_2 .$$

Proof. Following the proof of Lemma 6.5, one has:

$$\|g_t(\bar{\theta}_t) - g_t(\theta^\star_{t+1})\|_{V_t^{-2}} = \left\| V_t^{-1}\sum_{s=1}^{t}\gamma^{t-s}\alpha_s A_s A_s^\top(\theta^\star_s - \theta^\star_{t+1})\right\|_2$$

The remaining part follow the same arguments as Section 4.3.2.4 where there is a need to be cautious with the α_s term. We refer the reader to [Faury et al., 2021a, Lemma 7] for the complete proof. □

Appendix 6.B Regret Bound

6.B.1 Regret decomposition

Lemma 6.6. *The following holds:*

$$R(T) \leq \frac{2k_\mu}{c_\mu}\sum_{t=1}^{T}\beta_{t-1}(\delta)\left[\|A_t\|_{V_{t-1}^{-1}} - \|A_{t,\star}\|_{V_{t-1}^{-1}}\right] + \sum_{t=1}^{T}\left[\Delta_{t-1}(A_t) + \Delta_{t-1}(A_{t,\star})\right].$$

Proof. We recall that $A_{t,\star} = \text{argmax}_{a\in\mathcal{A}_t}\mu(a^\top\theta^\star_t)$. Note that:

$$\begin{aligned}
R(T) &= \sum_{t=1}^{T}\mu(A_{t,\star}^\top\theta^\star_t) - \mu(A_t^\top\theta^\star_t)\\
&= \sum_{t=1}^{T}\mu(A_{t,\star}^\top\theta^\star_t) - \mu(A_{t,\star}^\top\tilde{\theta}_{t-1}) + \mu(A_{t,\star}^\top\tilde{\theta}_{t-1}) - \mu(A_t^\top\tilde{\theta}_{t-1}) + \mu(A_t^\top\tilde{\theta}_{t-1}) - \mu(A_t^\top\theta^\star_t)\\
&= \sum_{t=1}^{T}\left[\mu(A_{t,\star}^\top\tilde{\theta}_{t-1}) - \mu(A_t^\top\tilde{\theta}_{t-1})\right] + \sum_{t=1}^{T}\left[\mu(A_{t,\star}^\top\theta^\star_t) - \mu(A_{t,\star}^\top\tilde{\theta}_{t-1})\right]\\
&\quad + \sum_{t=1}^{T}\left[\mu(A_t^\top\tilde{\theta}_{t-1}) - \mu(A_t^\top\theta^\star_t)\right]\\
&\leq \frac{2k_\mu}{c_\mu}\sum_{t=1}^{T}\beta_{t-1}(\delta)\left[\|A_t\|_{V_{t-1}^{-1}} - \|A_{t,\star}\|_{V_{t-1}^{-1}}\right]\\
&\quad + \sum_{t=1}^{T}\left[\mu(A_{t,\star}^\top\theta^\star_t) - \mu(A_{t,\star}^\top\tilde{\theta}_{t-1})\right] + \sum_{t=1}^{T}\left[\mu(A_t^\top\tilde{\theta}_{t-1}) - \mu(A_t^\top\theta^\star_t)\right] .
\end{aligned}$$

In the last inequality, we used the fact that $A_t = \text{argmax}_{a\in\mathcal{A}}\left\{\mu(a^\top\tilde{\theta}_{t-1}) + \frac{2k_\mu}{c_\mu}\beta_{t-1}(\delta)\|a\|_{V_{t-1}^{-1}}\right\}$. Using the definition of $\Delta_t(a)$ we con-

clude that:

$$R(T) \leq \frac{2k_\mu}{c_\mu} \sum_{t=1}^{T} \beta_t(\delta) \left[\|A_t\|_{V_{t-1}^{-1}} - \|A_{t,\star}\|_{V_{t-1}^{-1}} \right] + \sum_{t=1}^{T} \left[\Delta_{t-1}(A_t) + \Delta_{t-1}(A_{t,\star}) \right] .$$

□

6.B.2 Regret Bound

We now claim Theorem 6.1, bounding the regret of BVD-GLM-UCB.

Theorem 6.1. *Let $\delta \in (0,1]$ and $D \in \mathbb{N}^*$. Under Assumptions 6.1 -6.2-6.3 and Assumption 6.4 with probability at least $1-\delta$:*

$$R(T) \leq C_1 R_\mu \beta_T(\delta) \sqrt{dT} \sqrt{T \log(1/\gamma) + \log\left(1 + \frac{L^2(1-\gamma^T)}{\lambda d(1-\gamma)}\right)} + C_2 R_\mu \frac{\gamma^D}{1-\gamma} T + C_3 R_\mu D B_T$$

Further, setting $\gamma = 1 - (B_T/(dT))^{2/3}$ ensures:

$$R_T = \tilde{\mathcal{O}}\left(R_\mu d^{2/3} B_T^{1/3} T^{2/3}\right) \qquad w.h.p$$

Under general arm-set geometry and Assumptions 6.1-6.2-6.3, with probability at least $1-\delta$

$$\begin{aligned} R(T) \leq\ & C_1 R_\mu \beta_T(\delta) \sqrt{dT} \sqrt{T \log(1/\gamma) + \log\left(1 + \frac{L^2(1-\gamma^T)}{\lambda d(1-\gamma)}\right)} \\ & + C_4 R_\mu \frac{\gamma^D}{1-\gamma} T + C_5 k_\mu R_\mu \frac{\gamma^D}{(1-\gamma)^{3/2}} T + C_6 k_\mu R_\mu \sqrt{\frac{d}{1-\gamma}} D B_T + C_7 k_\mu R_\mu \frac{\sqrt{d}}{1-\gamma} D B_T \end{aligned}$$

Further, setting $\gamma = 1 - \frac{B_T^{2/5}}{d^{1/5} T^{2/5}}$ ensures:

$$R(T) = \tilde{\mathcal{O}}\left(R_\mu d^{9/10} B_T^{1/5} T^{4/5}\right) \qquad w.h.p$$

Proof. In the following, we assume that the event $\{\bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t), \forall t \geq 1\}$ holds, which happens with probability at least $1-\delta$ (Lemma 6.4). Thanks to Lemma 6.8, when Assumption 6.4 the following holds:

$$\begin{aligned} \Delta_t(A_{t+1}) + \frac{2k_\mu}{c_\mu} \beta_t(\delta) \|A_{t+1}\|_{V_t^{-1}} \leq\ & \frac{4k_\mu}{c_\mu} \beta_t(\delta) \|A_{t+1}\|_{V_t^{-1}} + \frac{4k_\mu^2 L^3 S}{c_\mu \lambda (1-\gamma)} \gamma^D \\ & + \frac{2k_\mu^2 L}{c_\mu} \sum_{s=t-D+1}^{t} \left\|\theta_s^\star - \theta_{s+1}^\star\right\|_2 . \end{aligned}$$

$$\Delta_t(A_{t+1,\star}) - \frac{2k_\mu}{c_\mu} \beta_t(\delta) \|A_{t+1,\star}\|_{V_t^{-1}} \leq \frac{4k_\mu^2 L^3 S}{c_\mu \lambda (1-\gamma)} \gamma^D + \frac{2k_\mu^2 L}{c_\mu} \sum_{s=t-D+1}^{t} \left\|\theta_s^\star - \theta_{s+1}^\star\right\|_2 .$$

Assembling this result with Lemma 6.6 yields:

$$R(T) \leq \underbrace{\sum_{t=1}^{T} \frac{4k_\mu}{c_\mu} \beta_{t-1}(\delta) \|A_t\|_{V_{t-1}^{-1}}}_{R_T^{\text{learn}}} + \underbrace{\sum_{t=1}^{T} \left[\frac{8k_\mu^2 L^3 S}{c_\mu \lambda (1-\gamma)} \gamma^D + \frac{4k_\mu^2 L}{c_\mu} \sum_{s=t-D}^{t-1} \left\|\theta_s^\star - \theta_{s+1}^\star\right\|_2 \right]}_{R_T^{\text{track}}} .$$

We now bound each term separately. Starting with R_T^{learn}:

$$
\begin{aligned}
R_T^{\text{learn}} &\leq \frac{4k_\mu}{c_\mu}\beta_T(\delta)\sum_{t=1}^{T}\|A_t\|_{V_{t-1}^{-1}} && (t \to \beta_t(\delta) \text{ increasing}) \\
&\leq \frac{4k_\mu}{c_\mu}\beta_T(\delta)\sqrt{T}\sqrt{\sum_{t=1}^{T}\|A_t\|^2_{V_{t-1}^{-1}}} && \text{(Cauchy-Schwarz)} \\
&\leq \frac{4k_\mu}{c_\mu}\beta_T(\delta)\sqrt{2T\max(1,L^2/\lambda)}\sqrt{dT\log(1/\gamma)+\log\left(\frac{\det V_T}{\lambda^d}\right)} && \text{(Lemma 5.30)} \\
&\leq \frac{4k_\mu}{c_\mu}\beta_T(\delta)\sqrt{2dT\max(1,L^2/\lambda)}\sqrt{T\log(1/\gamma)+\log\left(1+\frac{L^2(1-\gamma^T)}{\lambda d(1-\gamma)}\right)}\,. && \text{(Lemma 5.28)}
\end{aligned}
$$

The bounding of the tracking term is straight-forward:

$$
\begin{aligned}
R_T^{\text{track}} &= \frac{8k_\mu^2L^3S}{c_\mu\lambda(1-\gamma)}\gamma^D T + \frac{4k_\mu^2L}{c_\mu}\sum_{t=1}^{T}\sum_{s=t-D}^{t-1}\left\|\theta_s^\star-\theta_{s+1}^\star\right\|_2 \\
&\leq \frac{8k_\mu^2L^3S}{c_\mu\lambda(1-\gamma)}\gamma^D T + \frac{4k_\mu^2L}{c_\mu}DB_T\,.
\end{aligned}
$$

Assembling this two bounds (R_T^{learn} and R_T^{track}) yields the first announced result, with the following constants:

$$
\begin{aligned}
C_1 &= \sqrt{32\max(1,L^2/\lambda)}\,. \\
C_2 &= \frac{8k_\mu L^3 S}{\lambda}\,. \\
C_3 &= 4k_\mu L\,.
\end{aligned}
$$

The last part of the proof follows the asymptotic argument presented in Chapter 4. We assume that B_T is sub-linear and let:

$$
D = \frac{\log T}{1-\gamma}\,, \qquad \gamma = 1-\left(\frac{B_T}{dT}\right)^{2/3}\,.
$$

We therefore have the following asymptotic equivalences (omitting logarithmic dependencies):

$$
\begin{aligned}
\beta_T(\delta)\sqrt{dT}\sqrt{T\log(1/\gamma)} \sim dT\cdot\left(\frac{B_T}{dT}\right)^{1/3} &= d^{2/3}B_T^{1/3}T^{2/3} \\
\gamma^D T/(1-\gamma) \sim \exp(-\log T)T\left(\frac{B_T}{dT}\right)^{-2/3} &= d^{2/3}B_T^{-2/3}T^{2/3} \\
DB_T \sim B_T\left(\frac{B_T}{dT}\right)^{-2/3} &= d^{2/3}B_T^{1/3}T^{2/3}
\end{aligned}
$$

Merged with the regret-bound we just proved, this yields the announced result.

Without Assumption 6.4 similar results can be obtained. The main difference consists in using the upper-bound for the tracking term under general arm-set geometry which

is slightly more complicated. Plugging the bound from Lemma 6.8 and upper bounding $\sqrt{1+\frac{L^2}{\lambda(1-\gamma)}}$ by $1+\frac{L}{\sqrt{\lambda(1-\gamma)}}$ gives the announced regret decomposition. Let:

$$\gamma = 1 - \frac{B_T^{2/5}}{d^{1/5}T^{2/5}}$$

$$\begin{aligned}
\beta_T(\delta)\sqrt{dT}\sqrt{T\log(1/\gamma)} \sim dT \cdot \frac{B_T^{1/5}d^{-1/10}}{T^{1/5}} &= d^{9/10}B_T^{1/5}T^{4/5} \\
\gamma^D T/(1-\gamma)^{3/2} \sim \exp(-\log T)T\left(\frac{d^{1/5}T^{2/5}}{B_T^{2/5}}\right)^{3/2} &= d^{3/10}B_T^{-3/5}T^{3/5} \\
\frac{\sqrt{d}}{1-\gamma}DB_T \sim d^{1/2}B_T\left(\frac{d^{1/5}T^{2/5}}{B_T^{2/5}}\right)^{2} &= d^{9/10}B_T^{1/5}T^{4/5}
\end{aligned}$$

□

Appendix 6.C On the Projection Step

6.C.1 Equivalent Minimization Program

Recall the original minimization program for finding θ_t^p:

$$\theta_t^p \in \operatorname{argmin}_{\theta\in\mathbb{R}^d}\left\{\left\|g_t(\theta) - g_t(\hat{\theta}_t)\right\|_{V_t^{-2}} \text{ s.t } \Theta \cap \mathcal{E}_t^\delta(\theta) \neq \emptyset\right\}. \tag{P1}$$

Note that this minimum exists (0_d is feasible) and is indeed attained (the feasible set is compact and the objective smooth). The following reformulation is motivated by the fact that only $\tilde{\theta}_t \in \Theta \cap \mathcal{E}_t^\delta(\theta_t^p)$ is needed for the algorithm. To this end, we explicitly introduce $\tilde{\theta}_t$ in the program via a slack variable. Formally, we study:

$$\begin{pmatrix}\tilde{\theta}_t \\ \theta_t^p\end{pmatrix} \in \operatorname{argmin}_{\theta'\in\mathbb{R}^d,\theta\in\mathbb{R}^d}\left\{\left\|g_t(\theta) - g_t(\hat{\theta}_t)\right\|_{V_t^{-2}} \text{ s.t } \theta' \in \mathcal{E}_t^\delta(\theta) \cap \Theta\right\}. \tag{P1'}$$

We also introduce the following program:

$$\begin{pmatrix}\tilde{\theta}_t \\ \eta\end{pmatrix} \in \operatorname{argmin}_{\theta'\in\mathbb{R}^d,\eta\in\mathbb{R}^d}\left\{\left\|g_t(\theta') + \beta_t(\delta)\widetilde{V}_t^{1/2}\eta - g_t(\hat{\theta}_t)\right\|_{V_t^{-2}} \text{ s.t } \left\|\theta'\right\|_2 \leq S, \|\eta\|_2 \leq 1\right\}. \tag{P2}$$

We claim and prove the following result, which is an equivalent reformulation of Proposition 6.2.

Proposition 6.11. *The programs Equation (P1') and Equation (P2) are equivalent.*

Proof. The proof consists in building a bijection between the solutions of Equation (P1')

and Equation (**P2**). Let us introduce the mapping:

$$\begin{aligned} f : \Theta \times \mathbb{R}^d &\to \Theta \times \mathbb{R}^d \\ \begin{pmatrix} x \\ y \end{pmatrix} &\to \begin{pmatrix} f_1(x) \\ f_2(x,y) \end{pmatrix} = \begin{pmatrix} x \\ \beta_t^{-1}(\delta)\widetilde{V}_t^{-1/2}(g_t(y) - g_t(x)) \end{pmatrix} \end{aligned}$$

We now claim the following Lemma, which proof is deferred to Section 6.C.2.

Lemma 6.12. *The function:*

$$\begin{aligned} g_t : \mathbb{R}^d &\to \mathbb{R}^d \\ \theta &\to \sum_{s=1}^{t-1} \gamma^{t-s} \mu(A_s^\top \theta) A_s + \lambda c_\mu \theta \end{aligned}$$

is a bijection.

A straight-forward implication of this Lemma is the bijectivity of f. Let $(\tilde{\theta}^1, \theta^p)$ be a solution of Equation (**P1'**) and let:

$$\begin{pmatrix} \tilde{\theta}^2 \\ \eta^p \end{pmatrix} = f \begin{pmatrix} \tilde{\theta}^1 \\ \theta^p \end{pmatrix} .$$

We are going to show that $(\tilde{\theta}^2, \eta^p)$ is a solution of Equation (**P2**). Because $(\tilde{\theta}^1, \theta^p)$ is optimal for Equation (**P1'**), we have that:

$$\begin{aligned} &\|g_t(\theta^p) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \leq \|g_t(\theta) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \\ &\qquad\qquad \forall (\theta', \theta) \in \Theta \times \mathbb{R}^d \text{ s.t } \theta' \in \mathcal{E}_t^\delta(\theta) \\ \Leftrightarrow\ &\|g_t(\theta^p) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \leq \|g_t(\theta) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \qquad \text{(def. of } \mathcal{E}_t^\delta(\theta)) \\ &\qquad\qquad \forall (\theta', \theta) \in \Theta \times \mathbb{R}^d \text{ s.t } \|g_t(\theta') - g_t(\theta)\|_{\widetilde{V}_t^{-1}} \leq \beta_t(\delta) \\ \Leftrightarrow\ &\|g_t(\theta^p) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \leq \|g_t(\theta) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \\ &\qquad\qquad \forall (\theta', \theta) \in \Theta \times \mathbb{R}^d \text{ s.t } \|f_2(\theta', \theta)\|_2 \leq 1 \end{aligned}$$

Noticing that for all $(x, y) \in \Theta \times \mathbb{R}^d$ we have $g_t(y) = g_t(x) + \beta_t(\delta) V_t^{1/2} f_2(x, y)$ we therefore obtain:

$$
\begin{aligned}
&\|g_t(\tilde{\theta}^1) + \beta_t(\delta)\widetilde{V}_t^{1/2} f_2(\tilde{\theta}^1, \theta^p) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \leq \|g_t(\theta') + \beta_t(\delta)\widetilde{V}_t^{1/2} f_2(\theta', \theta) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \\
&\qquad\qquad \forall (\theta', \theta) \in \Theta \times \mathbb{R}^d \text{ s.t } \left\|f_2(\theta', \theta)\right\|_2 \leq 1 \\
\Leftrightarrow\ &\|g_t(\tilde{\theta}^1) + \beta_t(\delta)\widetilde{V}_t^{1/2} \eta^p - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \leq \|g_t(\theta') + \beta_t(\delta)\widetilde{V}_t^{1/2} f_2(\theta', \theta) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \\
&\qquad\qquad \forall (\theta', \theta) \in \Theta \times \mathbb{R}^d \text{ s.t } \left\|f_2(\theta', \theta)\right\|_2 \leq 1 \\
\Leftrightarrow\ &\|g_t(\tilde{\theta}^2) + \beta_t(\delta)\widetilde{V}_t^{1/2} \eta^p - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \leq \|g_t(\theta') + \beta_t(\delta)\widetilde{V}_t^{1/2} f_2(\theta', \theta) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \\
&\qquad\qquad \forall (\theta', \theta) \in \Theta \times \mathbb{R}^d \text{ s.t } \left\|f_2(\theta', \theta)\right\|_2 \leq 1 \qquad (\tilde{\theta}^1 = \tilde{\theta}^2) \\
\Leftrightarrow\ &\|g_t(\tilde{\theta}^2) + \beta_t(\delta)\widetilde{V}_t^{1/2} \eta^p - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \leq \|g_t(\theta') + \beta_t(\delta)\widetilde{V}_t^{1/2} f_2(\theta', \theta) - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \\
&\qquad\qquad \forall (\theta', \theta) \text{ s.t } \left\|f_2(\theta', \theta)\right\|_2 \leq 1, \|\theta'\|_2 \leq S \\
\Leftrightarrow\ &\|g_t(\tilde{\theta}^2) + \beta_t(\delta)\widetilde{V}_t^{1/2} \eta^p - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \leq \|g_t(\theta') + \beta_t(\delta)\widetilde{V}_t^{1/2} \eta - g_t(\hat{\theta}_t)\|_{V_t^{-2}} \\
&\qquad\qquad \forall (\theta', \eta) \text{ s.t } \|\eta\|_2 \leq 1, \|\theta'\|_2 \leq S
\end{aligned}
$$

where we last used the fact that f_2 spans $\mathbb{R}^d$ (surjectivity). Finally, we have that:

$$
\begin{aligned}
&\left\|\tilde{\theta}^2\right\|_2 \leq S && (\tilde{\theta}^2 = \tilde{\theta}^1 \in \Theta) \\
&\|\eta^p\|_2 = \beta_t^{-1}(\delta) \left\|g_t(\theta^p) - g_t(\tilde{\theta}^1)\right\|_{V_t^{-1}} \leq 1 && (\tilde{\theta}^1 \in \mathcal{E}_t^\delta(\theta^p))
\end{aligned}
$$

Combining the last two results proves that $(\tilde{\theta}^2, \eta^p)$ is feasible for Equation (**P2**), and optimal within the feasible set. As a consequence, $(\tilde{\theta}^2, \eta^p)$ is a solution of Equation (**P2**). Therefore, f is a bijection between the minimizers of Equation (**P1'**) and Equation (**P2**), which concludes the proof. □

6.C.2 Bijectivity of g_t

Lemma 6.12. *The function:*

$$
\begin{aligned}
g_t : &\mathbb{R}^d \to \mathbb{R}^d \\
&\theta \to \sum_{s=1}^{t-1} \gamma^{t-s} \mu(A_s^\top \theta) A_s + \lambda c_\mu \theta
\end{aligned}
$$

is a bijection.

Proof. Injectivity. Notice that $\forall \theta \in \mathbb{R}^d$:

$$
\nabla_\theta g(\theta) = \sum_{s=1}^{t} \gamma^{t-s} \dot{\mu}(A_s^\top \theta) A_s A_s^\top + \lambda c_\mu \mathbf{I_d} \succ 0 \,.
$$

Hence $\nabla_\theta g$ is P.S.D, and a simple integral Taylor expansion is enough to prove injectivity. Surjectivity Let $z \in \mathbb{R}^d$. Let $A = \text{Span}(A_1, .., A_t)$ be the vectorial space spanned by $\{A_s\}_{s=1}^t$. Let $z_\perp$ be the orthogonal projection of z on A and $z_\parallel = z - z_\perp$. Since $z_\perp \in A$, there exists

$\{\alpha_s\}_{s=1}^t \in \mathbb{R}^t$ such that:

$$z_\perp = \sum_{s=1}^{t} \alpha_s A_s \ .$$

Recall that $b(\cdot)$ is a primitive of μ, which is convex since μ is strictly increasing. Define:

$$L(\theta) = \sum_{s=1}^{t} \gamma^{t-s} \left[b(A_s^\top \theta) - \frac{\alpha_s}{\gamma^{t-s}} A_s^\top \theta \right] + \frac{\lambda c_\mu}{2} \left\| \theta - \frac{z_\parallel}{\lambda c_\mu} \right\|^2 \ .$$

which is a strictly convex, coercive function. Its minimum θ_z (which therefore exists and is uniquely defined) checks:

$$\begin{aligned}
&\nabla_\theta L(\theta_z) = 0 \\
&\Leftrightarrow \sum_{s=1}^{t} \gamma^{t-s} \left[\mu(A_s^\top \theta_z) - \frac{\alpha_s}{\gamma^{t-s}} \right] A_s + \lambda c_\mu \left(\theta_z - \frac{z_\parallel}{\lambda c_\mu} \right) = 0 \\
&\Leftrightarrow g(\theta_z) = \sum_{s=1}^{t} \alpha_s x_s + z_\parallel \\
&\Leftrightarrow g(\theta_z) = z_\perp + z_\parallel = z \ .
\end{aligned}$$

which proves surjectivity. □

Appendix 6.D BVD-GLM-UCB Algorithm

6.D.1 High-level ideas

In this part of the appendix, we denote $\gamma^\star$ as follows:

$$\gamma^\star = 1 - \frac{1}{2} \left(\frac{\mathcal{B}_T}{dT2S} \right)^{2/3} \ . \tag{6.23}$$

Remark 6.5. *$\gamma^\star$ as defined in Equation (6.23) has a different expression than the discount factor proposed in Theorem 6.1. This slight modification is to ensure that $\gamma^\star$ is larger than 1/2 and simplifies the finite time analysis of the regret. Yet, it has no consequence on the asymptotic bound.*

$\mathcal{B}_T$ being unknown, we cannot compute the optimal discount factor that depends on the parameter drift. The general idea is to use a set of different values for the discount factor (respectively the $\mathcal{B}_T$ values) called $\mathcal{H}$, covering the $[1/2, 1)$ space (respectively the $[0, 2ST)$ space). Then, we divide the time horizon T into different blocks of length H. Every H steps, we create **a new instance** of BVD-GLM-UCB with a γ that is chosen by a *master* algorithm: the EXP3 algorithm from [Auer et al., 2002b]. At the end of each block, this *master* algorithm receives the cumulative rewards from the instantiated *worker* and updates its probability distribution over the set $\mathcal{H}$. The objective of the master algorithm is to learn the most suitable value of γ so as to maximise the cumulative rewards in accordance with the dynamics of the environment. On the other side, the different *workers* algorithms act exactly as if the BVD-GLM-UCB algorithm was launched on a H-steps experiment. This setting is similar to the one presented in [Cheung et al., 2021] (respectively [Zhao et al., 2020]) with discount factors instead of sliding

windows (respectively restart parameters). This framework is called Bandit-over-Bandit (BOB) precisely because of this two-stage structure between the *master* and the *workers* algorithms.

6.D.2 Algorithm

The coverage $\mathcal{H}$ with the different discount factors is defined in the following way:

$$\mathcal{H} = \{\gamma_j = 1 - \mu_j | j = 1, \dots, N\} \tag{6.24}$$

$$\text{with } N = \left\lceil \frac{2}{3} \log_2 \left(2ST^{3/2}\right) \right\rceil + 1 \text{ and } \mu_j = \frac{1}{2} \frac{2^{j-1}}{d^{2/3} T (2S)^{2/3}} \,. \tag{6.25}$$

The *main* algorithm is an instance of the EXP3 algorithm from [Auer et al., 2002b] where the different arms correspond to the different discount factors. Following EXP3 analysis [Auer et al., 2002b], the probability of drawing γ_j for the block i is

$$p_i^{\gamma_j} = (1-\alpha) \frac{s_i^{\gamma_j}}{\sum_j s_i^{\gamma_j}} + \frac{\alpha}{N}, \ \forall j = 1, 2, \dots, N \,, \tag{6.26}$$

where α is defined as

$$\alpha = \min \left\{ 1, \sqrt{\frac{N \log(N)}{(e-1)\lceil T/H \rceil}} \right\} \tag{6.27}$$

and $s_i^{\gamma_j}$ is initialised at 1 and is updated at the end of each block **when selected** with

$$s_{i+1}^{\gamma_j} = s_i^{\gamma_j} \exp \left(\frac{\alpha}{N p_i^{\gamma_j}} \frac{\sum_{t=(i-1)H+1}^{\min\{iH,T\}} X_t}{mH} \right) . \tag{6.28}$$

Note that in Equation (6.28), X_t is the noisy reward obtained when the action A_t is selected with the BVD-GLM-UCB algorithm with parameter γ_j. Equation (6.26), Equation (6.27) and Equation (6.28) are the same as in [Auer et al., 2002b] except for the rescaling of the cumulative rewards on a block that is required to ensure that they lie in $[0, 1]$. Details on this rescaling part can be found in Proposition 6.15.

Input: Length H, time horizon T, regularization λ, confidence δ, inverse link function μ, constants S, L and m
Initialization: Create the covering space $\mathcal{H}$ as defined in Equation (6.24), set $s_1^{\gamma_i} = 1$, $\forall \gamma_i \in \mathcal{H}$.
for $i = 1, \dots, \lceil T/H \rceil$ **do**
  $\gamma_j \sim p_i^{\gamma}$, the probability vector defined in Equation (6.26).
  Start a BVD-GLM-UCB subroutine with parameter γ_j
  for $t = (i-1)H + 1, \dots, \min\{iH, T\}$ **do**
    Receive the action set $\mathcal{A}_t$.
    Select $A_t(\gamma_j) \in \mathcal{A}_t$ with BVD-GLM-UCB.
    Observe reward X_t.
  Update $s_{i+1}^{\gamma_j}$ according to Equation (6.28).
  Update $s_{i+1}^{\gamma} = s_i^{\gamma}$, $\forall \gamma \neq \gamma_j$.

Algorithm 15: BOB-BVD-GLM-UCB (detailed)

Remark 6.6. *We denote $A_t(\gamma)$ the action chosen with the BVD-GLM-UCB algorithm with a discount factor γ.*

6.D.3 Regret Guarantees

In this section, we give an upper-bound for the expected dynamic regret of BOB-BVD-GLM-UCB. By construction, it is natural to decompose the regret into two sources of errors. First the *master* error committed by the EXP3 algorithm by not choosing the best possible discount factor. Second the *worker* error inherent to the BVD-GLM-UCB algorithm. Note that there are two independent sources of randomness: the stochasticity of the rewards (whose expectation is denoted $\mathbb{E}_N$) and the randomness of the EXP3 algorithm (denoted $\mathbb{E}_{\text{EXP3}}$). Bringing things together,

$$\begin{aligned}\mathbb{E}[R(T)] &= \mathbb{E}_N\left[\sum_{t=1}^{T}\mu(A_{t,\star}^\top\theta_t^\star) - \mathbb{E}_{\text{EXP3}}[X_t]\right] \\ &= \underbrace{\mathbb{E}_N\left[\sum_{t=1}^{T}\mu(A_{t,\star}^\top\theta_t^\star) - \sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{\min\{iH,T\}}\mu(A_t(\widehat{\gamma})^\top\theta_t^\star)\right]}_{worker} \\ &+ \underbrace{\mathbb{E}_N\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{\min\{iH,T\}}\mu(A_t(\widehat{\gamma})^\top\theta_t^\star) - \mathbb{E}_{\text{EXP3}}\left[X_t\right]\right]}_{master}.\end{aligned} \tag{6.29}$$

The next step consists in upper-bounding the *worker* error and the *master* error from Equation (6.29) respectively.

Lemma 6.13. *With pavement $\mathcal{H}$ defined in Equation (6.24) for any unknown $\mathcal{B}_T > 0$, setting $k = \lfloor \frac{2}{3}\log_2(\mathcal{B}_T T^{1/2})\rfloor + 1$ yields*

$$\gamma_{k+1} \le \gamma^\star \le \gamma_k\,.$$

Proof. With assumption 6.1, we have $\mathcal{B}_T \le 2ST$. Using this, k (as defined in the statement of the lemma) is smaller than N. We have,

$$\begin{aligned}&k-1 \le \frac{2}{3}\log_2(\mathcal{B}_T T^{1/2}) \le k \\ \Leftrightarrow\; &-\frac{1}{2}\frac{2^{k-1}}{d^{2/3}T(2S)^{2/3}} \ge -\frac{1}{2}\left(\frac{\mathcal{B}_T}{dT2S}\right)^{2/3} \ge -\frac{1}{2}\frac{2^k}{d^{2/3}T(2S)^{2/3}}\,.\end{aligned}$$

Adding one for the different terms gives the result. □

For the rest of the section, we set $\widehat{\gamma} = \gamma_k$ with k defined in Lemma 6.13. We denote $\mathcal{B}_i = \sum_{t=(i-1)H+1}^{iH-1}\|\theta_{t+1}^\star - \theta_t^\star\|_2$ and

$$\beta_H^\star = \sqrt{\lambda}S + \frac{m}{2}\sqrt{2\log(T) + d\log\left(1 + \frac{2L^2}{\lambda d(1-\gamma^{\star 2})}\right)}\,. \tag{6.30}$$

Proposition 6.14. *The worker error can be upper-bounded in the following way:*

$$worker \leq m\frac{T}{H} + C_1 R_\mu \beta_H^\star \sqrt{dT}\sqrt{2T(1-\gamma^\star) + \frac{T}{H}\log\left(1+\frac{2L^2}{d\lambda(1-\gamma^\star)}\right)} + 2C_2 R_\mu \frac{1}{\sqrt{T}}\frac{1}{1-\gamma^\star} + \frac{3C_3 R_\mu}{\log(2)}\frac{\mathcal{B}_T \log(T)}{1-\gamma^\star} ,$$

with C_1, C_2, C_3 constant terms from Theorem 6.1 and $\beta_H^\star$ defined in Equation (6.30).

Proof. First, note that our objective here is to bound the expected regret whereas Theorem 6.1 bounds the regret and gives a high probability upper-bound. We denote $E_\delta^i = \{\bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t)$ for t s.t $(i-1)H+1 \leq t \leq \min\{iH, T\}\}$. This event holds with probability higher than $1-\delta$. When E_δ^i does not hold, the maximum regret could theoretically be suffered for all time instants.

As explained in the algorithm mechanism, a new instance of BVD-GLM-UCB will be launched every H steps with a discount factor selected by the EXP3 algorithm. Restarting a new algorithm and forgetting previous information comes at a cost in terms of regret. This is made explicit in the following decomposition of *worker*.

$$\begin{aligned} worker &= \mathbb{E}_N\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{\min\{iH,T\}} \mu(A_{t,\star}^\top\theta_t^\star) - \mu(A_t(\widehat{\gamma})^\top\theta_t^\star)\right] \\ &= \underbrace{\mathbb{E}_N\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{\min\{iH,T\}} \mu(A_{t,\star}^\top\theta_t^\star) - \mu(A_t(\widehat{\gamma})^\top\theta_t^\star)\Big|\{\cap_{i=1}^{\lceil T/H\rceil}E_\delta^i\}\right]\mathbb{P}\left(\cap_{i=1}^{\lceil T/H\rceil}E_\delta^i\right)}_{worker_1} \\ &+ \underbrace{\mathbb{E}_N\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{\min\{iH,T\}} \mu(A_{t,\star}^\top\theta_t^\star) - \mu(A_t(\widehat{\gamma})^\top\theta_t^\star)\Big|\{\cap_{i=1}^{\lceil T/H\rceil}E_\delta^i\}^c\right]\mathbb{P}\left(\{\cap_{i=1}^{\lceil T/H\rceil}E_\delta^i\}^c\right)}_{worker_2} \end{aligned}$$

Thanks to Lemma 6.4, E_δ^i holds with probability higher than $1-\delta$. By setting $\delta = 1/T$, we have

$$\mathbb{P}\left(\cup_{i=1}^{\lceil T/H\rceil}(E_\delta^i)^c\right) \leq \lceil T/H\rceil 1/T . \tag{6.31}$$

Under the event $\{\cup_{i=1}^{\lceil T/H\rceil}(E_\delta^i)^c\}$ not much can be said. The maximum regret $r_{\max} = m$ can be suffered at every time step. Therefore, using the upper-bound from Equation (6.31), we obtain

$$\begin{aligned} worker_2 &= \mathbb{E}_N\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{\min\{iH,T\}} \mu(A_{t,\star}^\top\theta_t^\star) - \mu(A_t(\widehat{\gamma})^\top\theta_t^\star)\Big|\{\cup_{i=1}^{\lceil T/H\rceil}(E_\delta^i)^c\}\right]\mathbb{P}\left(\cup_{i=1}^{\lceil T/H\rceil}(E_\delta^i)^c\right) \\ &\leq r_{\max}\lceil T/H\rceil . \end{aligned}$$

This term is related to the number of restarts of the algorithm. In the BOB framework, whatever the worker algorithm (sliding window, restart factor) a cost of order T/H will be paid due to the restarting of the *worker* at the beginning of each block.
On the contrary, under the event $\{\cap_{i=1}^{\lceil T/H\rceil} E_\delta^i\}$, using the assumption that the blocks are independent, we can follow the line of proof from Lemma 6.6 and Theorem 6.1 for every block. We introduce,

$$\beta_H = \sqrt{\lambda} S + \frac{m}{2}\sqrt{2\log(T) + d\log\left(1 + \frac{L^2(1-\gamma_k^{2H})}{\lambda d(1-\gamma_k^2)}\right)} \,. \tag{6.32}$$

$$\begin{aligned}
worker_1 &= \mathbb{E}_N\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{\min\{iH,T\}} \mu(A_{t,\star}^\top\theta_t^\star) - \mu(A_t(\widehat{\gamma})^\top\theta_t^\star)\Big|\{\cap_{i=1}^{\lceil T/H\rceil}E_\delta^i\}\right]\mathbb{P}\left(\cap_{i=1}^{\lceil T/H\rceil}E_\delta^i\right)\\
&\le \mathbb{E}_N\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{\min\{iH,T\}} \mu(A_{t,\star}^\top\theta_t^\star) - \mu(A_t(\widehat{\gamma})^\top\theta_t^\star)\Big|\{\cap_{i=1}^{\lceil T/H\rceil}E_\delta^i\}\right]\\
&\le \sum_{i=1}^{\lceil T/H\rceil}\left(C_1\beta_H\sqrt{dH}\sqrt{H\log(1/\widehat{\gamma}) + \log\left(1+\frac{L^2}{d\lambda(1-\widehat{\gamma})}\right)} + C_2\frac{\widehat{\gamma}^D}{1-\widehat{\gamma}}H + C_3\mathcal{B}_i D\right)\\
&\le C_1\beta_H\sqrt{dT}\sqrt{T\log(1/\widehat{\gamma}) + \frac{T}{H}\log\left(1+\frac{L^2}{d\lambda(1-\widehat{\gamma})}\right)} + C_2\frac{\widehat{\gamma}^D}{1-\widehat{\gamma}}T + C_3\mathcal{B}_T D \,,
\end{aligned}$$

where the second inequality is a consequence of Theorem 6.1. We set,

$$D = \frac{3/2\log(T)}{\log(1/\widehat{\gamma})} \,. \tag{6.33}$$

Hence,

$$\begin{aligned}
C_3\mathcal{B}_T D &\le \frac{3}{2}\frac{C_3\mathcal{B}_T\log(T)}{\log(1/\widehat{\gamma})}\\
&\le \frac{3C_3}{2\log(2)}\mathcal{B}_T\log(T)\frac{\widehat{\gamma}}{1-\widehat{\gamma}} \quad (\text{Using } \log(x)\ge\log(2)(x-1) \text{ for } x\in[1,2])\\
&\le \frac{3C_3}{2\log(2)}\frac{\mathcal{B}_T\log(T)}{1-\gamma_k} \quad (\widehat{\gamma}\le 1)\\
&\le \frac{3C_3}{\log(2)}\frac{\mathcal{B}_T\log(T)}{1-\gamma_{k+1}} \quad (\text{Definition of } \mathcal{H})\\
&\le \frac{3C_3}{\log(2)}\frac{\mathcal{B}_T\log(T)}{1-\gamma^\star} \quad (\text{Lemma 6.13}) \,.
\end{aligned}$$

We also have,

$$\begin{aligned}
C_2\frac{\widehat{\gamma}^D}{1-\widehat{\gamma}}T &\le C_2\frac{1}{\sqrt{T}}\frac{1}{1-\widehat{\gamma}} \quad (\text{Equation (6.33)})\\
&\le 2C_2\frac{1}{\sqrt{T}}\frac{2}{1-\gamma_{k+1}} \quad (\text{Definition of } \mathcal{H})\\
&\le 2C_2\frac{1}{\sqrt{T}}\frac{1}{1-\gamma^\star} \quad (\text{Lemma 6.13}) \,.
\end{aligned}$$

Finally, using $x \mapsto \log(x) \leq x - 1$ for $x > 1$ and Lemma 6.13, one has:

$$T\log(1/\widehat{\gamma}) + \frac{T}{H}\log\left(1 + \frac{L^2}{d\lambda(1-\widehat{\gamma})}\right) \leq T\frac{1-\widehat{\gamma}}{\widehat{\gamma}} + \frac{T}{H}\log\left(1 + \frac{2L^2}{d\lambda(1-\gamma^\star)}\right)$$
$$\leq 2T(1-\gamma^\star) + \frac{T}{H}\log\left(1 + \frac{2L^2}{d\lambda(1-\gamma^\star)}\right) .$$

Following similar steps, we can upper-bound β_H from Equation (6.32) by

$$\beta_H \leq \beta_H^\star .$$

Bringing things together, we have shown that under the event $\{\cap_{i=1}^{\lceil T/H\rceil}\mathcal{E}_i\}$ all the terms depending on $\widehat{\gamma}$ can be replaced by terms depending only on $\gamma^\star$ at the cost of multiplicative constant independent of T. Finally, one has

$$worker \leq m\frac{T}{H} + C_1 R_\mu \beta_H^\star \sqrt{dT}\sqrt{2T(1-\gamma^\star) + \frac{T}{H}\log\left(1 + \frac{2L^2}{d\lambda(1-\gamma^\star)}\right)}$$
$$+ \; 2C_2 R_\mu \frac{1}{\sqrt{T}}\frac{1}{1-\gamma^\star} + \frac{3C_3 R_\mu}{\log(2)}\frac{\mathcal{B}_T\log(T)}{1-\gamma^\star} .$$

□

The above proposition bounds the regret incurred if the same discount factor $\widehat{\gamma}$ is used for each block. To successfully upper bound BVD-GLM-UCB's regret, we need to upper bound the second part *master* which is the error due to the use of the EXP3 algorithm. This part can be controlled thanks to the analysis proposed in [Auer et al., 2002b]. Yet, two issues need to be overcome. (1) The rewards received at the end of a block does not lie in $[0, 1]$ which is required to use the result from [Auer et al., 2002b]. (2) We are in a stochastic environment with noisy rewards.

In the next proposition, we upper-bound the term of interest and explain how to deal with the two issues. The big picture is the following: using the assumption on the bounded rewards we can obtain an upper-bound for the maximum reward on a single block.

Proposition 6.15. *The regret due to the master algorithm can be bounded in the following way,*

$$\mathbb{E}_N\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{\min\{iH,T\}}\mu(A_t(\widehat{\gamma})^\top\theta_t^\star) - \mathbb{E}_{\text{EXP3}}\left[X_t\right]\right] \leq 2mH\sqrt{e-1}\sqrt{\frac{T}{H}\text{card}(\mathcal{H})\log(\text{card}(\mathcal{H}))} .$$

Proof. We denote γ_i the discount factor chosen by the EXP3 algorithm in the i-th block. The regret due to the use of the EXP3 *main* algorithm can be written as follows:

$$master = \mathbb{E}_N\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{\min\{iH,T\}}\mu(A_t(\widehat{\gamma})^\top\theta_t^\star) - \mathbb{E}_{\text{EXP3}}\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{\min\{iH,T\}}X_t\right]\right] .$$

We introduce $Q_i(\gamma_j) = \sum_{t=(i-1)H+1}^{\min\{iH,T\}} X_t(\gamma_j) = \sum_{t=(i-1)H+1}^{\min\{iH,T\}}\mu(A_t(\gamma_j)^\top\theta_t^\star) + \epsilon_t$, using Equation (6.9). This quantity corresponds to the reward obtained on the i-th block when using BVD-GLM-UCB with the discount factor γ_j. We also use $Q_i = \max_{\gamma\in\mathcal{H}} Q_i(\gamma)$.

Contrarily to existing works in the linear setting (e.g [Cheung et al., 2019, Lemma3]) our assumption on the bounded rewards is sufficient to solve both problems. We have, $|Q_i| \leq mH$ almost surely using $r_t \leq m$ for all time instants.

Let $\mathcal{U} = \{\forall t \leq T, 0 \leq r_t \leq m\}$. Thanks to Assumption 6.2, we have $\mathbb{P}(\mathcal{U}) = 1$.
One has,

$$\begin{aligned} master &\leq \mathbb{E}_N \left[\sum_{i=1}^{\lceil T/H \rceil} Q_i(\gamma_k) - \max_{\gamma \in \mathcal{H}} \sum_{i=1}^{\lceil T/H \rceil} Q_i(\gamma) + \max_{\gamma \in \mathcal{H}} \sum_{i=1}^{\lceil T/H \rceil} Q_i(\gamma) - \mathbb{E}_{\text{EXP3}} \left[\sum_{i=1}^{\lceil T/H \rceil} Q_i(\gamma_i) \right] \right] \\ &\leq \mathbb{E}_N \left[\max_{\gamma \in \mathcal{H}} \sum_{i=1}^{\lceil T/H \rceil} Q_i(\gamma) - \mathbb{E}_{\text{EXP3}} \left[\sum_{i=1}^{\lceil T/H \rceil} Q_i(\gamma_i) \right] \right] \\ &\leq \mathbb{E}_N \left[\max_{\gamma \in \mathcal{H}} \sum_{i=1}^{\lceil T/H \rceil} Q_i(\gamma) - \mathbb{E}_{\text{EXP3}} \left[\sum_{i=1}^{\lceil T/H \rceil} Q_i(\gamma_i) \right] \Big| \mathcal{U} \right] \mathbb{P}(\mathcal{U}) \, . \end{aligned}$$

We introduce

$$Y_i(\gamma_j) = \frac{Q_i(\gamma_j)}{mH} \, .$$

For all γ in $\mathcal{H}$, $Y_i(\gamma)$ lies in $[0, 1]$. Therefore,

$$master \leq mH\mathbb{E}_N \left[\max_{\gamma \in \mathcal{H}} \sum_{i=1}^{\lceil T/H \rceil} Y_i(\gamma) - \mathbb{E}_{\text{EXP3}} \left[\sum_{i=1}^{\lceil T/H \rceil} Y_i(\gamma_i) \right] \Big| \mathcal{U} \right] \, .$$

The last step consists in using [Auer et al., 2002b, Corollary 3.2]. We have,

$$\max_{\gamma \in \mathcal{H}} \sum_{i=1}^{\lceil T/H \rceil} Y_i(\gamma) \leq \frac{T}{H} \, .$$

All the conditions of Corollary 3.2 in [Auer et al., 2002b] are met and we obtain:

$$master \leq 2mH\sqrt{e-1}\sqrt{\frac{T}{H}\text{card}(\mathcal{H})\log(\text{card}(\mathcal{H}))} \, .$$

□

The two parts of regret in Equation (6.29) are bounded in Proposition 6.14 and Proposition 6.15 respectively. Combining them, we get our main result below:

Theorem 6.3. *Under Assumptions 6.1-6.2 and 6.4, the expected regret of* BOB-BVD-GLM-UCB *when setting* $H = \lfloor d\sqrt{T} \rfloor$ *satisfies:*

$$\mathcal{R}(T) = \tilde{\mathcal{O}}\left(R_\mu d^{2/3} T^{2/3} \max\left(\mathcal{B}_T, d^{-1/2} T^{1/4} \right)^{1/3} \right) \, .$$

Remark 6.7. *This theorem establishes an upper-bound for the expected regret in the Generalized Linear Bandits framework when the variational budget is unknown. When* $\mathcal{B}_T$ *is sufficiently large* ($\mathcal{B}_T \geq d^{-1/2}T^{1/4}$) *the obtained bound can not be improved. Yet, there is still a gap with the lower bound when the variation budget is small. This can be explained by the frequent restarts in the BOB framework.*

Proof. Using Proposition 6.15 and Proposition 6.14, we obtain:

$$\mathbb{E}\left[R(T)\right] \leq m\frac{T}{H} + C_1 R_\mu \beta_H^\star \sqrt{dT}\sqrt{2T(1-\gamma^\star) + \frac{T}{H}\log\left(1+\frac{2L^2}{d\lambda(1-\gamma^\star)}\right)}$$
$$+ \; C_2 R_\mu \frac{2}{\sqrt{T}}\frac{1}{1-\gamma^\star} + \frac{3C_3 R_\mu}{\log(2)}\frac{\mathcal{B}_T \log(T)}{1-\gamma^\star} + 2mH\sqrt{e-1}\sqrt{\frac{T}{H}\mathrm{card}(\mathcal{H})\log(\mathrm{card}(\mathcal{H}))}$$

First note that $\mathrm{card}(\mathcal{H}) = N$ defined in Equation (6.25) scales as $\log(T)$ and $\beta_H^\star$ scales as $\sqrt{d\log(T)}$. By plugging $H = \lfloor d\sqrt{T} \rfloor$ in the upper-bound we obtain:

$$\frac{T}{H} = \mathcal{O}(d^{-1/2}\sqrt{T}) \; .$$

$$\beta_H^\star \sqrt{dT}\sqrt{2T(1-\gamma^\star) + \frac{T}{H}\log\left(1+\frac{2L^2}{d\lambda(1-\gamma^\star)}\right)} = \widetilde{\mathcal{O}}\left(d\sqrt{T}\sqrt{\max\left(\frac{T\mathcal{B}_T^{2/3}}{d^{2/3}T^{2/3}}, \frac{T}{d\sqrt{T}}\right)}\right)$$
$$= d^{2/3}T^{2/3}\max(\mathcal{B}_T^{1/3}, d^{-1/6}T^{1/12})$$
$$= d^{2/3}T^{2/3}(\max(\mathcal{B}_T, d^{-1/2}T^{1/4}))^{1/3} \; .$$

$$\frac{1}{\sqrt{T}}\frac{1}{1-\gamma^\star} = \mathcal{O}\left(\frac{T^{1/6}}{d^{2/3}\mathcal{B}_T^{2/3}}\right) \; .$$

$$\frac{\mathcal{B}_T}{1-\gamma^\star} = \mathcal{O}\left(d^{2/3}\mathcal{B}_T^{1/3}T^{2/3}\right) \; .$$

$$H\sqrt{\frac{T}{H}\mathrm{card}(\mathcal{H})\log(\mathrm{card}(\mathcal{H}))} = \widetilde{\mathcal{O}}\left(d^{1/2}T^{3/4}\right) \; .$$

To conclude we notice that when $\mathcal{B}_T \leq d^{-1/2}T^{1/4}$,

$$d^{1/2}T^{3/4} = d^{2/3}T^{2/3}(\max(\mathcal{B}_T, d^{-1/2}T^{1/4}))^{1/3} \; .$$

On the contrary, when $\mathcal{B}_T \geq d^{-1/2}T^{1/4}$,

$$d^{1/2}T^{3/4} \leq d^{2/3}T^{2/3}(\max(\mathcal{B}_T, d^{-1/2}T^{1/4}))^{1/3} \; .$$

Finally, keeping the highest order term yields the announced result. □

Appendix 6.E Experimental Setup

This section is dedicated at providing useful details about the illustrative experiments presented in Section 6.6. The logistic setting at hand is characterized by the constants $S = L = 1$. At each round, the environment randomly draws 10 news arms, presented to the agent. All algorithms use the same ℓ_2 regularization parameter $\lambda = 1$. The sequence $\theta_t^\star$ evolves as follows: we let $\theta_t^\star = (0,1)$ for $t \in [1, T/3]$. Between $t = T/3$ and $t = 2T/3$ we smoothly rotate $\theta_t^\star$ from $(0,1)$ to $(1,0)$. Finally we let $\theta_t^\star = (0,1)$ for $t \in [2T/3, T]$. Easy computations show that the

total variation budget is

$$B_T = (2T/3)\sin\left(\frac{3\pi}{4T}\right) \simeq 1.5 \ .$$

We used the optimal value of γ recommended by the asymptotic analysis for D-LinUCB and BVD-GLM-UCB. We solve the projection step of GLM-UCB and BVD-GLM-UCB by (constrained) gradient-based methods, thanks to the SLSQP solver of `scipy`.

Remark 6.8. *In our experiments, we did not report the performance of the algorithms from [Russac et al., 2020, Russac et al., 2021a] that are using a projection step similar to the one used in [Filippi et al., 2010]). Because such algorithms are based on* discrete *switches of the reward signal, their behavior in this slowly-varying environment is largely sub-optimal. Indeed, in our experiment the number of abrupt-changes is* $\Gamma_T = 1000$. *For exponentially weighted algorithms, the recommended asymptotic value for the weights becomes* $\gamma \simeq 0.70$, *which in turns leads to algorithms that over-estimate the non-stationary nature of the problem, and perform poorly in practice.*

7 | Conclusion

In this thesis two different settings have been considered. In Chapter 2, we addressed a new pure exploration task. Its objective was to identify all the arms that are better than a control arm in the presence of subpopulations. We were able to quantify the complexity of the learning objective depending on the level of interaction of the learner with the different subpopulations.

The remaining chapters are devoted to the problem of reward maximization in non-stationary bandit models, with an increasing level of generality throughout the thesis. In Chapter 3, we considered the multi-armed bandit model, whereas Chapter 4 focused on the linear bandit model and Chapters 5 and 6 dealt with generalized linear models. In all of those settings, we proposed to combine forgetting mechanisms (through discount factors or with a sliding window) with subsampling (Chapter 3) or upper-confidence bound based techniques (Chapters 4, 5 and 6).

In the simpler multi-armed bandit model, we were able to obtain an asymptotically optimal algorithm in abruptly changing environments using a sliding window with the SW-LB-SDA algorithm (Chapter 3). Even when non-stationarity is measured through the variation budget, we have empirical evidence that the algorithm works well. An interesting future direction would be to extend the analysis to those more general non-stationary environments.

In the linear bandit model, we proposed D-LinUCB, an algorithm based on a weighted least squares estimator. With an additional assumption on the action sets, we established the asymptotic optimality of D-LinUCB. Yet, for general action sets, the upper-bound we obtained is larger than existing lower bounds that apply to specific instances of this setting. It is not clear if forgetting strategies are fundamentally suboptimal for general action sets or if this setting is fundamentally harder, which should be confirmed with a proper lower bound.

In the generalized linear model with abruptly changing environments, we obtained an asymptotically optimal (with respect to T and Γ_T) algorithm when adding an assumption on the gaps. With a new concentration inequality, we reduced the dependency in c_μ, a problem-dependent constant coming from the non-linearity of the model. Nevertheless, the analysis of the algorithms we designed suggests that the more non-linear the model, the harder the learning. [Abeille et al., 2021] recently came up with a new analysis with a different conclusion. They proposed an algorithm, termed OFU-GLB, where the effect of the non-linearity is a second order term for the regret and is tied to a transitory regime. For large time horizons, the effect of the non-linearity impacts the regret only through the reward sensitivity around the optimal action $\dot{\mu}(A_\star^\top \theta^\star)$. Interestingly, this suggests that some non-linear problems are much easier than their linear counterparts. These findings can be used in abruptly changing environments as proposed in [Faury, 2021, Chapter 4]. Understanding how to adapt these ideas to more general drifting environments is an interesting direction for future work.

For generalized linear bandits in drifting environments, we proposed a first complete analysis and uncovered mistakes made in several existing works. The nature of the difference between the

regret bounds from the linear bandit (of order $\widetilde{\mathcal{O}}(B_T^{1/4}T^{3/4})$) and those of the generalized linear bandit models (of order $\widetilde{\mathcal{O}}(B_T^{1/5}T^{4/5})$) in drifting environments is unsettled. We postulate that this is an artefact of the proof and that an improved analysis should yield the same rates.

Finally, all the algorithms that we have proposed for the regret minimization setting require some information about the non-stationarity of the environment. Obtaining optimal algorithms without this knowledge is still a domain under investigation [Chen et al., 2019, Auer et al., 2019, Wei and Luo, 2021]. However, doing without this knowledge about the non-stationarity comes at a cost: for the moment, these algorithms can not be implemented easily. None of the previously mentioned papers provide simulation to assess the empirical performance of the algorithms they propose. The ultimate goal for non-stationary bandits would be to simultaneously satisfy the three following requirements: (1) reaching optimality, (2) being agnostic to the non-stationarity of the environment and (3) being tractable in practice.

Bibliography

[Abbasi-Yadkori et al., 2011] Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320.

[Abeille et al., 2021] Abeille, M., Faury, L., and Calauzènes, C. (2021). Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3691–3699. PMLR.

[Abeille and Lazaric, 2017] Abeille, M. and Lazaric, A. (2017). Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184. PMLR.

[Achddou et al., 2021] Achddou, J., Cappé, O., and Garivier, A. (2021). Efficient algorithms for stochastic repeated second-price auctions. In *Algorithmic Learning Theory*, pages 99–150. PMLR.

[Agrawal, 1995] Agrawal, R. (1995). Sample mean based index policies by o (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078.

[Agrawal and Goyal, 2012] Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings.

[Agrawal and Goyal, 2013a] Agrawal, S. and Goyal, N. (2013a). Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, pages 99–107. PMLR.

[Agrawal and Goyal, 2013b] Agrawal, S. and Goyal, N. (2013b). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135. PMLR.

[Audibert et al., 2010] Audibert, J.-Y., Bubeck, S., and Munos, R. (2010). Best arm identification in multi-armed bandits. In *Conference on Learning Theory*, pages 41–53.

[Auer, 2002] Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

[Auer et al., 2002a] Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47.

[Auer et al., 2002b] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77.

[Auer et al., 2018] Auer, P., Gajane, P., and Ortner, R. (2018). Adaptively tracking the best arm with an unknown number of distribution changes. In *European Workshop on Reinforcement Learning, EWRL 2018*.

[Auer et al., 2019] Auer, P., Gajane, P., and Ortner, R. (2019). Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pages 138–158.

[Auer et al., 2008] Auer, P., Jaksch, T., and Ortner, R. (2008). Near-optimal regret bounds for reinforcement learning. In *Advances in neural information processing systems, NeurIPS*, pages 89–96.

[Bach, 2010] Bach, F. (2010). Self-concordant analysis for logistic regression. *Electron. J. Statist.*, 4:384–414.

[Bach, 2014] Bach, F. (2014). Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(19):595–627.

[Bach and Moulines, 2013] Bach, F. and Moulines, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in Neural Information Processing Systems*, pages 773–781.

[Baransi et al., 2014] Baransi, A., Maillard, O.-A., and Mannor, S. (2014). Sub-sampling for multi-armed bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 115–131. Springer.

[Baudry et al., 2020] Baudry, D., Kaufmann, E., and Maillard, O.-A. (2020). Sub-sampling for efficient non-parametric bandit exploration. *Advances in Neural Information Processing Systems*, 33.

[Baudry et al., 2021] Baudry, D., Russac, Y., and Cappé, O. (2021). On limited-memory subsampling strategies for bandits. In *International Conference on Machine Learning*, pages 727–737. PMLR.

[Besbes et al., 2014] Besbes, O., Gur, Y., and Zeevi, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems*, pages 199–207.

[Besbes et al., 2015] Besbes, O., Gur, Y., and Zeevi, A. (2015). Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244.

[Besbes et al., 2018] Besbes, O., Gur, Y., and Zeevi, A. (2018). Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards. *Available at SSRN 2436629*.

[Besson et al., 2020] Besson, L., Kaufmann, E., Maillard, O.-A., and Seznec, J. (2020). Efficient change-point detection for tackling piecewise-stationary bandits. Preprint.

[Bleakley and Vert, 2011] Bleakley, K. and Vert, J.-P. (2011). The group fused lasso for multiple change-point detection. *arXiv preprint arXiv:1106.4199*.

[Bouneffouf et al., 2012] Bouneffouf, D., Bouzeghoub, A., and Gançarski, A. L. (2012). A contextual-bandit algorithm for mobile context-aware recommender system. In *International Conference on Neural Information Processing*, pages 324–331. Springer.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge Univ Press.

[Bubeck et al., 2009] Bubeck, S., Munos, R., and Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer.

[Burnetas and Katehakis, 1996] Burnetas, A. and Katehakis, M. (1996). Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2).

[Cao et al., 2019] Cao, Y., Wen, Z., Kveton, B., and Xie, Y. (2019). Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 418–427. PMLR.

[Cappé et al., 2013] Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., Stoltz, G., et al. (2013). Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541.

[Carpentier and Munos, 2011] Carpentier, A. and Munos, R. (2011). Finite-time analysis of stratified sampling for monte carlo. In *NIPS-Twenty-Fifth Annual Conference on Neural Information Processing Systems*.

[Chan, 2020] Chan, H. P. (2020). The multi-armed bandit problem: An efficient nonparametric solution. *The Annals of Statistics*, 48(1):346–373.

[Chapelle and Li, 2011] Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257.

[Chen et al., 2017] Chen, L., Li, J., and Qiao, M. (2017). Nearly instance optimal sample complexity bounds for top-k arm selection. In *Artificial Intelligence and Statistics*, pages 101–110. PMLR.

[Chen et al., 2020] Chen, N., Wang, C., and Wang, L. (2020). Learning and optimization with seasonal patterns. *arXiv preprint arXiv:2005.08088*.

[Chen et al., 2019] Chen, Y., Lee, C.-W., Luo, H., and Wei, C.-Y. (2019). A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Conference on Learning Theory*, pages 696–726. PMLR.

[Chernoff, 1959] Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770.

[Cheshire et al., 2020] Cheshire, J., Menard, P., and Carpentier, A. (2020). The influence of shape constraints on the thresholding bandit problem. In *Conference on Learning Theory*, pages 1228–1275. PMLR.

[Cheung et al., 2019] Cheung, W. C., Simchi-Levi, D., and Zhu, R. (2019). Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1079–1087. PMLR.

[Cheung et al., 2021] Cheung, W. C., Simchi-Levi, D., and Zhu, R. (2021). Hedging the drift: Learning to optimize under nonstationarity. *Management Science*.

[Chu et al., 2011] Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings.

[Dani et al., 2008] Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. *Conference on Learning Theory*.

[De Rooij et al., 2014] De Rooij, S., Van Erven, T., Grünwald, P. D., and Koolen, W. M. (2014). Follow the leader if you can, hedge if you must. *The Journal of Machine Learning Research*, 15(1):1281–1316.

[Degenne and Koolen, 2019] Degenne, R. and Koolen, W. M. (2019). Pure exploration with multiple correct answers. In *32*, pages 14591–14600.

[Degenne et al., 2019] Degenne, R., Koolen, W. M., and Ménard, P. (2019). Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*, 32:14492–14501.

[Di Benedetto et al., 2020] Di Benedetto, G., Bellini, V., and Zappella, G. (2020). A linear bandit for seasonal environments. *arXiv preprint arXiv:2004.13576*.

[Diemert et al., 2017] Diemert, E., Meynet, J., Galland, P., and Lefortier, D. (2017). Attribution modeling increases efficiency of bidding in display advertising. In *Proceedings of the AdKDD and TargetAd Workshop, KDD, Halifax, NS, Canada, August, 14, 2017*. ACM.

[Ding et al., 2020] Ding, Q., Hsieh, C.-J., and Sharpnack, J. (2020). Multiscale non-stationary stochastic bandits. *arXiv preprint arXiv:2002.05289*.

[Dong and Van Roy, 2018] Dong, S. and Van Roy, B. (2018). An Information-Theoretic Analysis for Thompson Sampling with Many Actions. In *Advances in Neural Information Processing Systems*, pages 4157–4165.

[Dumitrascu et al., 2018] Dumitrascu, B., Feng, K., and Engelhardt, B. (2018). PG-TS: Improved Thompson Sampling for Logistic Contextual Bandits. In *Advances in Neural Information Processing Systems*, pages 4624–4633.

[Durand et al., 2018] Durand, A., Achilleos, C., Iacovides, D., Strati, K., Mitsis, G. D., and Pineau, J. (2018). Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine learning for healthcare conference*, pages 67–82. PMLR.

[Eliashberg and Jeuland, 1986] Eliashberg, J. and Jeuland, A. P. (1986). The impact of competitive entry in a developing market upon dynamic pricing strategies. *Marketing Science*, 5(1):20–36.

[Even-Dar et al., 2006] Even-Dar, E., Mannor, S., Mansour, Y., and Mahadevan, S. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6).

[Faury, 2021] Faury, L. (2021). *Variance-Sensitive Confidence Intervals for Parametric and Offline Bandits*. PhD thesis, Institut Polytechnique de Paris.

[Faury et al., 2020] Faury, L., Abeille, M., Calauzènes, C., and Fercoq, O. (2020). Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR.

[Faury et al., 2021a] Faury, L., Russac, Y., Abeille, M., and Calauzènes, C. (2021a). Regret bounds for generalized linear bandits under parameter drift. *arXiv preprint arXiv:2103.05750*.

[Faury et al., 2021b] Faury, L., Russac, Y., Abeille, M., and Calauzènes, C. (2021b). A technical note on non-stationary parametric bandits: Existing mistakes and preliminary solutions. In *Algorithmic Learning Theory*.

[Filippi et al., 2010] Filippi, S., Cappé, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems, NeurIPS 2010*, pages 586–594.

[Flajolet and Jaillet, 2017] Flajolet, A. and Jaillet, P. (2017). Real-time bidding with side information. In *Advances in neural information processing systems*, pages 5168–5178.

[Gabillon et al., 2012] Gabillon, V., Ghavamzadeh, M., and Lazaric, A. (2012). Best arm identification: A unified approach to fixed budget and fixed confidence. In *NIPS-Twenty-Sixth Annual Conference on Neural Information Processing Systems*.

[Gajane et al., 2018] Gajane, P., Ortner, R., and Auer, P. (2018). A sliding-window algorithm for markov decision processes with arbitrarily changing rewards and transitions. *arXiv preprint arXiv:1805.10066*.

[Garivier and Kaufmann, 2016] Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR.

[Garivier et al., 2019] Garivier, A., Ménard, P., and Stoltz, G. (2019). Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399.

[Garivier and Moulines, 2008] Garivier, A. and Moulines, E. (2008). On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*.

[Garivier and Moulines, 2011] Garivier, A. and Moulines, E. (2011). On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188. Springer.

[Gittins, 1974] Gittins, J. (1974). A dynamic allocation index for the sequential design of experiments. *Progress in statistics*, pages 241–266.

[Goldenshluger and Zeevi, 2013] Goldenshluger, A. and Zeevi, A. (2013). A linear response bandit problem. *Stoch. Syst.*, 3(1):230–261.

[Gorre et al., 2001] Gorre, M. E., Mohammed, M., Ellwood, K., Hsu, N., Paquette, R., Rao, P. N., and Sawyers, C. L. (2001). Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science*, 293(5531):876–880.

[Gupta et al., 2011] Gupta, N., Granmo, O.-C., and Agrawala, A. (2011). Thompson sampling for dynamic multi-armed bandits. In *2011 10th International Conference on Machine Learning and Applications and Workshops*, volume 1. IEEE.

[Hariri et al., 2015] Hariri, N., Mobasher, B., and Burke, R. (2015). Adapting to user preference changes in interactive recommendation. In *IJCAI*, volume 15, pages 4268–4274.

[Hartland et al., 2006] Hartland, C., Gelly, S., Baskiotis, N., Teytaud, O., and Sebag, M. (2006). Multi-armed bandit, dynamic environments and meta-bandits. *Preprint. hal-archives-ouvertes.fr*.

[Hinkley, 1971] Hinkley, D. V. (1971). Inference about the change-point from cumulative sum tests. *Biometrika*, 58(3):509–523.

[Honda and Takemura, 2015] Honda, J. and Takemura, A. (2015). Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *Journal of Machine Learning Research*, 16:3721–3756.

[Johari et al., 2015] Johari, R., Pekelis, L., and Walsh, D. J. (2015). Always valid inference: Bringing sequential analysis to A/B testing. *arXiv preprint arXiv:1512.04922*.

[Jun et al., 2017] Jun, K.-S., Bhargava, A., Nowak, R., and Willett, R. (2017). Scalable Generalized Linear Bandits: Online Computation and Hashing. In *Advances in Neural Information Processing Systems*, pages 99–109.

[Jun et al., 2021] Jun, K.-S., Jain, L., Nassif, H., and Mason, B. (2021). Improved confidence bounds for the linear logistic model and applications to bandits. In *International Conference on Machine Learning*, pages 5148–5157. PMLR.

[Kalyanakrishnan et al., 2012] Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012). PAC subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662.

[Karnin and Anava, 2016] Karnin, Z. S. and Anava, O. (2016). Multi-armed bandits: Competing with optimal sequences. *Advances in Neural Information Processing Systems*, 29:199–207.

[Kaufmann et al., 2014] Kaufmann, E., Cappé, O., and Garivier, A. (2014). On the complexity of A/B testing. In *Conference on Learning Theory*, pages 461–481.

[Kaufmann et al., 2016] Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42.

[Kaufmann and Koolen, 2018] Kaufmann, E. and Koolen, W. M. (2018). Mixture martingales revisited with applications to sequential tests and confidence intervals. Preprint.

[Kaufmann et al., 2012] Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer.

[Keskin and Zeevi, 2017] Keskin, N. B. and Zeevi, A. (2017). Chasing demand: Learning and earning in a changing environment. *Mathematics of Operations Research*, 42(2):277–307.

[Kirschner and Krause, 2018] Kirschner, J. and Krause, A. (2018). Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pages 358–384. PMLR.

[Kleinberg and Immorlica, 2018] Kleinberg, R. and Immorlica, N. (2018). Recharging bandits. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 309–319. IEEE.

[Kocsis and Szepesvári, 2006a] Kocsis, L. and Szepesvári, C. (2006a). Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer.

[Kocsis and Szepesvári, 2006b] Kocsis, L. and Szepesvári, C. (2006b). Discounted ucb. *In: 2nd Pascal Challenge Workshop*.

[Komiyama et al., 2021] Komiyama, J., Fouché, E., and Honda, J. (2021). Finite-time analysis of globally nonstationary multi-armed bandits. *arXiv preprint arXiv:2107.11419*.

[Korda et al., 2013] Korda, N., Kaufmann, E., and Munos, R. (2013). Thompson sampling for 1-dimensional exponential family bandits. In *Advances in neural information processing systems*, pages 1448–1456.

[Krishnamurthy and Gopalan, 2021] Krishnamurthy, R. and Gopalan, A. (2021). On slowly-varying non-stationary bandits. *arXiv preprint arXiv:2110.12916*.

[Kveton et al., 2019a] Kveton, B., Szepesvari, C., Ghavamzadeh, M., and Boutilier, C. (2019a). Perturbed-history exploration in stochastic multi-armed bandits. *Proceedings of the 28th International Joint Conference on Artificial Intelligence*.

[Kveton et al., 2019b] Kveton, B., Szepesvari, C., Vaswani, S., Wen, Z., Lattimore, T., and Ghavamzadeh, M. (2019b). Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 3601–3610. PMLR.

[Kveton et al., 2020] Kveton, B., Zaheer, M., Szepesvari, C., Li, L., Ghavamzadeh, M., and Boutilier, C. (2020). Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2066–2076. PMLR.

[Lai and Robbins, 1985] Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

[Lattimore and Szepesvári, 2020] Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

[Lattimore and Szepesvari, 2020] Lattimore, T. and Szepesvari, C. (2020). *Bandit Algorithms*. Cambridge University Press.

[Levine et al., 2017] Levine, N., Crammer, K., and Mannor, S. (2017). Rotting bandits. *Neural Information Processing Systems*.

[Li et al., 2010] Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM.

[Li et al., 2017] Li, L., Lu, Y., and Zhou, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR.

[Liu et al., 2018] Liu, F., Lee, J., and Shroff, N. (2018). A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

[Locatelli et al., 2016] Locatelli, A., Gutzeit, M., and Carpentier, A. (2016). An optimal algorithm for the thresholding bandit problem. In *International Conference on Machine Learning*, pages 1690–1698. PMLR.

[Luo et al., 2018] Luo, H., Wei, C.-Y., Agarwal, A., and Langford, J. (2018). Efficient contextual bandits in non-stationary worlds. In *Conference On Learning Theory*, pages 1739–1776. PMLR.

[Luo et al., 2021] Luo, Y., Gupta, V., and Kolar, M. (2021). Dynamic regret minimization for control of non-stationary linear dynamical systems. *arXiv preprint arXiv:2111.03772*.

[Mason et al., 2020] Mason, B., Jain, L., Tripathy, A., and Nowak, R. (2020). Finding all $\{\epsilon\}$-good arms in stochastic bandits. *Advances in Neural Information Processing Systems*.

[Mintz et al., 2020] Mintz, Y., Aswani, A., Kaminsky, P., Flowers, E., and Fukuoka, Y. (2020). Nonstationary bandits with habituation and recovery dynamics. *Operations Research*, 68(5):1493–1516.

[Mukherjee and Maillard, 2019] Mukherjee, S. and Maillard, O.-A. (2019). Distribution-dependent and time-uniform bounds for piecewise iid bandits. *arXiv preprint arXiv:1905.13159*.

[Munos, 2014] Munos, R. (2014). From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning.

[Nguyen, 2020] Nguyen, K. T. (2020). A bandit learning algorithm and applications to auction design. In *Advances in Neural Information Processing Systems*.

[Page, 1954] Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.

[Peña et al., 2008] Peña, V. H., Lai, T. L., and Shao, Q.-M. (2008). *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media.

[Pike-Burke and Grunewalder, 2019] Pike-Burke, C. and Grunewalder, S. (2019). Recovering bandits. *Advances in Neural Information Processing Systems*, 32:14122–14131.

[Radlinski et al., 2008] Radlinski, F., Kleinberg, R., and Joachims, T. (2008). Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*, pages 784–791. ACM.

[Raj and Kalyani, 2017] Raj, V. and Kalyani, S. (2017). Taming non-stationary bandits: A bayesian approach. *arXiv preprint arXiv:1707.09727*.

[Riou and Honda, 2020] Riou, C. and Honda, J. (2020). Bandit algorithms based on Thompson sampling for bounded reward distributions. In *Algorithmic Learning Theory*, pages 777–826. PMLR.

[Rusmevichientong and Tsitsiklis, 2010] Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411.

[Russac et al., 2020] Russac, Y., Cappé, O., and Garivier, A. (2020). Algorithms for non-stationary generalized linear bandits. *arXiv preprint arXiv:2003.10113*.

[Russac et al., 2021a] Russac, Y., Faury, L., Cappé, O., and Garivier, A. (2021a). Self-concordant analysis of generalized linear bandits with forgetting. In *International Conference on Artificial Intelligence and Statistics*, pages 658–666. PMLR.

[Russac et al., 2021b] Russac, Y., Katsimerou, C., Bohle, D., Cappé, O., Garivier, A., and Koolen, W. M. (2021b). A/B/n testing with control in the presence of subpopulations. *Advances in Neural Information Processing Systems*, 34.

[Russac et al., 2019] Russac, Y., Vernade, C., and Cappé, O. (2019). Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems*, 32:12040–12049.

[Russo and Van Roy, 2014] Russo, D. and Van Roy, B. (2014). Learning to Optimize via Posterior Sampling. *Mathematics of Operations Research*, 39(4):1221–1243.

[Seznec et al., 2019] Seznec, J., Locatelli, A., Carpentier, A., Lazaric, A., and Valko, M. (2019). Rotting bandits are no harder than stochastic ones. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2564–2572. PMLR.

[Seznec et al., 2020] Seznec, J., Menard, P., Lazaric, A., and Valko, M. (2020). A single algorithm for both restless and rested rotting bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3784–3794. PMLR.

[Silver et al., 2016] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

[Srivastava et al., 2014] Srivastava, V., Reverdy, P., and Leonard, N. E. (2014). Surveillance in an abruptly changing world via multiarmed bandits. In *53rd IEEE Conference on Decision and Control*, pages 692–697. IEEE.

[Tewari and Murphy, 2017] Tewari, A. and Murphy, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer.

[Thompson, 1933] Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

[Touati and Vincent, 2020] Touati, A. and Vincent, P. (2020). Efficient learning in non-stationary linear markov decision processes. *arXiv preprint arXiv:2010.12870*.

[Tracà et al., 2021] Tracà, S., Rudin, C., and Yan, W. (2021). Regulating greed over time in multi-armed bandits. *Journal of Machine Learning Research*, 22:3–1.

[Trovo et al., 2020] Trovo, F., Paladino, S., Restelli, M., and Gatti, N. (2020). Sliding-window Thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*, 68:311–364.

[Valko et al., 2014] Valko, M., Munos, R., Kveton, B., and Kocák, T. (2014). Spectral bandits for smooth graph functions. In *Proceedings of the 31st International Conference on Machine Learning, ICML 2014*, pages 46–54.

[Wang et al., 2017] Wang, Y., Ouyang, H., Wang, C., Chen, J., Asamov, T., and Chang, Y. (2017). Efficient ordered combinatorial semi-bandits for whole-page recommendation. In *AAAI*, pages 2746–2753.

[Wei et al., 2016] Wei, C.-Y., Hong, Y.-T., and Lu, C.-J. (2016). Tracking the best expert in non-stationary stochastic environments. In *Advances in neural information processing systems*, pages 3972–3980.

[Wei and Luo, 2021] Wei, C.-Y. and Luo, H. (2021). Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. *Conference on Learning Theory*.

[Wei and Srivatsva, 2018] Wei, L. and Srivatsva, V. (2018). On abruptly-changing and slowly-varying multiarmed bandit problems. In *2018 Annual American Control Conference (ACC)*, pages 6291–6296. IEEE.

[Whittle, 1988] Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298.

[Wu et al., 2018] Wu, Q., Iyer, N., and Wang, H. (2018). Learning contextual bandits in a non-stationary environment. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 495–504.

[Xu et al., 2020] Xu, X., Dong, F., Li, Y., He, S., and Li, X. (2020). Contextual-bandit based personalized recommendation with time-varying user interests. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6518–6525.

[Yang et al., 2017] Yang, F., Ramdas, A., Jamieson, K., and Wainwright, M. J. (2017). A framework for Multi-A(rmed)/B(andit) testing with online FDR control. In *Advances in Neural Information Processing Systems*.

[Yu and Mannor, 2009] Yu, J. Y. and Mannor, S. (2009). Piecewise-stationary bandit problems with side observations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1177–1184. ACM.

[Yue and Joachims, 2009] Yue, Y. and Joachims, T. (2009). Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM.

[Zhao and Zhang, 2021] Zhao, P. and Zhang, L. (2021). Non-stationary linear bandits revisited. *arXiv preprint arXiv:2103.05324*.

[Zhao et al., 2020] Zhao, P., Zhang, L., Jiang, Y., and Zhou, Z.-H. (2020). A simple approach for non-stationary linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 746–755. PMLR.

[Zhou et al., 2020] Zhou, H., Wang, L., Varshney, L. R., and Lim, E.-P. (2020). A near-optimal change-detection based algorithm for piecewise-stationary combinatorial semi-bandits. *AAAI*.

[Zhou et al., 2019] Zhou, Z., Xu, R., and Blanchet, J. (2019). Learning in generalized linear contextual bandits with stochastic delays. In *Advances in Neural Information Processing Systems*.

RÉSUMÉ

La version classique du modèle de bandit suppose que les distributions de probabilité des récompenses sont indépendantes et identiquement distribuées. Pour autant, cette hypothèse est restrictive dans de nombreux cas, puisqu'elle ne permet pas de prendre en compte d'éventuels changements de comportements. Dans le domaine médical, l'efficacité d'un traitement peut diminuer au cours du temps. Pour un site internet d'information en temps réel, le taux de consultation d'une page diminue à raison de sa date d'ancienneté. Les tendances de mode et les préférences des consommateurs évoluent rapidement. Un algorithme de recommendation ignorant ces formes de non-stationarité est alors susceptible de faire des suggestions sous-optimales. Ainsi, l'objet de cette thèse est l'étude des algorithmes de bandits stochastiques dans des environnements non-stationnaires. La non-stationarité peut être incorporée de plusieurs manières dans les modèles de bandits. Dans un premier temps, nous étudions une variante du problème d'identification du meilleur bras. Cette variante correspond à un système d'apprentissage qui cherche à identifier l'ensemble des options qui sont meilleures qu'un bras de contrôle, et ce en présence de sous-populations. Entre autres, l'utilisation de sous-populations permet la modélisation de l'évolution temporelle des différents bras. Nous proposons ensuite des algorithmes avec des garanties théorique fortes pour la minimisation du regret et étudions le compromis exploration-exploitation pour de tels environnements. Nos recherches portent sur trois modèles différents: le bandit classique multi-bras, le bandit linéaire ou encore le bandit linéaire généralisé. Nous examinons les spécificités de chacun de ces trois modèles, ainsi que les défis techniques liés à la non-stationarité.

MOTS CLÉS

Apprentissage séquentiel, algorithmes de bandits, environnements non stationnaires, minimisation du regret.

ABSTRACT

The vanilla bandit model assumes that the rewards are independent and identically distributed. However, this assumption is restrictive: it prevents from modeling evolving behaviors that are common in real-world applications. In the medical domain, the efficiency of a treatment is likely to diminish over time. The opening rate of news articles fades for aging news. Fashion trends and consumers preferences evolve rapidly. Any recommender system ignoring the non-stationarity of the distributions of rewards is likely to make suboptimal choices. The objective of this thesis is the study of stochastic bandit algorithms in non-stationary environments. There are several ways to include non-stationarity into bandit models. We first study a variant of the best arm identification problem where the learner seeks to identify the set of arms that are better than a control arm in the presence of subpopulations. Those subpopulations can encode a temporal information (e.g. day of the week) and properly using them makes it possible to include non-stationarity in the pure exploration setting. We characterize the complexity of this learning task and propose optimal algorithms for solving it. We then propose theoretically grounded algorithms for minimizing the regret and discuss the exploration-exploitation trade-off the learner is facing in dynamically changing environments. Our findings concern three different settings: the well-known multi-armed bandit, the more general linear bandit but also generalized linear bandit. For each of those settings, we identify the technical challenges brought by non-stationarity.

KEYWORDS

Sequential learning, bandit algorithms, non-stationary environments, regret minimization.