



INSTITUT  
POLYTECHNIQUE  
DE PARIS

NNT : 2021IPPAT046

Thèse de doctorat



# Variance-Sensitive Confidence Intervals for Parametric and Offline Bandits

Thèse de doctorat de l'Institut Polytechnique de Paris  
préparée à Télécom Paris

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)  
Spécialité de doctorat : Mathématiques Appliquées

Thèse présentée et soutenue à Paris, le 11/10/2021, par

**LOUIS FAURY**

Composition du Jury :

Francis Bach Professeur, ENS Paris	Président
Émilie Kaufmann Chargée de recherche, Inria Lille	Rapporteuse
Tor Lattimore Chercheur, DeepMind	Rapporteur
Thomas Bonald Professeur, TélécomParis	Examineur
Aurélien Garivier Professeur, ENS Lyon	Examineur
Olivier Fercoq Maître de conférence, TélécomParis	Directeur de thèse
Marc Abeille Chercheur, Criteo	Invité

**Short English abstract.** In this dissertation we present recent contributions to the problem of optimization under bandit feedback through the design of variance-sensitive confidence intervals. We tackle two distinct topics: **(1)** the regret minimization task in Generalized Linear Bandits (GLBs for short, a broad class of non-linear parametric bandits) and **(2)** the problem of off-line policy optimization under bandit feedback. For **(1)** we study the effects of non-linearity in GLBs and challenge the current understanding that a high level of non-linearity is detrimental to the exploration-exploitation trade-off. We introduce improved algorithms as well as a novel analysis that prove that if correctly handled the regret minimization task in GLBs is not necessarily harder than for their linear counterparts. It can even be *easier* for some important members of the GLB family such as the Logistic Bandit. Our approach leverages a new confidence set which captures the non-linearity of the reward signal through its variance, along with a local treatment of the non-linearity through a so-called self-concordance analysis. For **(2)** we leverage results from the distributionally robust optimization framework to construct asymptotic variance-sensitive confidence intervals for the counterfactual evaluation of policies. This allows to ensure conservatism (sought out by risk-averse agents) while searching off-line for promising policies. Our confidence intervals lead to new counterfactual objectives which, contrary to their predecessors, are more suited for practical deployment thanks to their convex and composite natures.

**Titre en Français.** Intervalles de Confiance Sensibles à la Variance: Applications aux Bandits Paramétrique et Bandits Hors Ligne.

**Résumé court en Français.** Cette thèse présente des contributions récentes au problème d'optimisation sous feedback bandit, à travers la construction d'intervalles de confiance sensibles à la variance. Nous traitons deux aspects distincts du problème: **(1)** la minimisation du regret pour les bandits à modèle linéaire généralisé (GLBs), une large classe de bandits paramétriques non-linéaires et **(2)** le problème d'optimisation de politique hors ligne sous signal bandit. Concernant **(1)** nous étudions les effets de la non-linéarité dans les GLBs et remettons en question la compréhension actuelle selon laquelle des hauts niveaux de non-linéarité ne peuvent être que préjudiciables à l'équilibre exploration-exploitation. Des algorithmes améliorés suivis d'une nouvelle méthode d'analyse montre que lorsque correctement manipulé, le problème de minimisation du regret dans les GLBs n'est pas nécessairement plus dur que pour leur contrepartie linéaire. Il peut même être significativement facilité pour certains membres importants de la famille GLB comme le bandit logistique. Notre approche utilise de nouveaux ensembles de confiance sensibles à la non-linéarité au travers de la variance qu'elle impose à la fonction récompense, accompagnés d'un traitement local de la non-linéarité au travers d'une analyse dite auto-concordante. Concernant **(2)** nous utilisons des résultats de la littérature de l'optimisation robuste afin de construire des intervalles de confiance asymptotiques sensibles à la variance pour l'évaluation contrefactuel de politiques. Cela permet d'assurer du conservatisme (désirable pour des agents averses au risque) lors de la recherche hors-ligne de politiques prometteuses. Cet interval de confiance engendre de nouveaux objectifs contrefactuels qui sont plus adaptés à des applications pratiques, car convexes et de nature composites.

**Key words:** machine learning, confidence intervals, decision-making, bandit algorithm, linear bandit, generalized linear bandit, counterfactual estimation, off-line optimization.

**Mots-clefs:** apprentissage automatique, processus décisionnel, algorithme bandit, bandit linéaire, bandit linéaire généralisé, estimation contrafactuelle, optimisation hors-ligne.

# Remerciements - Acknowledgements

---

Avant toute chose, je veux exprimer ma sincère et profonde gratitude envers mes deux encadrants de thèse, Marc et Olivier. Votre accompagnement, vos conseils et votre enthousiasme pour la recherche m'ont été d'une aide précieuse tout au long de ce doctorat et je vous en suis extrêmement reconnaissant.

Olivier, merci pour la patience et l'optimisme qui te caractérise. Je te remercie en particulier pour la liberté que tu m'as laissé d'explorer différents sujets de recherche, et l'enthousiasme que tu as manifesté au début de chaque nouveau projet. Je mesure et apprécie la confiance que tu m'as accordé à chaque étape de la thèse - j'espère en avoir fait bon usage.

Marc, un grand merci pour m'avoir en premier montrer les problèmes de bandit, puis pour avoir passer de longues heures au tableau blanc afin de m'en expliquer les subtilités. Merci d'avoir insisté, et de ne pas avoir abandonné après les (très) nombreuses bêtises et fautes grossières que j'ai pu écrire pendant mes premiers pas dans le monde des bandits. Merci pour cette grande complicité qui s'est ensuite établie entre nous, et qui a contribué à rendre ces longues heures à gratter des mathématiques ou à écrire des articles en des moments de bonne humeur et souvent, de bonne rigolade. Je ne peux en garder que d'excellents souvenirs, même si quelques peu embrumés parce que passés à des heures parfois très tardives, notamment avant certaines deadlines.. À ce sujet, je me dois aussi de remercier ton épouse Marion et tes enfants à qui je t'ai souvent emprunté à des heures indues parce que quelques heures avant la soumission "mais si je te jure la preuve est cassée!!". J'ai appris à ton contact la rigueur, l'exigence et l'excellence scientifique; si je peux suite à ce doctorat faire preuve ne serait-ce que d'un dixième de tes talents dans ces domaines, je considérerai ces trois dernières années rentabilisées.

Je tiens également à remercier Clément pour être un manager d'équipe de recherche hors-pair. J'ai conscience que nos vies de chercheurs sont rendues bien plus faciles grâce à toi, qui t'assures sans relâche que nous bénéficions du meilleur cadre de travail possible. Tes conseils, ta patience et ta pédagogie ont été des des éléments clefs de ma thèse - je t'en remercie. J'espère continuer à apprendre en travaillant à tes côtés dans les mois prochains; j'espère aussi continuer à trouver du temps pour te mettre des raclées à Battlefront et Rayman lors de longues soirées jeux vidéos!

Il convient également de remercier Criteo pour avoir mis en place un écosystème de recherche particulièrement efficace et agréable. Il m'a permis de faire une thèse à dominante théorique, tout en évoluant dans un environnement industriel riche et plein d'inspiration pour des sujets de recherche. C'est une chance tout particulière que d'avoir pu compléter ma thèse parmi les équipe de recherche de Criteo. Je dois cette chance à Flavian, à qui je souhaite exprimer toute ma gratitude pour m'avoir en premier accorder sa confiance, il y a maintenant plus de trois ans dans le cadre d'un stage de recherche. Un grand merci aussi à Ugo, pour des innombrables parties de ping-pong, de baby-foot et d'échecs, entre lesquelles il nous est arrivé de travailler ensemble, pour mon plus grand plaisir. Merci aux autres membres de mon équipe de recherche à Criteo (Vianney, Jérémie, et l'ensemble de l'équipe des thésards) pour des discussions scientifiques

toujours stimulantes. Merci également à Yoan, mon super collaborateur de confinement toujours plein d'optimisme; merci pour ta bonne humeur, bien utiles à certain moments dans les derniers mois!

Thanks to Francis Bach, Thomas Bonald, Aurélien Garivier, Émilie Kaufmann and Tor Lattimore for accepting to be part of my committee. A special thanks to Émilie and Tor for accepting to review my dissertation; it is a great honor for me.

Je remercie mes proches pour leur indéfectible soutien; en particulier, mes parents pour m'avoir encouragé à poursuivre des études scientifiques qui me permettent de m'épanouir pleinement aujourd'hui. Finally, I owe a huge thanks to Caroline for the happiness she brings me everyday. Thank you for supporting me through this project, as well as all the others.

# Introduction

---

## 1 High level presentation of the thesis

This dissertation presents recent contributions to the problem of sequential decision-making under uncertainty, an important formalization of countless real-world situations where one repeatedly interacts with an unknown environment in order to achieve a specific goal. In this contexts any decision-making agent inevitably faces a fundamental difficulty: how to efficiently achieve one's goal while simultaneously reasoning about the potential outcomes (often corrupted by noise) of the available actions? Behind this broad question lies a fundamental dilemma of sequential decision-making, known as the *exploration-exploitation* trade-off. To succeed the agent must achieve a careful balance between two conflicting objectives: increase its knowledge by probing the environment (*exploration*) and leverage the information acquired so far to enhance its performance (*exploitation*). Designing intelligent agents that achieve such a trade-off is central to the development of artificial intelligence, and at the heart of machine learning fields such as bandit optimization - the topic of this dissertation. When learning under bandit feedback, the agent enters a repeated game with the environment; it sequentially plays some available actions of which the outcomes are noisy random variables referred to as *rewards*. The goal of the agent can be, for instance, to maximize its cumulative reward over time (*regret minimization*) or identify the best actions within as few interactions as possible (*pure exploration*). Albeit adopting a seemingly simplistic setting, the bandit optimization problem compactly captures many learning-theoretic difficulties of sequential decision-making under uncertainty. More importantly, it allows for a neat and precise theoretical treatment, which has brought forward clear and portable principles. This portability was particularly well-illustrated through the rise of Internet technologies and services, which prompted the deployment of bandit algorithms for a great variety of tasks - *e.g* A/B testing, news/movies recommendation and ad-placement.

A natural idea when playing with an unknown environment is to *estimate* the environment's response to one's actions. This task is made harder when this response is stochastic, as one must then reason about the *plausible* environments that are likely to have generated the sequences of rewards so far observed. This rationale has driven most of the research in sequential decision making under uncertainty - the bandit optimization problem being no exception. It is for instance central to address the regret minimization problem thanks to the *optimism-in-face-of-uncertainty* principle, which in order to efficiently balance exploration and exploitation prescribes playing the action that appears to be the most rewarding in the most favorable plausible environment. At this point the notion of plausible environment may still seem abstract to the reader; throughout the bandit literature it has been formalized through the construction of *confidence intervals*. Designing tight confidence intervals and appropriately leveraging them for the task at hand has been instrumental in the development of efficient bandit algorithms.

This thesis follows this rationale and presents contributions to two distinct and almost or-

thogonal aspects of the bandit optimization problem: (1) the sequential optimization (*i.e.* regret minimization) of generalized linear models under bandit feedback and (2) the offline evaluation and optimization of policies under logged bandit feedback. Our results rely on new confidence intervals, each adapted to the learning-theoretic challenges of each task. They illustrate how the careful construction of appropriate confidence intervals can simultaneously bring new theoretical insight and lead to the design of algorithms with improved theoretical guarantees. This naturally translates into enhanced practical performances and sometimes substantially simpler and more efficient algorithms. We present our results into two distinct parts, for they fundamentally lie at two different ends of the spectrum of bandit optimization.

## Part I. Non-linearity in parametric bandits : an unavoidable curse?

Countless real-world problems are fundamentally non-linear; in order to address them numerous efforts have been made within the machine learning community to introduce ever more expressive models. Such efforts have been particularly successful in the supervised learning setting, where the advent of deep neural networks has revolutionized the field. A similar success-story is however still missing for sequential decision-making, where a finer understanding of such models is necessary to efficiently tackle many real-world environments. To achieve this an important missing part of the puzzle consists in describing the effect of non-linearities on the exploration-exploitation trade-off. The first part of this thesis aims at answering this question through the careful study of a generic class of parametric bandits known as Generalized Linear Bandits (GLBs). Indeed, it offers a simple and uncluttered framework to isolate and study the effects of non-linearity, while still being rich enough to be relevant for many practical applications.

A GLB model postulates that the expected reward associated to an action is obtained by computing a linear transformation of the action (seen as a vector in an Euclidian space), to which is then applied a *non-linear* map  $\mu$ . While such models have already undergone meticulous scrutiny in the literature, the current conclusion is that the more non-linear  $\mu$  is, the harder the learning problem is. Indeed, the regret guarantees of existing algorithms all scale linearly with  $\kappa$ , a constant measuring the function  $\mu$ 's degree of non-linearity. Unfortunately  $\kappa$  is disproportionately large even for reasonable models, which suggests that non-linearity is highly *detrimental* to the algorithms' performance. Whether such behavior is unavoidable (*i.e.* it is inherited from a fundamental difficulty of the problem) or whether it is the consequence of a sub-optimal algorithmic design (and/or a loose analysis) is still unknown. Our results fill this theoretical gap by carefully analyzing the effects of non-linearity in the GLB framework through the lens of improved algorithms based on the celebrated optimism-in-face-of-uncertainty principle. At the heart of our approach is the design of refined confidence intervals (sensitive to the variation in variance of the reward signal, acting in GLBs as a proxy for the level of non-linearity), along with a careful treatment of non-linearity where the notion of *local* information is paramount. Our algorithms enjoy exponentially smaller regret than previous approaches and unveil a much more nuanced aspect on the effects of non-linearity. We show that asymptotically such effects fade away as algorithms enter regimes where the reward signal appears (locally) as linear. In most favorable cases this yields a surprising conclusion: some highly non-linear problems turn out to be dramatically easier than their linear counterparts. We also tie non-linearity's detrimental potential to a necessary *burn-in* phase during which any algorithm must uniformly explore its environment. In some worst-case configuration, a high level of non-linearity can significantly impact the length of this transitory phase. Below we sum-up the outline of this part of the dissertation and our different contributions on the study of GLBs.

**Chapter 1. From Multi-Armed Bandits to Generalized Linear Bandits** The goal of this chapter is to formally introduce important notions of sequential decision-making in the

bandit setting and present the state-of-the-art material that will be needed for the rest of the dissertation. We first present the well-known multi-armed bandit setting, which here serves mostly to introduce notations and key quantities (*e.g.* regret, confidence intervals, ...). We use this occasion to provide some basic intuition on the exploration-exploitation trade-off and the optimism in face of uncertainty principle. We examine the limitations of this framework; this leads us to consider the parametric bandit setting. We dedicate some time discussing the linear bandit problem and some key insights from the related literature; namely the optimism principle for parametric bandits and the construction of confidence sets through appropriate tail-inequalities. Again we examine the limitations of this setting; the need to understand richer reward structures and to cover reward distributions of greater practical relevance motivates our study of generalized linear bandits. We review existing approaches for this framework and identify an important weakness; high levels of non-linearity dramatically hinders the performance of existing algorithms. We discuss the challenges in alleviating this behavior and give a brief summary of our approach and contributions. We finish this chapter by detailing some preliminary technical results that will be used throughout the dissertation.

**Chapter 2. Variance-Aware Confidence Sets for GLBs** This chapter aims at deriving improved confidence sets for GLBs. We first provide some intuition on a “candidate” set that fits our requirements, obtained by an asymptotical analysis in a random-design setting. To prove its validity in the general bandit setting we provide a new concentration result based on the theory of self-normalized process: a Bernstein-like tail-inequality for self-normalized martingales. We review its ties, similarities and differences with previous work before applying it to the design of an improved confidence set for GLBs. The main feature of this confidence set resides in its local variance sensitivity which captures the effective level of non-linearity in the environment. This feature is central to the rest of our contributions as it allows for a refined local treatment of the non-linearity. We present several variants as well as an extension to a weighted martingale version which will be used for non-stationary environments. This chapter contains a detailed presentation of results published in:

- Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq (2020a). “Improved Optimistic Algorithms for Logistic Bandits” in Proceedings of the 37th International Conference on Machine Learning (ICML).
- Marc Abeille, Louis Faury and Clément Calauzènes (2021). “Instance-Wise Minimax Optimal Algorithm for Logistic Bandits” in Proceedings of the 24th International Conference of Artificial Intelligence and Statistics (AISTATS).

**Chapter 3. Locality-Sensitive Algorithms for GLBs** In this chapter we apply our confidence set from Chapter 2 to the design of improved self-concordant GLB algorithms. We introduce two algorithms which both rely on this enhanced confidence set but differ in how they enforce optimism. For both algorithms we prove regret upper-bounds that tell a nuanced story about the effects of non-linearity. Such effects are indeed deferred to a second-order term of the regret, tied to a *transitory* regime during which the algorithms search for highly rewarding areas of the action set. The regret suffered during this phase is still negatively impacted by the non-linearity but becomes *negligeable* for large horizons as the algorithms enter a *permanent* regime. Non-linearity then no longer plays a role; only the reward sensitivity around the optimal action does. In addition to such a contrasting conclusion, our algorithms display a dramatic improvement over previous approaches as they enjoy regret upper-bounds that are *exponentially smaller* w.r.t problem-dependent constants. The end of the chapter is dedicated to the Logistic Bandit setting, for which we identify configurations where non-linearity does not even impact

the transitory phase. This ultimately removes its detrimental effects from the regret bounds, even for short horizons. We also derive a problem-dependent regret lower-bound in the Logistic Bandit setting, proving that in the permanent regime our algorithms are *minimax-optimal* w.r.t the dimension  $d$ , the horizon  $T$  and a constant  $\kappa$  that embodies the effects of non-linearity. We conclude this chapter with some numerical experiments illustrating our theoretical findings. This chapter contains and improves results from:

- Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq (2020a). “Improved Optimistic Algorithms for Logistic Bandits” in Proceedings of the 37th International Conference on Machine Learning (ICML).
- Marc Abeille, Louis Faury and Clément Calauzènes (2021). “Instance-Wise Minimax Optimal Algorithm for Logistic Bandits” in Proceedings of the 24th International Conference of Artificial Intelligence and Statistics (AISTATS).

**Chapter 4. Extension to Non-Stationary Environments** The goal of this chapter is to extend the lessons learned in the stationary case to non-stationary environments. We first study the case of piece-wise stationary environments and extend existing approaches from non-stationary linear bandits, based on the *forgetting* of old interactions. To do so we leverage the weighted version of our tail-inequality for self-normalized martingale from Chapter 2, and show that the conclusions from the stationary setting gracefully extend to this non-stationary case. We then turn our attention to environments where non-stationarity is much less structured - for instance environments where the reward model *drifts* over time. In this setting the learning challenges are mixed with tracking difficulties, which hardens the treatment of non-linearity. We propose an algorithm which simultaneously addresses both challenges, however leaving open the question of the optimal scaling w.r.t  $\kappa$  in this general non-stationary setting. This chapter contains and improves results from:

- Yoan Russac, Louis Faury, Olivier Cappé, Aurélien Garivier (2021). “Self-Concordant Analysis of Generalized Linear Bandits under Forgetting” in Proceedings of the 24th International Conference of Artificial Intelligence and Statistics (AISTATS).
- Louis Faury, Yoan Russac, Marc Abeille, Clément Calauzènes (2021). “A Technical Note on Non-Stationary Parametric Bandits: Existing Mistakes and Preliminary Solutions” in Proceedings of the 31st International Conference on Algorithmic Learning Theory (ALT).

**Chapter 5. Summary and Future Work** In this chapter, we summarize our contributions and review directions for future research and remaining open questions.

## Part II. Robust learning for learning under logged bandit feedback

The online setting studied in the previous part is rich in theoretical insights and teachings, however is not always best fitted for real-world applications for several reasons. For instance, closely following the level of exploration recommended by theory is often costly in the short term. While an appropriate amount of exploration is optimal in the long-term, it can also be the source of an immediate loss of revenue (compared to a greedy approach) - often prohibitive for industrial applications which have to comply with short term revenue constraints. Decision-makers are more inclined to have a “hands-on” control of exploration and prefer being able to manipulate it easily, in light of their current constraints. Second, few decision-makers truly face *cold-start* problems like the ones we described in the first part; by leveraging extraneous data and expert knowledge of the problem at hand, they can design reasonable strategies even before their first

interaction with the environment. The main challenge then shifts to the improvement of such policies with data-driven approaches. However, the risk-averse nature of many industrial actors which face bandit problems often speaks in disfavor of off-the-shelf *online* bandit optimization solution. It is frequently desirable in practical applications to ensure some *stability* in the decision-making process; before updating a recommendation engine, one would like to guarantee that this new version will generate at least as much revenue as its predecessors. This requires the development of a *counterfactual* reasoning, and the construction of specific proxy which allows to answer the question: “what revenue could have I hoped to collect if I had acted differently?”. This requires a consequent amount of data, and modifications to the decision-making process cannot be made on-the-fly. A typical illustration of this learning problem arises when an initial *policy* (a strategy for selecting actions) has already been deployed, and that its interactions (*i.e.* the played actions and their outcomes) with the environment have been recorded. The goal of the learner is to leverage such interactions to learn a better strategy.

This problem, often referred to as learning from logged bandit feedback or off-line policy optimization, is the subject of the second part of this dissertation. The main challenge inherent to this learning problem originates from the fact that available observations are biased towards actions favored by the initial policy - the one that was effectively deployed. The current blueprint for addressing this issue is to design *counterfactual* estimators. With only input the recorded feedback, their purpose is to forecast the performance of any other policy as if it was taking the actions by itself. This enables the search for an optimal system without having to exhaustively try-out online all the possible alternative strategies. The main drawback of this approach is that counterfactual estimators come with large variance; directly using them as an absolute criterion to decide for the deployment of a new policy is risky, and can come with severe post-decision surprise (*i.e.* the actual performance of the selected policy is much worse than what was forecasted). Previous work alleviated this issue by resorting to confidence intervals quantifying the variance of counterfactuals estimators, and by penalizing policies for which performance evaluation is subject to high-variance. This however comes at the cost of several *practical* limitations; mainly, the derived objectives for optimizing the initial policy are non-convex and do not undergo stochastic optimization, necessary to handle large logged datasets.

In this part of the dissertation we present an alternative approach which relies on the distributionally robust principle. It allows us to circumvent these practical issues altogether without sacrificing performance. At the heart of this new approach are asymptotic confidence intervals obtained through the distributionally robust optimization (DRO) framework. Under the DRO formulation, one replaces the empirical distribution of observations by the most *pessimistic* distribution over the set of distributions coherent with the empirical observations - the *ambiguity* set. We show that this allows to derive confidence intervals for the performance of any policy, which properties match the challenges of off-line policy evaluation. Furthermore, we prove that they yield better behaved policy optimization objectives; namely *convex* objectives which bear stochastic optimization. We now sum-up the outline of this part and our different contributions.

**Chapter 6. Learning from Logged Bandit Feedback** The goal of this chapter is to introduce the problem of learning from logged feedback. After providing a formal definition, we review existing approaches based on propensity importance re-weighting and its variants. We highlight the issues posed by the high variance of the resulting estimators for risk-averse decision makers. We then focus on describing the counterfactual risk minimization principle as a conservative solution to limit this drawback and avoid post-decision surprise. We review the advantages and limitations of this approach, which we will try to overcome in the following chapter.

**Chapter 7. Distributionally Robust Policy Evaluation and Optimization** We start this chapter by detailing the distributionally robust framework, with a particular focus on robust formulation relying on  $f$ -divergence based ambiguity sets. We discuss how the existing guarantees of robust estimators gracefully meet with the challenges of the learning from logged bandit feedback problem. From these guarantees we also construct asymptotic confidence intervals for offline policy evaluation, computable through solving convex problems. We apply this tool for policy optimization; after providing approximate algorithms for Kullback-Leibler based ambiguity sets, we give a general policy optimization strategy relying on generic  $f$ -divergence based ambiguity sets. This strategy relies on a convex policy optimization objective, which by its *composite* nature is compatible with stochastic optimization. From a practical stand-point, this is a clear improvement over previous approaches. Furthermore, we show that it enjoys good empirical results and competes with more computationally demanding alternatives. This chapter is adapted from the following publications:

- Louis Faury, Ugo Tanielian, Elvis Dohmatob, Elena Smirnova and Flavian Vasile (2020b). “Distributionally Robust Counterfactual Risk Minimization” in Proceedings of the AAAI Conference on Artificial Intelligence.
- Otmane Sakhi, Louis Faury and Flavian Vasile (2020). “Improving Offline Contextual Bandits with Distributional Robustness” in Proceedings of the RecSys Workshop on Reinforcement Learning and Robust Estimators for Recommendation Systems (REVEAL’20).

**Chapter 8 Summary and Future Work** In this chapter, we summarize our contributions and review directions for future research.

## 2 Présentation du contenu de la thèse

Cette thèse présente de récentes contributions au problème de décision séquentielle contre l’incertain, une formalisation importante de nombreuses situations réelles où un agent agit de manière répétée avec un environnement inconnu dans le but de remplir un certain objectif. Sans connaissance précise de l’environnement cet agent se heurte inévitablement à une difficulté fondamentale: comment remplir efficacement son but tout en raisonnant quant aux conséquences (potentiellement corrompues par un bruit inhérent à l’environnement) de ses actions? Derrière cette question presque rhétorique se cache un dilemme fondamental de la prise de décision séquentielle, connu sous le nom de compromis exploration-exploitation. Ce compromis reflète le besoin de l’agent à atteindre une balance délicate entre deux principes contradictoires; augmenter sa connaissance en sondant uniformément l’environnement (*exploration*) tout en utilisant l’information acquise jusqu’ici pour améliorer sa performance (*exploitation*). Construire des agents autonomes qui remplissent ce compromis est central au développement de l’intelligence artificielle, et est au cœur de nombreux domaines de l’apprentissage automatique comme celui des bandits - le sujet de cette thèse. Lorsqu’il apprend d’un signal de type bandit un tel agent s’engage dans un jeu répété avec l’environnement; séquentiellement, il joue une action parmi un ensemble fini et fixe d’actions disponibles et reçoit comme récompense la réalisation d’une variable aléatoire. Le but de l’agent, par exemple, est de maximiser la somme des récompenses perçues dans le temps (minimisation du regret) ou d’identifier la meilleure action après un minimum d’interactions (exploration pure). Malgré son aspect simpliste à première apparence, ce modèle d’apprentissage capture de nombreuses difficultés fondamentales du problème de décision séquentielle contre l’incertain. Plus important encore, il permet un traitement théorique soigné et précis qui a déjà accouché de principes clairs et polyvalents. Cette polyvalence est particulièrement bien illustrée à travers l’avènement des technologies et services de l’Internet,

qui a déclenché le déploiement d'algorithmes de bandit pour des tâches diverses et variées - par exemple l'A/B testing, la recommandation en ligne d'articles de presse ou de publicités.

Lors d'un jeu contre un environnement inconnu, une idée naturelle consiste à estimer et prédire la réponse de l'environnement à chacune des différentes actions. Cette tâche est rendue difficile lorsque cette réponse est stochastique, ce qui oblige à raisonner quant à l'ensemble des environnements *plausibles* qui auraient pu générer (avec haute probabilité) l'ensemble des récompenses observées jusqu'ici. Ce raisonnement a motivé une grande partie de la recherche sur la prise de décision séquentielle contre l'incertain - le problème du bandit ne faisant pas exception.

Cette idée est en effet primordiale pour adresser le problème de minimization du regret grâce au principe d'*optimisme face à l'incertain*, qui afin d'atteindre un compromis efficace entre exploration et exploitation prescrit de jouer l'action qui apparaît la plus gratifiante à la vue de l'ensemble des environnements plausibles. La notion d'environnement plausible peut ici paraître floue au lecteur; dans la littérature des bandits, elle est formalisée par des *intervalles de confiance*. Construire des intervalles de confiance étroits et les utiliser à meilleur escient a été déterminant pour le développement d'algorithmes de bandit efficaces.

Cette thèse poursuit cet effort et présente de récentes contributions dans ce cadre, pour deux aspects distincts et presque orthogonaux du problème d'optimisation bandit; (1) la minimization du regret pour les bandits linéaires généralisés et (2) l'évaluation et l'optimisation hors-ligne de politiques sous un signal bandit historique et enregistré. Nos résultats se basent sur de nouveaux intervalles de confiance, chacun adapté aux difficultés fondamentales d'apprentissage propres à chacune de ces tâches. Ils illustrent comment la construction délicate d'intervalles de confiance adaptés peut simultanément apporter de nouveaux aperçus théoriques et amener à la construction de nouveaux algorithmes dont les garanties théoriques sont améliorées. Ces algorithmes sont plus performants que leur prédécesseurs, et parfois plus simples à déployer. Nos résultats portant sur deux parties fondamentalement éloignées sur le spectre des problèmes de bandit, nous les présentons dans deux parties distinctes.

## Non-linéarité dans les bandit paramétrique: un mal inévitable?

La grande majorité des environnements réels sont fondamentalement non-linéaires, et un effort continu dans le domaine de l'apprentissage automatique a consisté à introduire des modèles de plus en plus expressifs afin de modéliser cette non-linéarité. Cet effort a été particulièrement couronné de succès en ce qui concerne l'apprentissage supervisé, avec l'avènement des réseaux de neurones qui ont indéniablement révolutionné le domaine. Ce succès ne s'est pas généralisé (pour l'instant) aux problèmes de décision séquentielle, où une compréhension plus fine de ces modèles est nécessaire (et manquante) afin d'adresser efficacement des problèmes pratiques. Pour atteindre ce but, une importante pièce manquante du puzzle consiste à comprendre les interactions entre non-linéarité et compromis exploration-exploitation. Cette première partie de la thèse s'efforce de répondre à cette question, à travers l'étude attentive d'une classe générique de bandit paramétriques: les bandits à modèle linéaire généralisé (GLBs, d'après l'acronyme anglais). Cette classe de problèmes offre une formulation permettant d'isoler et d'étudier les effets de la non-linéarité de façon particulièrement nette et précise, tout en restant suffisamment riche pour être pertinente pour un grand nombre de cas pratiques.

Dans un GLB, la récompense moyenne associée à une action (comprise comme un certain vecteur dans un espace Euclidien) est modélisée au travers d'une structure linéaire sous-jacente (un produit scalaire avec un vecteur inconnu représentant l'environnement) à laquelle est appliquée une fonction *non-linéaire*  $\mu$ . Si de tels modèles ont déjà fait l'objet d'études théoriques poussées dans la littérature, la conclusion de ces études est quelque peu décevante puisqu'elle suggère que plus la fonction  $\mu$  est non-linéaire, plus le problème d'apprentissage est dur. En

effet, les garanties de regret des algorithmes existants évoluent toutes linéairement avec une constante  $\kappa$  qui mesure le degré de non-linéarité de  $\mu$ . Cette constante est malheureusement disproportionnée pour de nombreux modèles, ce qui suggère que la non-linéarité est particulièrement préjudiciable à la bonne performance des algorithmes. Savoir si un tel comportement est inévitable (*i.e* il est lié à une difficulté fondamentale du problème d'apprentissage) ou s'il est la conséquence d'une conception sous-optimale des algorithmes est encore une question ouverte. Nos résultats comblent cet écart théorique au travers d'une analyse méticuleuse des effets de la non-linéarité dans le cas GLB et d'algorithmes améliorés basés sur le célèbre principe d'optimisme face à l'incertain. Au coeur de notre approche se trouve la dérivation de nouveaux ensembles de confiances, sensibles à la non-linéarité, ainsi qu'un traitement précis de la non-linéarité où la notion d'information *locale* est primordiale. Nos résultats dévoilent des aspects de la non-linéarité qui sont nettement plus nuancés; en particulier ils prouvent que pour des interactions suffisamment longues, la non-linéarité peut être *bénéfique* au compromis exploration-exploitation. En cela, nos algorithmes bénéficient de garanties théoriques améliorant l'état de l'art par des facteurs exponentiels dans plusieurs cadres d'apprentissage (*e.g* stationnaire et non-stationnaires). D'un autre côté, nous montrons que la non-linéarité peut tout de même impacter de manière négative la durée d'une nécessaire phase d'exploration initiale durant laquelle n'importe quel algorithme doit explorer son environnement uniformément. Nous résumons le déroulé de cette partie ainsi que nos contributions dans les prochaines lignes.

**Chapitre 1. Du Bandit Manchot au Bandit à Modèle Linéaire Généralisé** Le but de ce chapitre est de formellement introduire plusieurs notions importantes du problème de décision séquentielle dans le cadre bandit, et de présenter l'état de l'art nécessaire au reste de cette dissertation. Nous commençons par présenter le problème bien connu du bandit manchot, qui sert ici d'introduction à certaines notations et quantités clés (*e.g* regret, intervalle de confiance). Nous utilisons également cette occasion pour fournir de l'intuition sur le compromis exploration-exploitation et le principe d'optimisme face à l'incertain. Nous discutons ensuite les limitations du bandit manchot; cela nous amène à une importante extension, le bandit paramétrique. Nous dédions une partie importante au bandit linéaire et à certains principes importants émergents de cette littérature; en particulier, le principe d'optimisme dans les bandits paramétriques à travers la construction d'ensembles de confiance par des inégalités de concentration appropriées. À nouveau, nous discutons les limitations de cette modélisation; le besoin d'utiliser des structures de récompense plus riches et de couvrir un plus grand nombre de distributions de récompense nous poussent à nous intéresser aux bandits à modèle linéaire généralisé. Nous présentons ce cadre ainsi que les approches existantes en détails, examinons leurs désavantages ainsi que les défis restants qui seront adressés dans cette thèse. Nous donnons un bref résumé de notre approche et de nos contributions, et terminons ce chapitre en détaillant certains résultats préliminaires qui seront utilisés au travers de cette dissertation.

**Chapitre 2. Ensembles de Confiance Réactifs à la Variance pour les GLBs** Ce chapitre a pour but de dériver de nouveaux et plus étroits ensembles de confiance pour les GLBs. Nous fournissons d'abord de l'intuition sur un ensemble de confiance "candidat" qui remplit nos critères, obtenu par un argument asymptotique et sous un cadre dit de design aléatoire. Pour prouver sa validité dans le cadre général du bandit, nous prouvons un nouveau résultat de concentration basé sur la théorie des processus auto-normalisés: une inégalité de type Bernstein sur les queues de distributions pour les martingales auto-normalisée. Nous discutons ses liens, similarités et différences avec l'état de l'art avant de l'appliquer à la construction d'ensembles de confiance améliorés pour les GLBs. La principale caractéristique de cet ensemble de confiance réside dans sa sensibilité "locale" à la variance de la fonction récompense, qui capture son

niveau de non-linéarité effectif. Nous présentons plusieurs variantes ainsi qu’une extension aux martingales pondérées qui sera utilisée pour traiter des environnements non-stationnaires. Ce chapitre contient une présentation détaillée des résultats publiés dans les articles suivants:

- Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq (2020a). “Improved Optimistic Algorithms for Logistic Bandits” in Proceedings of the 37th International Conference on Machine Learning (ICML).
- Marc Abeille, Louis Faury and Clément Calauzènes (2021). “Instance-Wise Minimax Optimal Algorithm for Logistic Bandits” in Proceedings of the 24th International Conference of Artificial Intelligence and Statistics (AISTATS).

**Chapitre 3. Algorithmes GLBs avec Sensibilité Locale** Dans ce chapitre nous appliquons l’ensemble de confiance dérivé au Chapitre 2 à la construction d’algorithmes performants pour le problème des bandits GLBs dits auto-concordants. Nous introduisons deux algorithmes qui reposent sur ce nouvel ensemble de confiance mais diffèrent par la façon dont ils imposent la notion d’optimisme. Nous prouvons des bornes majorant le regret de chaque algorithme; elles explicitent les véritables effets de la non-linéarité, plus nuancés que l’état de l’art ne le suggère. Ces effets sont délégués à un terme de second ordre du regret, associé à une phase *transitoire* durant laquelle les algorithmes sondent l’environnement pour trouver des zones d’action à forte récompense. Le regret subi pendant cette période est toujours négativement impacté par la non-linéarité, mais devient négligeable pour de larges horizons alors que les algorithmes rentrent dans un régime *permanent*. La non-linéarité arrête dès lors de jouer un rôle et seule la sensibilité locale de la fonction récompense autour de l’action optimale compte. En plus de cette conclusion en fort contraste avec l’état de l’art, nos algorithmes bénéficient d’une amélioration significative par rapport à leur prédécesseurs. Leurs bornes de regret sont en effet réduites par des facteurs *exponentiellement* grands en des constantes liées à la géométrie du problème. La fin de ce chapitre est dédiée à une étude plus poussée du bandit logistique, pour lequel nous identifions des configurations où le rôle de la non-linéarité est davantage réduit puisqu’elle n’impacte même plus la phase transitoire. En cela les effets négatifs de la non-linéarité sont complètement effacés, même pour des horizons courts. Nous dérivons également une borne minorante de regret pour le cas du bandit logistique qui prouve que nos algorithmes sont *minimax* optimaux par rapport à la dimension du problème  $d$ , l’horizon  $T$  ainsi qu’une constante  $\kappa$  qui incarne les effets de la non-linéarité. Nous concluons ce chapitre avec une brève étude numérique illustrant nos résultats théoriques. Ce chapitre contient et améliore des résultats publiés dans:

- Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq (2020a). “Improved Optimistic Algorithms for Logistic Bandits” in Proceedings of the 37th International Conference on Machine Learning (ICML).
- Marc Abeille, Louis Faury and Clément Calauzènes (2021). “Instance-Wise Minimax Optimal Algorithm for Logistic Bandits” in Proceedings of the 24th International Conference of Artificial Intelligence and Statistics (AISTATS).

**Chapitre 4. Extension aux Environnements Non-Stationnaires** Le but de ce chapitre est d’étendre les différents enseignements du cas stationnaire aux environnements non-stationnaires. Nous commençons par l’étude du cas stationnaire par morceaux pour lequel les approches existantes pour le bandit stationnaires sont basées sur le principe d’*oubli*. Pour cela, nous utilisons la version pondérée de notre inégalité de concentration du Chapitre 2 et montrons que les conclusions obtenues dans le cas stationnaire s’étendent à ce type de non-stationnarité. Nous nous

penchons ensuite sur des environnements où la non-stationnarité est moins structurée - par exemple des environnements où la récompense évolue continuellement à travers l’horizon. Dans ce cadre, les difficultés d’apprentissage se mêlent à des défis de suivi, ce qui complique le traitement fin de la non-linéarité. Nous proposons un algorithme qui adresse simultanément les deux difficultés du problème mais laisse ouverte la question de l’optimalité en  $\kappa$  pour ce problème plus général de non-stationnarité. Ce chapitre contient et améliore des résultats publiés dans les articles:

- Yoan Russac, Louis Faury, Olivier Cappé, Aurélien Garivier (2021). “Self-Concordant Analysis of Generalized Linear Bandits under Forgetting” in Proceedings of the 24th International Conference of Artificial Intelligence and Statistics (AISTATS).
- Louis Faury, Yoan Russac, Marc Abeille, Clément Calauzènes (2021). “A Technical Note on Non-Stationary Parametric Bandits: Existing Mistakes and Preliminary Solutions” in Proceedings of the 31st International Conference on Algorithmic Learning Theory (ALT).

**Chapitre 5. Résumé et Perspectives** Dans ce chapitre, nous résumons nos contributions et discutons les directions et perspectives de recherche futures.

### Optimisation hors-ligne de politique: une formulation robuste en distribution

Le problème en ligne étudié dans la partie précédente est riche en aperçus et enseignements théoriques, mais n’est malheureusement pas toujours le plus adapté à des applications pratiques et ce pour plusieurs raisons. Par exemple, suivre finement le niveau d’exploration recommandé par la théorie est souvent coûteux sur le court terme. Bien qu’une proportion appropriée d’exploration soit optimale sur le long terme, elle est aussi la source d’une perte de revenu immédiat (en comparaison à l’approche gloutonne), ce qui est parfois prohibitif dans des applications industrielles où l’agent se doit de répondre à des contraintes de revenu immédiats (par exemple pour rester profitable). Dans de telles applications, les preneurs de décisions sont davantage attirés par un contrôle plus “à la main” de l’exploration et préfèrent être capables de la manipuler facilement et à la lumière de leurs contraintes à court terme. De plus, peu d’agents se retrouvent face à un problème de démarrage “à froid” comme celui décrit dans la partie précédente; en utilisant des données extérieures ou la connaissance fine de ses experts, la plupart des acteurs industriels sont capables de construire des stratégies efficaces avant même leur première interaction avec l’environnement. Le défi principal est alors de réussir à améliorer ces stratégies grâce aux données qu’elles collectent. Là encore, la nature averse au risque de ces acteurs industriels lorsque confrontés à un problème de bandit joue souvent en défaveur des approches de bandits en ligne. Il est en effet désirable lors d’applications pratiques d’assurer une certaine stabilité du processus de prise de décision. Par exemple, avant de déployer une nouvelle version d’un moteur de recommandation il est souhaitable de pouvoir garantir que cette nouvelle version va générer au moins autant de revenu que la précédente. Cela requiert le développement d’un raisonnement dit *contrafactuel* et la construction de critères intermédiaires qui permettent de répondre à la question: “quel revenu aurais-je pu espérer recevoir si j’avais agis différemment?”. En pratique cela demande une quantité de données conséquente et les modifications du processus de prise de décision ne peuvent être effectuées en ligne. Une situation typique illustrant cette problématique à lieu lorsqu’une *politique* (une stratégie selon laquelle des actions sont choisies) a déjà été déployée et que ses interactions avec son environnement ont été enregistrées. Le but de l’agent est alors d’utiliser cette donnée (souvent extrêmement volumineuse) afin de raffiner sa prise de décision et de construire une politique encore plus profitable.

Ce problème, communément appelé apprentissage sous données bandit enregistrées ou optimisation hors-ligne de politique, est le sujet de cette seconde partie de la dissertation. Le principal

défi inhérent à ce problème d'apprentissage est lié au fait que les observations disponibles sont *biaisées* vers les actions jugées comme supérieures par la politique initiale - celle qui a été déployée. Une des stratégies principales pour répondre à cette problématique est de construire des estimateurs *contrafactuels*. Avec comme seule entrée l'interaction enregistrée leur objectif est de dé-biaser cette donnée afin de prédire la performance de n'importe quelle politique. Cela permet de rechercher hors-ligne une stratégie optimale sans avoir à déployer en ligne toutes les politiques potentielles. Le défaut principal de cette approche vient de l'importante variance des estimateurs contrafactuels; les utiliser sans précautions comme un critère de décision absolu est particulièrement risqué puisque l'écart entre la performance prédite et la performance réelle d'une politique peut être important. Certains travaux adressent ce problème en employant des intervalles de confiance quantifiant la variance des estimateurs contrafactuels et en pénalisant les politiques pour lesquelles cette variance est grande. Cette approche connaît cependant des limitations *pratiques*; notamment, les objectifs numériques associés permettant d'optimiser la politique initiale sont non-convexes et ne sont pas compatibles avec des approches d'optimisation stochastiques, nécessaires pour la gestion de grands volumes de données enregistrées.

Dans cette partie de la dissertation nous présentons une approche alternative qui repose sur le principe d'optimisation robuste en distribution. Elle permet de résoudre les problèmes susmentionnés simultanément, tout en conservant les garanties théoriques des approches existantes. À nouveau, le cœur de notre méthode repose sur un nouvel intervalle de confiance (cette fois ci asymptotique) obtenu dans le cadre de l'optimisation robuste en distribution. Sous cette formulation, la distribution empirique des données est remplacée par la distribution la plus pessimiste parmi l'ensemble des distributions cohérentes avec les données. Ce principe permet la construction d'intervalles de confiance sur la performance de nouvelles politiques répondant aux défis de l'optimisation hors-ligne (*e.g* aversion au risque, grande variance, ..). De plus, ces derniers engendrent des objectifs d'optimisation bien plus adaptés à des situations pratiques; ces objectifs sont en effets *convexes* et compatibles avec les méthodes d'optimisation stochastiques classiques. Nous présentons ci-dessous un bref résumé de l'organisation de cette partie.

**Chapitre 6. Apprentissage sous Données Bandit Enregistrées** Le but de ce chapitre est d'introduire formellement le problème d'apprentissage sous données bandit enregistrées. Après une définition formelle du problème, nous discutons les différentes approches existantes se basant sur la re-pondération des données et ses variantes. Nous illustrons les problèmes posés par la variance importante de ces estimateurs pour les agents averses au risque. Nous nous focalisons ensuite sur le principe de minimisation contrafactuelle du risque comme solution à ce problème. Nous exposons les avantages ainsi que les limitations de cette approche que nous nous efforcerons de résoudre dans le chapitre suivant.

**Chapitre 7. Évaluation et Optimisation Robuste en Distribution de Politiques** Nous commençons ce chapitre en détaillant le principe d'optimisation robuste en distribution, avec une attention particulière aux formulations robustes se basant sur des  $f$ -divergences. Nous illustrons comment les garanties des estimateurs robustes répondent aux défis de l'apprentissage sous données bandit enregistrées. Grâce à ces garanties se construisent des intervalles de confiance asymptotiques pour l'évaluation hors-ligne de politique, calculables en résolvant de simples problèmes convexes. Nous appliquons ensuite cet outil à l'optimisation hors-ligne de politique; après avoir détaillé un premier algorithme reposant sur la divergence de Kullback-Leibler et illustrant notre approche, nous donnons une stratégie générale basée sur des  $f$ -divergences génériques. Nos méthodes reposent sur des objectifs convexes et *composites*, compatibles avec les méthodes classiques d'optimisation stochastique. D'un point de vue pratique cela constitue un avantage clair sur les approches existantes, ce qui se traduit par de meilleurs résultats empiriques. Ce

chapitre est adapté des publications suivantes:

- Louis Faury, Ugo Tanielian, Elvis Dohmatob, Elena Smirnova and Flavian Vasile (2020b). “Distributionally Robust Counterfactual Risk Minimization” in Proceedings of the AAAI Conference on Artificial Intelligence.
- Otmame Sakhi, Louis Faury and Flavian Vasile (2020). “Improving Offline Contextual Bandits with Distributional Robustness” in Proceedings of the RecSys Workshop on Reinforcement Learning and Robust Estimators for Recommendation Systems (REVEAL’20).

**Chapitre 8. Résumé et Perspectives** Ce chapitre résume nos contributions et expose des futures directions de recherche.

### 3 Notations

We introduce here some mathematical notations that we use throughout the thesis.

**Functions and Operators.** Let  $f$  and  $g$  be two univariate, real valued functions and  $x, y \in \mathbb{R}$ .

$\dot{f}$	first derivative of $f$ .
$\ddot{f}$	second derivative of $f$ .
$f = \mathcal{O}(g)$	there exists $t_0 \in \mathbb{R}$ and $c \in \mathbb{R}^+$ such that $f(t) \leq c \cdot g(t)$ for all $t \geq t_0$ .
$f = \tilde{\mathcal{O}}(g)$	$f = \mathcal{O}(g \times \text{polylog}(\cdot))$ .
$f = \Omega(g)$	there exists $t_0 \in \mathbb{R}$ and $c \in \mathbb{R}^+$ such that $f(t) \geq c \cdot g(t)$ for all $t \geq t_0$ .
$\delta_x$	Dirac delta function at $x \in \mathbb{R}$ .
$x \vee y$	$\max\{x, y\}$ .
$x \wedge y$	$\min\{x, y\}$ .

For any integer  $d \geq 2$  and any function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  which is Fréchet differentiable at  $x \in \mathbb{R}^d$ :

$\nabla f _x \in \mathbb{R}^d$	Fréchet derivative of $f$ at $x$ .
--------------------------------	------------------------------------

**Linear Algebra.** Let  $d \in \mathbb{N}$ ,  $x \in \mathbb{R}^d$  and  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times d}$  two positive semi-definite  $d \times d$  matrices.

$\ x\ $	$\ell_2$ -norm of $x$ , <i>i.e.</i> $\ x\  = \sqrt{x^\top x}$ .
$\mathcal{B}_d(x, r)$	$d$ -dimensional $\ell_2$ -ball centered at $x$ and of radius $r$
$\mathcal{S}_d(x, r)$	$d$ -dimensional $\ell_2$ -sphere centered at $x$ and of radius $r$ .
$\ x\ _{\mathbf{A}}$	weighted $\ell_2$ -seminorm of $x$ by $\mathbf{A}$ , <i>i.e.</i> $\ x\ _{\mathbf{A}} = \sqrt{x^\top \mathbf{A} x}$ .
$\mathbf{I}_d$	$d \times d$ identity matrix.
$\mathbf{1}_d, \mathbf{0}_d$	$d$ -dimensional vector which entries are all 1 (resp. 0).
$\mathbf{A} \succeq \mathbf{B}$	Löwner ordering of $\mathbf{A}$ and $\mathbf{B}$ , <i>i.e.</i> $\mathbf{A} - \mathbf{B}$ is positive semi-definite.
$\mathbf{A} \succ \mathbf{B}$	$\mathbf{A} - \mathbf{B}$ is positive definite.

**Sets.** Let  $k \in \mathbb{N} \setminus \{0\}$  and  $S$  a set.

$[k]$	the set of integers from 1 to $k$ .
$ S $	cardinality of $S$ .
$S^c$	complement of $S$ (when there is no ambiguity about the ground set).
$\Delta_k$	$k$ -dimensional simplex.

**Randomness.** For an event  $\mathcal{E}$  we denote  $\mathbf{1}(\mathcal{E})$  the indicator function of  $\mathcal{E}$  - *i.e.*  $\mathbf{1}(\mathcal{E}) = 1$  if  $\mathcal{E}$  holds and 0 otherwise. Let  $\{X_s\}_{s=1}^t$  be a sequence of random variables taking values in  $\mathbb{R}^d$ ; then  $\sigma(X_1, \dots, X_t)$  refers to the  $\sigma$ -algebra generated by  $X_1, \dots, X_t$ .

# Contents

---

<b>Introduction</b>	<b>5</b>
1 High level presentation of the thesis . . . . .	5
2 Présentation du contenu de la thèse . . . . .	10
3 Notations . . . . .	16
<b>I Non-Linearity in Generalized Linear Bandits</b>	<b>20</b>
<b>1 From Multi-Armed Bandits to Generalized Linear Bandits</b>	<b>21</b>
1.1 Multi-armed bandits . . . . .	22
1.1.1 Learning problem . . . . .	22
1.1.2 Optimism in face of uncertainty . . . . .	25
1.2 Linearly parametrized bandits . . . . .	27
1.2.1 Learning problem and algorithms . . . . .	27
1.2.2 Limits of the linear model . . . . .	31
1.3 Beyond linearity: Generalized Linear Bandits . . . . .	31
1.3.1 Learning problem . . . . .	32
1.3.2 Quantifying non-linearity. . . . .	33
1.3.3 Linearization approach . . . . .	35
1.3.4 Limitations and challenges . . . . .	38
1.4 Our approach: a self-concordant analysis for Generalized Linear Bandits. . . . .	40
1.4.1 Setting: self-concordant GLBs . . . . .	40
1.4.2 Brief summary of contributions . . . . .	40
1.4.3 Notations and first technical results . . . . .	41
<b>2 Variance-Aware Confidence Sets for Generalized Linear Bandits</b>	<b>48</b>
2.1 Towards improved confidence sets . . . . .	49
2.2 Bernstein-like tail-inequality for self-normalized martingales . . . . .	50
2.2.1 Result and discussion . . . . .	50
2.2.2 Proof of the main theorem . . . . .	51
2.3 Application to the design of confidence-sets for GLBs . . . . .	54
2.3.1 Confidence set . . . . .	54
2.3.2 A convex relaxation . . . . .	56
2.4 An extension to weighted self-normalized martingales . . . . .	57
<b>3 Locality-Sensitive Algorithms for GLBs</b>	<b>65</b>
3.1 Tighter and local exploration bonuses . . . . .	66

3.1.1	Algorithm and regret upper-bound . . . . .	66
3.1.2	Discussion . . . . .	67
3.1.3	Proof of the regret bound . . . . .	68
3.2	The parameter-search alternative . . . . .	71
3.2.1	Algorithm and regret upper-bound . . . . .	71
3.2.2	Discussion . . . . .	72
3.2.3	Proof of the regret bound . . . . .	72
3.2.4	A tractable algorithm: convex relaxation . . . . .	73
3.3	Non-linearity and transitory regret . . . . .	74
3.3.1	Transitory regret and detrimental arms . . . . .	75
3.3.2	Non-linearity in LogB: a blessing? . . . . .	77
3.4	Optimality of the permanent regret . . . . .	79
3.4.1	Regret lower-bound . . . . .	79
3.4.2	Proof of the lower-bound . . . . .	80
3.5	Numerical simulations . . . . .	84
<b>4</b>	<b>Extensions to Non-Stationary Environments</b>	<b>103</b>
4.1	The learning problem . . . . .	104
4.1.1	Setting and non-stationary metrics . . . . .	104
4.1.2	Forgetting strategies . . . . .	104
4.2	Piece-wise stationary environments . . . . .	105
4.2.1	Confidence sets . . . . .	105
4.2.2	Algorithms and regret upper-bounds . . . . .	107
4.2.3	Sketch of proof for the sliding-window strategy . . . . .	108
4.3	Drifting environments . . . . .	110
4.3.1	Motivation and Challenges . . . . .	110
4.3.2	A linearization algorithm . . . . .	112
4.3.3	Sketch of proof . . . . .	114
<b>5</b>	<b>Summary and Future Work</b>	<b>121</b>
5.1	Summary . . . . .	121
5.2	Remaining challenges and open questions . . . . .	123
5.2.1	Simultaneous statistical and computational efficiency . . . . .	123
5.2.2	Best-arm identification . . . . .	123
5.2.3	Open question: optimality of forgetting mechanisms . . . . .	124
<b>II</b>	<b>Offline Bandits: Robust Policy Evaluation and Optimization</b>	<b>125</b>
<b>6</b>	<b>Learning from Logged Bandit Feedback</b>	<b>126</b>
6.1	Motivation and formalization . . . . .	127
6.1.1	The ad-placement case . . . . .	127
6.1.2	The learning problem . . . . .	128
6.2	Counterfactual estimation . . . . .	129
6.2.1	Counterfactual estimators . . . . .	129
6.2.2	Confidence intervals . . . . .	131
6.3	The Counterfactual Risk Minimization principle . . . . .	132
6.3.1	A variance-regularized objective . . . . .	133
6.3.2	Limitations . . . . .	135

---

<b>7</b>	<b>Distributionally Robust Policy Evaluation and Optimization</b>	<b>138</b>
7.1	Distributionally Robust Optimization . . . . .	139
7.1.1	High-level presentation . . . . .	139
7.1.2	Asymptotic guarantees of generalized empirical likelihood estimators . . .	141
7.2	Application to offline policy evaluation . . . . .	142
7.2.1	Asymptotic confidence interval on policy risk . . . . .	142
7.2.2	Computing the confidence interval . . . . .	142
7.2.3	Numerical simulations . . . . .	144
7.3	Distributionally robust policy optimization algorithms . . . . .	145
7.3.1	An approximation for Kullback-Leibler ambiguity sets . . . . .	146
7.3.2	Robust policy optimization for coherent $f$ -divergences . . . . .	150
7.3.3	Extensions . . . . .	152
<b>8</b>	<b>Conclusion</b>	<b>154</b>

## Part I

# Non-Linearity in Generalized Linear Bandits

# From Multi-Armed Bandits to Generalized Linear Bandits

---

The goal of this chapter is to formally introduce important notions of sequential decision-making in the bandit setting and present the state-of-the-art material that will be needed for the rest of the dissertation. We first present the well-known multi-armed bandit setting, which here serves mostly to introduce notations and key quantities (*e.g* regret, confidence intervals, ..). We use this occasion to provide some basic intuition on the exploration-exploitation trade-off and the optimism in face of uncertainty principle. We examine the limitations of this framework; this leads us to consider the parametric bandit setting. We dedicate some time discussing the linear bandit problem and some key insights from the related literature (*e.g* the construction of confidence sets through appropriate tail-inequalities). We examine the limitations of this setting; the need to understand richer reward structures and to cover reward distributions of greater practical relevance motivates our study of generalized linear bandits. We review existing approaches for this framework and identify an important weakness; high levels of non-linearity dramatically hinders the performance of existing algorithms. We discuss the challenges in alleviating this behavior and give a brief summary of our approach and contributions. We finish this chapter by detailing some preliminary technical results that will be use throughout the dissertation.

## Outline

---

1.1	Multi-armed bandits . . . . .	22
1.1.1	Learning problem . . . . .	22
1.1.2	Optimism in face of uncertainty . . . . .	25
1.2	Linearly parametrized bandits . . . . .	27
1.2.1	Learning problem and algorithms . . . . .	27
1.2.2	Limits of the linear model . . . . .	31
1.3	Beyond linearity: Generalized Linear Bandits . . . . .	31
1.3.1	Learning problem . . . . .	32
1.3.2	Quantifying non-linearity. . . . .	33
1.3.3	Linearization approach . . . . .	35
1.3.4	Limitations and challenges . . . . .	38
1.4	Our approach: a self-concordant analysis for Generalized Linear Bandits. . . . .	40
1.4.1	Setting: self-concordant GLBs . . . . .	40
1.4.2	Brief summary of contributions . . . . .	40
1.4.3	Notations and first technical results . . . . .	41

---

## 1.1 Multi-armed bandits

The stochastic Multi-Armed Bandit (MAB) framework describes a sequential-decision making problem with a finite and static set of actions (also called arms). Each action has its own intrinsic reward which can be observed when the action is played, and only up to some noise. The learning agent enters a repeated game with the environment and aims to maximize her expected cumulative pay-off without *a-priori* knowledge of the different actions' rewards. To fulfill this objective she must use her current knowledge of the world to play highly rewarding actions (*exploitation*) and simultaneously increase her global knowledge of the environment by refining her evaluation of poorly estimated actions (*exploration*). Such a situation is ubiquitous in real-world problems and its study has a long history; it was first formalized to study clinical trials and since then has also been applied to various fields such as allocation in finance, website optimization, web-routing, ... From a theoretical standpoint, it has established itself as a prominent framework in the sequential decision-making literature. This can reasonably be attributed to its simplicity (in the positive sense of the term); it offers an immaculate setting to study the exploration-exploitation dilemma, stripped of any unessential complexity. As a result, there exists a substantial literature on the topic and numerous different approaches to solve this problem. The goal of this section is not to provide an exhaustive review on MABs and in-depth description of existing algorithms; instead we will use it to introduce some key concepts (*e.g.* exploration-exploitation dilemma, regret, optimism). This will be useful to introduce and discuss parametric bandits (such as the linear bandit) which are the main focus of this part.

### 1.1.1 Learning problem

**Setting.** We describe here the stochastic MAB learning problem (see *e.g.* [Auer et al., 2002](#)). Let  $K \in \mathbb{N}$  and denote  $\mathcal{A} = [K]$  the available actions. Each action  $k \in \mathcal{A}$  has an associated reward distribution  $\nu_k$  which is assumed to have a finite mean denoted  $\mu_k$ . At each round  $t \geq 1$  the agent plays an action  $a_t \in \mathcal{A}$  and receives a stochastic reward  $r_{t+1}$  drawn according to the distribution  $\nu_{a_t}$ . The agent's decision is taken in an adaptive fashion: her behavior can be represented by a policy  $\pi: \mathcal{F}_t \rightarrow \Delta_K$  where  $\mathcal{F}_t = \sigma(\{a_s, r_{s+1}\}_{s=1}^{t-1})$  encodes the information obtained so far.<sup>1</sup> Of course, the set of distributions  $\nu = \{\nu_k\}_{k \in [K]}$  is unknown to the agent and therefore so is the optimal action (in terms of expected pay-off). We denote this action  $a_\star$  and its associated expected reward  $\mu_\star$ . Formally:

$$a_\star = \arg \max_{k \in [K]} \mu_k \quad \text{and} \quad \mu_\star = \max_{k \in [K]} \mu_k = \mu_{a_\star} .$$

The goal of the agent is to maximize her cumulative expected reward  $\mathbb{E}[\sum_{t=1}^T \mu_{a_t}]$  where  $T$  is the length of the game; equivalently, she tries to minimize her expected *cumulative pseudo-regret* which is the difference in expectation between the payoff gathered by playing the optimal arm  $a_\star$  at each round and the cumulative rewards that were actually observed. Formally:

$$\text{Regret}_\nu^\pi(T) := T\mu_\star - \mathbb{E} \left[ \sum_{t=1}^T \mu_{a_t} \right] ,$$

By introducing the sub-optimality gaps  $\Delta_k := \mu_\star - \mu_k$  and  $T_k(t) := \sum_{s=1}^t \mathbb{1}(a_s = k)$  the number of times that action  $k$  was played, the regret can be re-written as:

$$\text{Regret}_\nu^\pi(T) = \mu_\star \sum_{k=1}^K \mathbb{E}[T_k(T)] - \mathbb{E} \left[ \sum_{k=1}^K \mu_k T_k(T) \right] = \sum_{k=1}^K \Delta_k \mathbb{E}[T_k(T)] . \quad (1.1)$$

<sup>1</sup>This definition of a policy is informal to keep the discussion concise. See ([Lattimore and Szepesvári, 2020](#), Section 4.6) for a rigorous measure-theoretic definition of a policy.

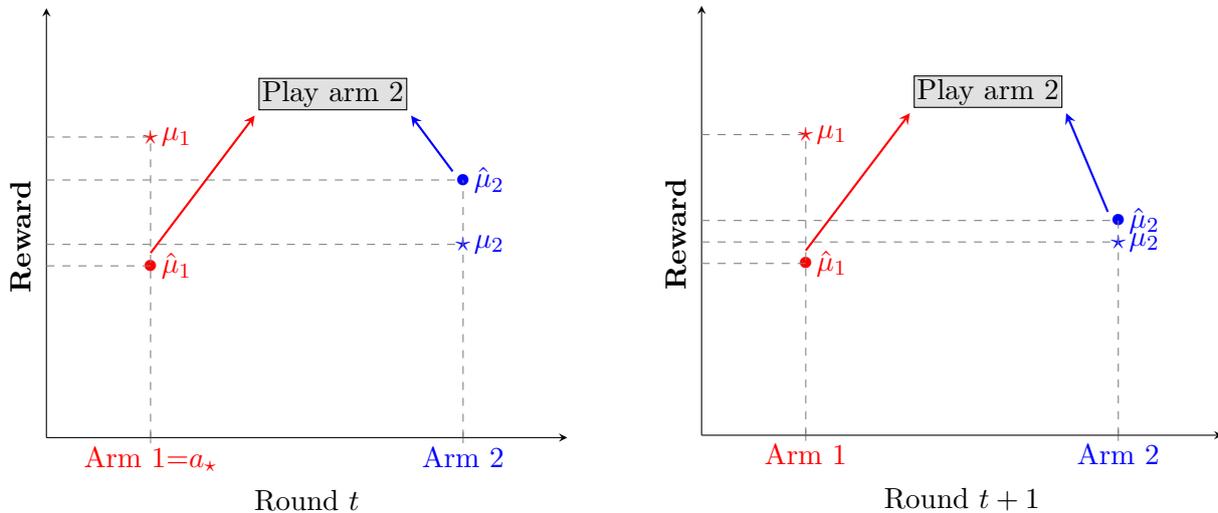


Figure 1.1: Illustration of the behavior of the greedy policy. The optimal arm is the first arm since  $\mu_1 > \mu_2$ . In this case, the random realization of the rewards have led to the following situation: at round  $t$ , the estimate  $\hat{\mu}_1$  of the best arm is smaller than the estimate  $\hat{\mu}_2$  of the second arm. A greedy policy will therefore pick the second arm, and suffer an instantaneous regret  $\mu_1 - \mu_2 > 0$ . The agent refines its knowledge about the second arm and the estimate  $\hat{\mu}_2$  improves. The situation for the first arm does not change and it remains poorly estimated. Under a greedy strategy this might never change; the first arm will possibly never be played since  $\hat{\mu}_1 < \mu_2$ .

Broadly speaking, the goal of the agent is to follow a policy  $\pi$  such that the regret is *sub-linear* in  $T$ . Before discussing algorithms that achieve this objective, we need to discuss the main challenge they need to address: the exploration-exploitation dilemma.

**Exploration-exploitation dilemma.** The agent does not know the distributions  $\{\nu_k\}_{k=1}^K$  and therefore is ignorant of the highest mean  $\mu_*$  or its associated action  $a_*$ . She builds her knowledge of the environment through the actions  $a_1, a_2, \dots$  she played and the rewards  $r_2, r_3, \dots$  she received. She therefore faces an exploration-exploitation dilemma, which involves a careful balance between two *conflicting* objectives: play the most rewarding action given her current knowledge to minimize the instantaneous regret (*exploitation*) and play poorly estimated arms to refine her knowledge and collect information that may be useful in the future (*exploration*). Focusing only on exploration trivially leads to a linear regret (since it essentially involves playing all actions uniformly) while pure exploitation can lead to the same result if the expected reward associated with the optimal arm is poorly estimated (which would lead a greedy algorithm to always play a sub-optimal arm and henceforth suffer linear regret). We illustrate in Fig. 1.1 this phenomenon in the case of a two-arm bandit problem.

**Lower-bounds.** Before introducing algorithmic principles that address this trade-off, we can highlight the hardness of the problem (from a learning-theoretic point of view) by recalling existing lower-bounds. The lower-bound of (Lai and Robbins, 1985) characterizes, in the asymptotic regime, the number of suboptimal arm-pulls that any *consistent* policy must perform. Informally, a policy is called consistent for a class  $\mathcal{V}$  of distributions if for any  $\nu \in \mathcal{V}^K$  it enjoys at

worst a logarithmic asymptotic regret. Formally,  $\pi$  is consistent for  $\mathcal{V}$  if:

$$\forall \nu \in \mathcal{V}^K, \forall \alpha > 0, \lim_{T \rightarrow \infty} \frac{\text{Regret}_\nu^\pi(T)}{T^\alpha} = 0. \quad (1.2)$$

**Theorem 1.1.1** (Lai and Robbins (1985)). *Let  $\mathcal{V}$  be a class of distributions parametrized by their means, and let  $\pi$  be a consistent policy for  $\mathcal{V}$ . Then under mild conditions over  $\mathcal{V}$ , for every  $\nu \in \mathcal{V}^K$  one has:*

$$\liminf_{T \rightarrow \infty} \frac{\text{Regret}_\nu^\pi(T)}{\log(T)} = \sum_{k=1}^K \frac{\Delta_k}{\text{KL}(\nu_k || \nu_{a_*})}.$$

where  $\text{KL}(\nu_k || \nu_{a_*})$  is the Kullback-Leibler divergence between the reward's distributions of the  $k^{\text{th}}$  arm and the optimal arm.

**Theorem 1.1.1** states that any *consistent* policy (i.e an allocation strategy satisfying Eq. (1.2)) must suffer, asymptotically, a logarithmic regret. In other words if  $\pi$  is asymptotically performant on *all* instances  $\nu$ , then it must explore enough to tell those instances apart and therefore its regret cannot be bounded (whatever the considered instance  $\nu$ ). This necessary level of exploration is quantified by **Theorem 1.1.1**: together with Eq. (1.1) it informs us that asymptotically, each suboptimal arm must be played proportionally to  $\log(T)$  times. The exact proportion is a function of the sub-optimality gap and the discrepancy (measured by a Kullback-Leibler divergence) between the considered arm's reward distribution and the distribution of highest mean. To gain some intuition on this constant, let us consider the following example; for all  $k \in [K]$ , let  $\nu_k$  be a Bernoulli distribution of mean  $\mu_k$ . In such a case we have that  $\text{KL}(\nu_k || \nu_{a_*}) \approx \Delta_k^2$  and the number of sub-optimal pulls prescribed by **Theorem 1.1.1** becomes  $1/\Delta_k$ . Therefore the smaller (resp. higher) the sub-optimal gap, the more (resp. less) a sub-optimal arm must be played. This makes intuitive sense; if a sub-optimal arm  $k$  is associated to a distribution  $\nu_k$  which mean  $\mu_k$  is close to  $\mu_*$ , then it must be played quite often so the agent can actually tell them apart with high confidence, and vice-versa.

It is natural to wonder how such conclusions transfer to the finite-time setting and without restricting the considered class of policies. It is actually possible to exhibit collections of *hard* MAB instances, on which no algorithm can achieve uniformly good finite-time performance. Such constructions are at the origins of *problem-independent* lower-bounds, which essentially quantify the *worst-case* hardness of the MAB problem. We recall here a lower-bound appearing in Lattimore and Szepesvári (2020) which essentially shows that for any policy we can find a Gaussian MAB problem such that the regret of the considered policy is  $\Omega(\sqrt{KT})$ .

**Theorem 1.1.2** (Theorem 15.2 of Lattimore and Szepesvári (2020)). *Let  $T \geq K$  and  $\mathcal{V}$  be the class of normal distributions with unit variance. For any policy  $\pi$  it exists  $\nu \in \mathcal{V}^K$  such that:*

$$\text{Regret}_\nu^\pi(T) \geq c\sqrt{KT}$$

where  $c > 0$  is a universal constant - i.e independent of the problem.

Such results set the bar for what one can expect in the worst-case of the strategies designed to solve the MAB problem. A worthwhile goal is therefore to design policies which worst-case performances effectively matches this lower-bound. This is the topic of the next section, which describes the optimism-in-face-of-uncertainty principle, a popular heuristic at the root of many strategies fulfilling this goal.

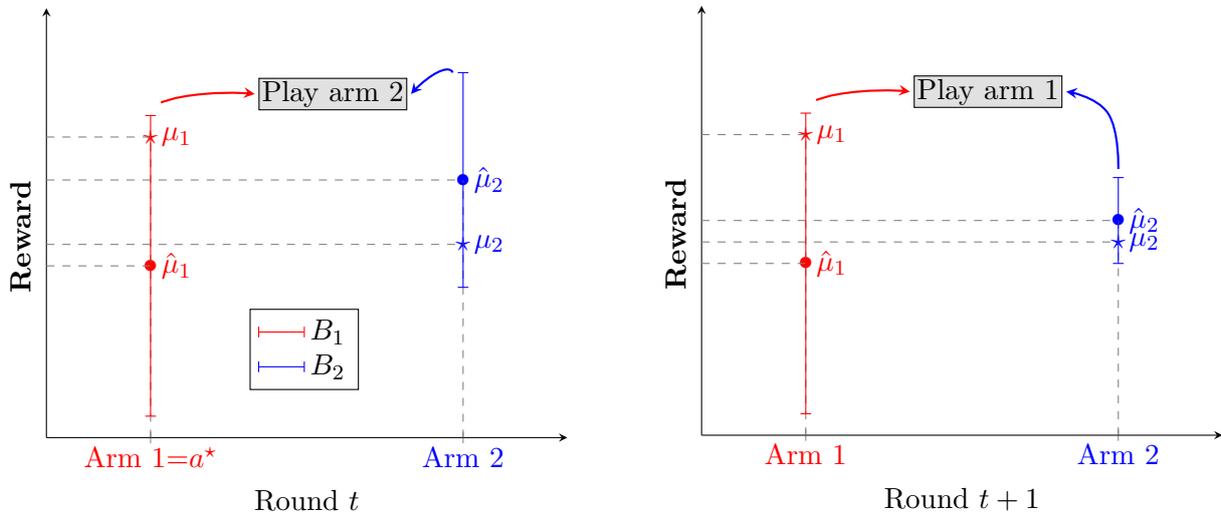


Figure 1.2: Illustration of the OFU principle, in opposition with the behavior of the greedy strategy as illustrated in Fig. 1.1. The errors bars represent the confidence intervals for the arms. At round  $t$  the OFU principle prescribes playing action 2 as the confidence intervals suggests that it could be the most rewarding action. After observing the associated reward, the estimation of the arm's expected reward improves and the width of the confidence interval decreases. At round  $t + 1$  the OFU principles now recommends arm 1 which is the optimal arm. Compared to the greedy policy, optimism allows to balance the exploration-exploitation trade-off and ensures small cumulative regret.

### 1.1.2 Optimism in face of uncertainty

**Intuition.** The main idea behind the optimism-in-face-of-uncertainty (OFU) principle is to play *greedily* according to the most optimistic of *plausible* environments. More precisely, it prescribes considering all the environments which are coherent with the rewards observed so far (*i.e.* that are plausible) and play the best action of the environment which has the best possible payoff. By doing so, one either selects the best action of the true environment (which comes at zero regret) or select a sub-optimal action that carries valuable information for the future. In the MAB setting the set of plausible environments can be quantified through confidence intervals  $B_k^t$  for the true means  $\{\mu_k\}_{k=1}^K$ , *i.e.* such that at each round  $t \geq 1$ :

$$\forall k \in [K], \mu_k \in B_k^t \text{ with high probability.}$$

An optimistic algorithm plays the action which has the highest upper confidence bound:

$$\text{play } a_t = \arg \max_{k \in [K]} \sup_{\tilde{\mu}_k \in B_k^t} \tilde{\mu}_k. \quad (1.3)$$

Naturally the width of the confidence intervals  $\{B_k^t\}_{k=1}^K$  must gracefully degrades in time so the set of plausible environments eventually concentrates around the true environment - this is the topic of the following paragraph. We illustrate the OFU principle in Fig. 1.2.

**Upper Confidence Bounds algorithms.** Algorithms that plan according to the OFU principle have usually been called Upper-Confidence Bounds (UCB) algorithms. This principle has bred a myriad of algorithms which differ conceptually by the confidence interval they rely on to quantify the set of plausible environments. A short (far from exhaustive) list includes: UCB1

**Algorithm 1** UCB1 (Auer et al., 2002)

---

**input:** Arms  $\{1, \dots, K\}$ , failure level  $\delta$ .

**for**  $k \in [K]$  **do**

    Play arm  $k$ , observe reward.  $\triangleright$  *initialization*

**end for**

  Update empirical means.

**for**  $t \in [K + 1, T]$  **do**

    Play the arm  $a_t$  with highest upper-confidence bound (see Eq. (1.4)).  $\triangleright$  *planning*

    Observe reward  $r_{t+1}$ , update empirical means and confident upper-bounds.  $\triangleright$  *learning*

**end for**

---

(Auer et al., 2002), UCB-v (Audibert et al., 2009) and KL-UCB (Cappé et al., 2013). We describe here the UCB1 of Auer et al. (2002) for the sake of illustration and refer the interested reader to (Lattimore and Szepesvári, 2020, Part II) for an in-depth introduction and discussion on UCB algorithms for the MAB problem. To ease the exposition we will assume that the distributions  $\{\nu_k\}_{k \in [K]}$  all have support in  $[0, 1]$  so the associated random variables are all 1/2-sub-Gaussian (see Definition A.2 and Lemma A.3). At any round  $t$  and for a given confidence level  $\delta \in (0, 1]$  the UCB algorithm follows the policy:

$$\text{play } a_t = \arg \max_{k \in [K]} \left\{ \hat{\mu}_k^t + \sqrt{\frac{\log(1/\delta)}{2T_k(t-1)}} \right\}, \quad (1.4)$$

where  $\hat{\mu}_k^t = \sum_{s=1}^{t-1} \mathbb{1}(a_s = k) r_{s+1} / T_k(t-1)$  is the empirical mean of  $\mu_k$ . This is motivated by the Chernoff-Hoeffding concentration inequality (see Lemma A.4) which states that (up to the fact that  $T_k(t-1)$  is a random variable, which will require a slight refinement - *e.g* an union bound):

$$\mathbb{P} \left( \mu_k \leq \hat{\mu}_k^t + \sqrt{\frac{\log(1/\delta)}{2T_k(t-1)}} \right) \leq 1 - \delta \text{ for all } k \in [K].$$

The exploration bonus over the empirical mean that appears in Eq. (1.4) can therefore be written as the upper-bound of a confidence interval. This feature is at the heart of the proof's behind the theoretical guarantees of UCB1.

**Theorem 1.1.3** (Regret of UCB1, Theorems 7.1 and 7.2 of (Lattimore and Szepesvári, 2020)). *Let  $\mathcal{V}$  be the family of distributions with support in  $[0, 1]$ . Given any horizon  $T$ , setting  $\delta = 1/T^2$  ensures that for a given bandit instance  $\nu \in \mathcal{V}^K$  the regret of UCB1 satisfies:*

$$\text{Regret}_\nu(T) \leq 3 \sum_{k=1}^K \Delta_k + \sum_{k: \Delta_k \neq 0} \frac{16 \log(T)}{\Delta_k}.$$

Furthermore, the worst-case regret of UCB1 satisfies;

$$\max_{\nu \in \mathcal{V}^K} \text{Regret}_\nu(T) = \mathcal{O} \left( \sqrt{KT \log(T)} \right).$$

Note that the worst-case regret matches (up to logarithmic term) with the lower-bound of Theorem 1.1.2. Refining the concentration bounds (*e.g* by relying on higher order moments of the distribution (Audibert et al., 2009) or on its complete description (Cappé et al., 2013)) leads to tighter confidence intervals and improved regret upper-bounds.

## 1.2 Linearly parametrized bandits

**From MAB to parametric bandits.** From a theoretical standpoint, the main interest behind the MAB framework is to provide a minimalistic framework to study and address the exploration-exploitation dilemma. Its simplicity surely is the reason for its popularity, but comes at a price as it leaves many questions open when it comes to more realistic sequential decision-making problems. For instance, it is common for a practitioner to have the *a-priori* knowledge that some actions are *similar* and will roughly have the same payoff. For instance, a recommendation engine company knows that the appetite of a given user for different pairs of shoes is sensibly the same, but is independent of its interest in plasma screens. In the MAB setting that we have described so far the knowledge of interrelated payoffs is discarded and the arms are all treated independently. Intuitively this feels like a miss; allowing to transfer the information brought by playing an arm to similar arms should ease the exploration-exploitation task as not every arm needs to be independently explored. This calls for bandit settings with additional structure and in particular motivates the study of the so-called *parametric* bandits. In this setting actions are represented by vectors of a given metric space  $E$ , with the idea that the closer two actions are, the more similar their reward. Each arm's reward distribution is assumed to be tied to a shared parameter  $\theta_\star \in \Theta$  and dictated by a common function  $f : E \times \Theta \rightarrow \mathbb{R}$ . Formally;

$$\forall a \in \mathcal{A}, \quad \mathbb{E}[r|a] = f(a, \theta_\star).$$

The function  $f$  is assumed known by the agent while  $\theta_\star$  is the unknown quantity of interest. The same way a MAB algorithm ultimately tries to estimate the vector of means  $\{\mu_k\}_{k \in [K]}$ , a parametric bandit algorithm will try to learn  $\theta_\star$  while balancing exploration and exploitation. For this reason it is reasonable to study cases where *estimation* or learning is possible, which typically restricts the class of function of the reward function  $f$ . A prototypical instance of such a model is the linear bandit (LB) where  $f$  is assumed to be the scalar product over  $E$ :

$$\forall a \in \mathcal{A}, \quad \mathbb{E}[r|a] = a^\top \theta_\star.$$

The linear bandit has been extensively studied in the literature (e.g. Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011; Abeille and Lazaric, 2017, and references therein). It compactly sums-up some key challenges of parametric bandits and is the topic of the rest of this section.

**Remark 1.2.1** (Other reasons to study parametric bandits). *We motivated the linear bandit setting by the need to share information across similar arms. It also allows to incorporate extraneous contextual information which may impact the payoff of the actions (e.g some user features in a recommendation task). Indeed, the problem remains fundamentally the same if:*

$$\forall a \in \mathcal{A}, \forall x \in \mathcal{X}, \quad \mathbb{E}[r|a, x] = \phi(x, a)^\top \theta_\star.$$

where  $\mathcal{X}$  denotes the space of possible contextual information and  $\phi$  a given joint feature map. Finally we will see that parametric bandits allow to handle infinite (e.g continuous) actions sets which is not possible under a MAB description.

**Remark 1.2.2** (MAB as a parametric bandit). *The MAB setting is a special case of the more general parametric setting. Indeed, if  $\mathcal{A}$  lies on an orthogonal basis of some  $\mathbb{R}^K$ , the associated LB problem is actually a  $K$ -arm MAB with mean vector  $\theta_\star$ .*

### 1.2.1 Learning problem and algorithms

The goal of this section is to present existing work on the LB and insist on some key component of LB algorithms (e.g confidence sets) that will be useful for the following chapters.

**Reward model.** The action set  $\mathcal{A}$  is potentially infinite yet embedded in a  $d$ -dimensional Euclidean space. At each round  $t$ , the agent plays  $a_t \in \mathcal{A}$  and receives a reward  $r_{t+1}$  such that:

$$r_{t+1} = a_t^\top \theta_\star + \eta_{t+1}, \quad (1.5)$$

where  $\theta_\star \in \mathbb{R}^d$  and  $\eta_{t+1}$  is a zero-mean noise. More precisely, let:

$$\mathcal{F}_t := \sigma(a_1, r_2, a_2, \dots, r_t, a_t),$$

be the  $\sigma$ -field encoding all the information available before  $r_{t+1}$  is observed. It will be assumed that  $\mathbb{E}[\eta_{t+1}|\mathcal{F}_t] = 0$ , or equivalently that  $\mathbb{E}[r_{t+1}|\mathcal{F}_t] = a_t^\top \theta_\star$ . The LB analysis is conducted under the following assumption, which essentially ensures that the mean rewards are bounded.

**Assumption 1.2.1** (Bounded decision set). *For any  $a \in \mathcal{A}$  we have  $\|a\| \leq 1$ .<sup>2</sup> Furthermore, the unknown parameter  $\theta_\star$  satisfies  $\|\theta_\star\| \leq S$  where  $S$  is known.*

As in the MAB setting the noise  $\eta$  must satisfy basic distributional properties in order for estimators to concentrate and  $\theta_\star$  to be learnable.

**Assumption 1.2.2** (Sub-Gaussian noise). *Conditionally to  $\mathcal{F}_t$  the noise  $\eta_{t+1}$  is sub-Gaussian with variance-proxy  $\sigma^2$ :*

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}[\exp(\lambda \eta_{t+1})|\mathcal{F}_t] \leq \exp(\lambda^2 \sigma^2 / 2) \quad a.s.$$

The cumulative pseudo-regret suffered by an agent when following a policy  $\pi$  writes:

$$\text{Regret}_{\theta_\star}^\pi(T) := T a_\star(\theta_\star)^\top \theta_\star - \sum_{t=1}^T a_t^\top \theta_\star,$$

where  $a_\star(\theta_\star) := \arg \max_{a \in \mathcal{A}} a^\top \theta_\star$  denotes the optimal action.

**Remark 1.2.3** (Time-varying action sets). *In all generality, the action set can be time varying, for instance to account for the contextual nature of the sequential decision-making problem at hand. This does not change the overall message we try to highlight in this section so we will stick to constant action sets.*

**Lower-bound.** Before discussing algorithms for LB we first highlight the inherent hardness of the problem by recalling the following regret lower-bound.

**Theorem 1.2.1** (Theorem 24.2 of [Lattimore and Szepesvári \(2020\)](#)). *Let  $d < 2T$  and  $\mathcal{A} = \mathcal{B}_2(0, 1)$ . For any policy  $\pi$  it exists  $\theta \in \mathbb{R}^d$  such that:*

$$\mathbb{E}[\text{Regret}_\theta^\pi(T)] = \Omega(d\sqrt{T}).$$

Similarly to the MAB setting this lower-bound states that the worst-case regret grows at least as fast as  $\sqrt{T}$ , and linearly with the dimension  $d$  of the problem.

<sup>2</sup>This assumption can be easily replaced by  $\|a\| \leq A$ , or enforced by pre-scaling the action set.

**Sketching an optimistic approach.** Similarly to the MAB setting, an important principle is to estimate  $\theta_*$  based on past interactions. Provided an estimator  $\hat{\theta}_t$  of  $\theta_*$  we obtain estimates  $a^\top \hat{\theta}_t$  for the mean reward of each arm  $a \in \mathcal{A}$  - the ground truth being  $a^\top \theta_*$ . Mimicking the optimistic approach for MAB we can shoot for the design of exploration bonuses  $\varepsilon_t(\cdot)$  to build upper confidence bounds for  $a^\top \theta_*$ , *i.e* such that:

$$\forall a \in \mathcal{A}, a^\top \theta_* \leq a^\top \hat{\theta}_t + \varepsilon_t(a) \quad \text{w.h.p.}$$

Equivalently, we can ask for the exploration bonus to upper-bound the *prediction error*  $\Delta_t(a) = |a^\top(\theta_* - \hat{\theta}_t)|$  of the estimator  $\hat{\theta}_t$ . In other words:

$$\forall a \in \mathcal{A}, \varepsilon_t(a) \geq \Delta_t(a) \quad \text{w.h.p.} \quad (1.6)$$

Assuming for now that we can design such exploration bonuses, we can play greedily according to the highest upper confidence bound:

$$\text{play } a_t \in \arg \max_{a \in \mathcal{A}} a^\top \hat{\theta}_t + \varepsilon_t(a) .$$

This action selection mechanism ensures that  $\text{Regret}_{\theta_*}(T) \leq 2 \sum_{t=1}^T \varepsilon_t(a_t)$  (with high probability) which we can expect to be sub-linear if the exploration bonuses vanish fast enough. From [Eq. \(1.6\)](#) we can see that such bonuses can be obtained by uniformly bounding the prediction error of  $\hat{\theta}_t$ , or equivalently by controlling the deviation of  $\hat{\theta}_t$  from  $\theta_*$  in every direction. This calls for the design of *confidence sets* for  $\theta_*$ .

**Confidence set for LB.** In the LB setting, the regularized least-square estimator is a good candidate for the design of confidence sets for it comes with strong theoretical guarantees. It is defined as follows:

$$\hat{\theta}_t := \arg \min_{\theta \in \mathbb{R}^d} \left\{ \sum_{s=1}^{t-1} (r_{s+1} - a_s^\top \theta)^2 + \lambda \|\theta\|^2 / 2 \right\} ,$$

where  $\lambda > 0$ . It can be computed in closed form:

$$\hat{\theta}_t = \mathbf{V}_t^{-1} \sum_{s=1}^{t-1} r_{s+1} a_s \quad \text{where} \quad \mathbf{V}_t = \sum_{s=1}^{t-1} a_s a_s^\top + \lambda \mathbf{I}_d . \quad (1.7)$$

[Abbasi-Yadkori et al. \(2011\)](#) proved the following confidence set for  $\theta_*$  based on this estimator by leveraging the theory of self-normalized processes.

**Theorem 1.2.2** (Theorem 2 of [Abbasi-Yadkori et al. \(2011\)](#)). *Let  $\delta \in (0, 1]$ . The set:*

$$\mathcal{E}_t(\delta) = \left\{ \theta, \|\hat{\theta}_t - \theta\|_{\mathbf{V}_t} \leq \sqrt{\lambda} S + \sigma \sqrt{d \log \left( \frac{1 + t/\lambda}{\delta} \right)} \right\} . \quad (1.8)$$

*is an anytime confidence set for  $\theta_*$  at level at least  $1 - \delta$ . In other words:*

$$\mathbb{P}(\theta_* \in \mathcal{E}_t(\delta) \text{ for all } t \geq 1) \geq 1 - \delta .$$

The radius of this confidence set is  $\mathcal{O}(\sqrt{d \log(t/\delta)})$ . The main challenge in proving this result comes from the complex correlations between the rewards and the actions, as the randomness in the reward directly impacts which arms will be played in the future. We leave a detailed discussion behind the design and proof of this confidence set for the next chapter where we will derive similar results for generalized linear bandits.

---

**Algorithm 2** OFUL (Abbasi-Yadkori et al., 2011)
 

---

**input:** Arms set  $\mathcal{A}$ , regularization coefficient  $\lambda$ , failure level  $\delta$ , norm upper-bound  $S$ .

Set  $\mathbf{V}_1 \leftarrow \lambda \mathbf{I}_d$  and  $\hat{\theta}_1 \leftarrow 0_d$ .  $\triangleright$  initialization

**for**  $t \in [1, T]$  **do**

Play the arm  $a_t$  according to Eq. (1.9).  $\triangleright$  planning

Observe reward  $r_{t+1}$ .

Update  $\hat{\theta}_{t+1}$  (Eq. (1.7)) and the confidence set  $\mathcal{E}_{t+1}(\delta)$  (Eq. (1.8)).  $\triangleright$  learning

**end for**

---

**An optimal algorithm.** Abbasi-Yadkori et al. (2011) leverage this confidence set to build OFUL, an optimistic algorithm for the LB problem. The rationale remains the same as for MAB, that is play the most rewarding action for the most optimistic environment:

$$\text{play } a_t \in \arg \max_{a \in \mathcal{A}} \max_{\theta \in \mathcal{E}_t(\delta)} a^\top \theta . \quad (1.9)$$

**Remark 1.2.4** (Parameter-search vs. exploration bonus). *The strategy followed by OFUL might appear dissonant with the exploration bonus approach that we discussed to motivate the need for a confidence set for  $\theta_*$ . Indeed, OFUL enforces optimism via parameter-search over the confidence set  $\mathcal{E}_t(\delta)$ . In the linear case, the two approaches are actually strictly equivalent, as one can show that the action selection process of OFUL can be rewritten as:*

$$\text{play } a_t \in \arg \max_{a \in \mathcal{A}} a^\top \hat{\theta}_t + \varepsilon_t(a) .$$

where  $\varepsilon_t(a) = \beta_t(\delta) \|a\|_{\mathbf{V}_t^{-1}}$ . The function  $\beta_t(\delta)$  corresponds to the radius of the confidence set from Theorem 1.2.2. This equivalence is classical, and is reminiscent of the fact that a confidence set for  $\theta_*$  directly induces high-probability bounds on the prediction error  $\Delta_t$  of any estimator in that set. This distinction between parameter-search and exploration bonus will be important in the next chapter; we shall see that they are no longer equivalent for non-linear reward models and induce important algorithmic distinctions.

The OFUL algorithm (pseudo-code is provided in Algorithm 2) matches the lower-bound of Theorem 1.2.1 up to logarithmic factors and is therefore (almost) minimax-optimal.

**Theorem 1.2.3** (Theorem 3 of Abbasi-Yadkori et al. (2011)). *With probability at least  $1 - \delta$  the regret of OFUL satisfies:*

$$\text{Regret}_{\theta_*}(T) = \mathcal{O} \left( d\sqrt{T} \log(T/\delta) \right) .$$

The cardinality of the action set  $|\mathcal{A}|$  no longer appears in the regret bound. By learning directly a *shared* parameter, the LB approach allows to handle a large (potentially infinite) number of arms, only at the price of the dimension  $d$  of the problem. Furthermore, for well-structured infinite arm-sets (e.g the unit ball) the planning program of Eq. (1.9) can be efficiently solved.

This concludes the introduction of the main ingredients to analyze the LB setting, which we will use as reference in this part of the dissertation. There exist many refinements and extensions to the vanilla LB problem (e.g sparse LB); we won't cover them here for the sake of conciseness but refer the interested reader to the excellent book of (Lattimore and Szepesvári, 2020, Part V) for an in-depth discussion of the LB problem and existing algorithms.

### 1.2.2 Limits of the linear model

The previous section attests that the LB is a compact yet powerful parametric bandit framework as it allows to neatly isolate the main challenges that come with parametrization. In particular, the design of appropriate confidence sets have emerged as essential in the design of optimal algorithms. From a learning-theoretic stand-point, an important message carried by the LB is that provided a learnable reward structure the parametric bandit is not fundamentally *harder* than MAB (similar regret rates are achieved). This is an encouraging conclusion as parametric bandits are more general and allow to address more challenging sequential problems.

Nonetheless, and because of its relative simplicity, the LB problem still leaves open some important theoretical questions. For instance, can we expect the same kind of guarantees with *richer* reward models? In particular, how does non-linearity impact the results obtained under linear parametrization? Precisely answering this question stands as an important step towards a better understanding of complex, non-linear environments which are ubiquitous in real-world situations. This last remark leads us to another important downside of the LB setting, this time from a practical point of view. Indeed, the reward model established in Eq. (1.5) prescribes *continuous* rewards. It therefore excludes cases of great practical importance such as binary or categorical rewards. For instance, binary rewards are omnipresent in virtually any web-related applications of bandits where rewards are typically measured by clicks or sales.

As we shall see in the following section such limitations can be addressed altogether by studying the so-called Generalized Linear Bandits (GLBs). This framework stands as a *minimalistic* non-linear extension to the LB setting; it is close enough to the LB setting that we can use the same fundamental tools (at least conceptually) and observe how their results are affected by non-linearity. The GLB setting therefore allows to neatly single out the potential issues raised by non-linearity and its effects on the exploration-exploitation trade-off. Finally, the GLB family encloses a diverse range of reward distributions - *e.g* binary or categorical. Designing efficient GLB algorithms is hence important from a practical standpoint, in order to handle the many real-world situations where the LB is critically misspecified.

## 1.3 Beyond linearity: Generalized Linear Bandits

The GLB framework was originally introduced and studied by [Filippi et al. \(2010\)](#); it generalizes the LB framework by adding a single non-linear activation  $\mu$  on top of the scalar product:

$$\forall a \in \mathcal{A}, \quad \mathbb{E}[r|a] = \mu(a^\top \theta_\star).$$

The goal of this section is to provide a thorough definition of the GLB learning problem. We will review related work in details and highlight some important limitations of existing approaches which motivated our contributions on this topic.

**Exponential distributions and generalized linear models.** To properly introduce the learning problem we first discuss the idea behind generalized linear models (GLMs) ([McCullagh and Nelder, 1989](#)). Let  $r$  be some response variable to some feature vector  $x$ . Its conditional distribution is said to belong to the *canonical exponential* family if its density  $p(\cdot|x)$  (w.r.t to a given reference measure) is of the form:

$$p(r|x) \propto \exp(r\beta_x - b(\beta_x)),$$

up to normalization. In the above definition,  $\beta_x$  is a scalar parameter (indexed by the feature vector  $x$ ) and  $b$  a real-valued univariate function assumed to be twice continuously differentiable.

Standard computations yields that:

$$\mathbb{E}[r|x] = \dot{b}(\beta_x) \quad \text{and} \quad \text{Var}(r|x) = \ddot{b}(\beta_x). \quad (1.10)$$

The function  $\mu := \dot{b}$  is commonly referred to as the *inverse link* function; from the above identity, we get that  $b$  is strictly convex and equivalently that  $\mu$  it is a strictly increasing function. The canonical exponential distribution family contains, for instance, the Gaussian (with known variance) and Exponential distributions (with Lebesgue reference measure) as well as the Bernoulli and Poisson distributions (with the counting reference measure). In a GLM it is assumed that the  $\{\beta_x\}_x$  share a common linear structure:

$$\beta_x = x^\top \theta_\star \quad \text{where } \theta_\star \in \mathbb{R}^d.$$

### 1.3.1 Learning problem

**Reward model.** The GLB framework of [Filippi et al. \(2010\)](#) imposes rewards to be described by a GLM. Strictly speaking it does not require the rewards distribution to belong to a canonical exponential family but only that the first moment condition in [Eq. \(1.10\)](#) is checked. The reward model can therefore be described by the following statement; when at round  $t$  the agent plays an arm  $a_t$  it receives the reward  $r_{t+1}$  which satisfies:

$$\mathbb{E}[r_{t+1}|\mathcal{F}_t] = \mu(a_t^\top \theta_\star). \quad (1.11)$$

The function  $\mu$  is strictly increasing and continuously differentiable. As for the LB some boundedness assumptions (over the arm-set, the unknown parameter and the noise) are made to ensure the soundness of the learning problem.

**Assumption 1.3.1** (Bounded decision set). *For any  $a \in \mathcal{A}$  we have  $\|a\| \leq 1$ . Furthermore, the unknown parameter  $\theta_\star$  satisfies  $\|\theta_\star\| \leq S$  where  $S$  is known.*

**Assumption 1.3.2** (Bounded noise). *There exist  $\sigma > 0$  such that for all  $t \geq 1$ :*

$$\left| r_{t+1} - \mu(a_t^\top \theta_\star) \right| \leq \sigma \quad \text{a.s.}$$

[Assumption 1.3.1](#) leads us to define the set  $\Theta := \mathcal{B}_d(0_d, S)$  which will be referred to as the *admissible* parameter set.

**Regret.** The cumulative pseudo-regret incurred when following a policy  $\pi$  writes:

$$\text{Regret}_{\theta_\star}^\pi(T) = T\mu(a_\star(\theta_\star)^\top \theta_\star) - \sum_{t=1}^T \mu(a_t^\top \theta_\star),$$

where  $a_\star(\theta_\star) = \arg \max_{a \in \mathcal{A}} \mu(a^\top \theta_\star)$ . Note that the function  $\mu$  being strictly increasing this can be simplified to match the definition of the optimal arm for LB, *i.e.*  $a_\star(\theta_\star) = \arg \max_{a \in \mathcal{A}} a^\top \theta_\star$ .

**Illustration: the Logistic Bandit problem.** Over the many bandit problems covered by the GLB framework surely one of the most interesting (at least from a practical standpoint) is the Logistic Bandit (LogB). The LogB will be our *red thread* throughout this part of the dissertation and we will extensively use it to illustrate discussions and findings. Paragraphs dedicated to the LogB are indicated by the  symbol - see below.

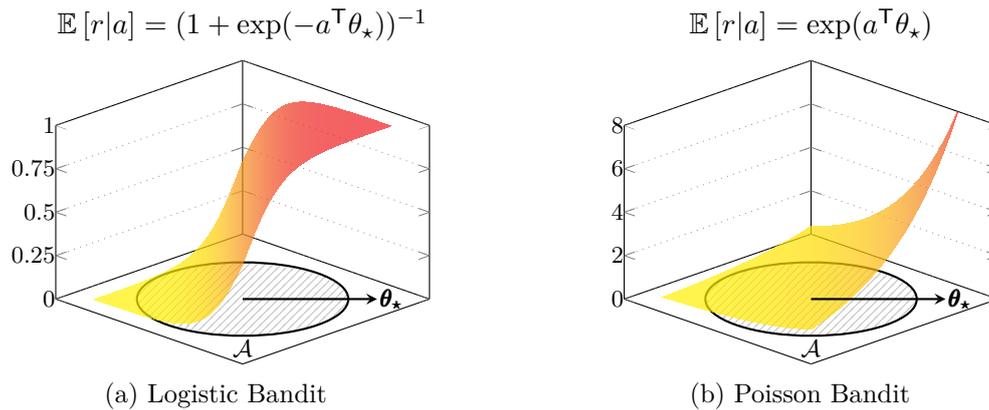


Figure 1.3: Graphical representation of two-dimensional GLBs; (*left*) the Logistic Bandit. The reward signal displays flat tails on the boundaries of the decision set. (*right*) the Poisson bandit. The reward signal has a flat left ( $x^\top \theta_\star < 0$ ) tail and an exploding right tail.

 The LogB arises when rewards are binary and sampled according to a Bernoulli distribution with linear log-odds. Formally, it allows to tackle the case where the reward obtained after playing an arm  $a$  is drawn according to a Bernoulli distribution with mean  $\mu(a^\top \theta_\star)$  where  $\mu$  is the *logistic function*:  $\mu(z) = (1 + \exp(-z))^{-1}$ . In other words:

$$r_{t+1} \sim \text{Bernoulli}(\mu(a_t^\top \theta_\star))$$

We provide a visual illustration of a two-dimensional LogB instance in Fig. 1.3a. The practical advantage of this model is to handle in a principled way binary feedback, which is ubiquitous in real-world problems (*e.g.* click/no-click in computational advertisement, recovery/no-recovery in clinical trials, and generally any kind of success/failure feedback). From a theoretic perspective the LogB captures the many challenges brought by non-linearity in GLBs, and as we shall see in the following chapters is the GLB for which our analysis brings the most distinctive learning-theoretic message compared to previous works. For this reason most of our results and findings that hold for self-concordant GLBs will be illustrated through the LogB case.

**Remark 1.3.1** (The Poisson Bandit: another important model). *The Poisson Bandit concerns the situation where the reward is drawn according to a Poisson distribution with mean  $\exp(a^\top \theta_\star)$ . In this case the inverse link function is  $\mu(z) = \exp(z)$ . The main difference with the LogB is that the Poisson Bandit exhibit both flat and exploding tails (cf Fig. 1.3b) which will lead to sensibly different learning-theoretic conclusions.*

### 1.3.2 Quantifying non-linearity.

As anticipated in previous sections one of the main motivation to study GLBs is to understand and address the potential issues raised by non-linearity. However, the nature and level of the non-linearity differs across reward models (*cf.* Fig. 1.3), and can even sensibly differs across several instances of the same model (*cf.* Fig. 1.4). The level of non-linearity is therefore highly *problem-dependent*; to compactly evaluate its impact on different bandit instances it is desirable to resort to non-linearity metrics which measures the “distance” of a given GLB problem to its LB counterpart (the linear model serves as a reference point when evaluating the effects of non-linearity).

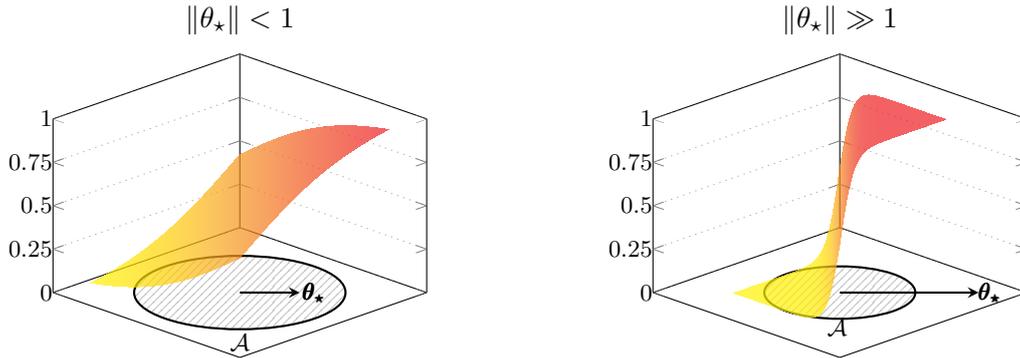


Figure 1.4: Illustration of the problem-dependent nature of non-linearity on two LogB problems. The level of non-linearity is tied to the regimes of the link function  $\mu$  covered by the set  $\{a^\top \theta_*, a \in \mathcal{A}\}$ . For both problems we have  $\mathcal{A} = \mathcal{B}_d(0_d, 1)$ ; if  $\|\theta_*\|$  is small (*left*) for all actions  $a^\top \theta_*$  lie close to the origin where the logistic function is almost linear. When  $\|\theta_*\|$  is large (*right*) many actions lay in flat regions of the logistic function; the resulting reward signal is highly non-linear.

**Reward sensitivity.** The GLB framework allows for such a measure through the analysis of the local sensitivity of the reward signal across the arm set. Formally, for a given link function  $\mu$  let us define the minimal and maximum effective sensitivity of the reward model as:

$$\ell_\mu(\mathcal{A}, \theta_*) := \min_{a \in \mathcal{A}} \dot{\mu}(a^\top \theta_*) \quad \text{and} \quad L_\mu(\mathcal{A}, \theta_*) := \max_{a \in \mathcal{A}} \dot{\mu}(a^\top \theta_*) .$$

The link function  $\mu$  being strictly increasing their ratio is well-defined;

$$\kappa_\mu(\mathcal{A}, \theta_*) := L_\mu(\mathcal{A}, \theta_*) / \ell_\mu(\mathcal{A}, \theta_*) \geq 1 .$$

This last quantity compactly captures the discrepancy between the GLB model attached to the tuple  $(\mu, \mathcal{A}, \theta_*)$  and its LB counterpart (note that if  $\mu(z) = z$  which is the linear case we have  $\kappa_\mu(\mathcal{A}, \theta_*) \equiv 1$ ). Indeed *the more non-linear the reward model, the greater the mismatch between  $L_\mu$  and  $\ell_\mu$  and the greater  $\kappa_\mu$ .*<sup>3</sup> For this reason, we informally identify it as a measure of non-linearity. We illustrate this in the following paragraph regarding the LogB case, for which such relationship between the model's level of non-linearity and  $\kappa_\mu$  is particularly remarkable.

(✂) In the LogB case, the maximum sensitivity  $L_\mu$  is bounded by the Lipschitz constant of the logistic function; in most cases of interests  $L_\mu = 1/4$  and hence  $\kappa_\mu = 4/\ell_\mu$ . The main quantity of interest is therefore the minimum sensitivity  $\ell_\mu$  which quantifies the *flatness* of the reward signal near the boundaries of the arm set. The tail of the logistic function being exponentially flat (as illustrated in Fig. 1.5) the ratio  $\kappa_\mu$  is often *very* large, even under reasonable configurations. Indeed, from the definition of the sigmoid function, one easily obtains that:

$$1/\ell_\mu \geq \exp\left(\max_{a \in \mathcal{A}} |a^\top \theta_*|\right).$$

The quantity  $a^\top \theta_*$  is directly tied to the mean reward obtained when playing  $a$ . This lower bound stresses that  $\kappa_\mu$  will be exponentially large as soon as there exists bad (resp. good) arms  $a$  associated with a low (resp. high) probability of receiving a positive reward.

<sup>3</sup>We drop the dependency w.r.t  $\theta_*$  and  $\mathcal{A}$  whenever obvious from context in order to reduce clutter.

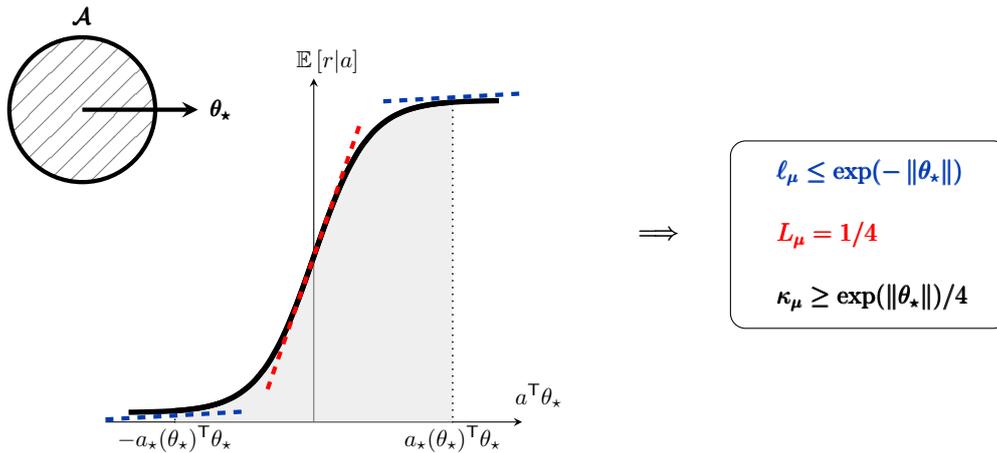


Figure 1.5: Illustration of  $\ell_\mu$ ,  $L_\mu$  and  $\kappa_\mu$  for a LogB with  $\mathcal{A} = \mathcal{B}_d(0, 1)$ . The larger  $\|\theta_\star\|$ , the flatter the tails, the higher the non-linearity level and the larger the effective sensitivity ratio  $\kappa_\mu$ . This growth is exponential as  $\kappa_\mu \approx \exp(\|\theta_\star\|)$ .

**Computable alternatives.** The minimum and maximum effective sensitivities  $\ell_\mu$  and  $L_\mu$  directly depend on the unknown parameter  $\theta_\star$  - and therefore so does their ratio  $\kappa_\mu$ . They therefore cannot in all generality be computed, but one can obtain proxys thanks to [Assumption 1.3.1](#). Indeed, it ensures that  $\theta_\star \in \Theta$  where  $\Theta = \mathcal{B}_d(0, S)$  is the set of admissible parameters. One can therefore compute the worst-case reward sensitivities over all potential reward signals:

$$\bar{\ell}_\mu(\mathcal{A}, \Theta) = \min_{a \in \mathcal{A}, \theta \in \Theta} \dot{\mu}(a^\top \theta) \quad \text{and} \quad \bar{L}_\mu(\mathcal{A}, \Theta) = \max_{a \in \mathcal{A}, \theta \in \Theta} \dot{\mu}(a^\top \theta), \quad (1.12)$$

as well as the associated ratio:

$$\bar{\kappa}_\mu(\mathcal{A}, \Theta) = \bar{\ell}_\mu(\mathcal{A}, \Theta) / \bar{L}_\mu(\mathcal{A}, \Theta). \quad (1.13)$$

We have the following trivial inequalities between the effective and worst-case sensitivities;<sup>4</sup>

$$\bar{\ell}_\mu \leq \ell_\mu, \quad \bar{L}_\mu \geq L_\mu \quad \text{and} \quad \bar{\kappa}_\mu \geq \kappa_\mu.$$

The tightness of those quantities (w.r.t to the *effective* non-linearity metrics) is a function of the true geometry (characterized by  $\mathcal{A}$  and  $\theta_\star$ ) as well as the tightness of the inequality  $\|\theta_\star\| \leq S$ .

 For the LogB one has that  $\bar{\kappa}_\mu \geq \exp(S)$ , which can be significantly larger than  $\kappa_\mu$ .

### 1.3.3 Linearization approach

We are now ready to start discussing existing work on the GLB problem. The main objective of this section is to provide a presentation of the main conceptual ideas and technical tools introduced by [Filippi et al. \(2010\)](#) in their seminal work on GLBs. The main idea behind their approach is to *linearize* the reward signal of GLBs in order to resort to LB recipes. Other recent works rely on similar mechanisms and carry the same learning-theoretic conclusions w.r.t to the effects of non-linearity. They are briefly discussed in the end of this section.

<sup>4</sup>We drop the dependency w.r.t  $\mathcal{A}$  and  $\Theta$  whenever clear from context to reduce clutter.

**Learning and confidence set.** [Filippi et al. \(2010\)](#) suggest to use the maximum likelihood principle in order to estimate  $\theta_*$ . Formally, given the regularized log-loss (recall that  $b$  is a primitive of the inverse link function  $\mu$ ):

$$\mathcal{L}_t(\theta) := \sum_{s=1}^{t-1} \left[ b(a_s^\top \theta) - r_{s+1} a_s^\top \theta \right] + \lambda \|\theta\|^2 / 2,$$

they compute the regularized maximum-likelihood estimator (MLE)  $\hat{\theta}_t := \arg \min_{\theta \in \mathbb{R}^d} \mathcal{L}_t(\theta)$ . This estimator is well-defined and unique given the strongly convex nature of the log-loss  $\mathcal{L}_t$  and for the ease of exposition we assume for now that the event  $\{\forall t \geq 1, \hat{\theta}_t \in \Theta\}$  holds. To design an optimistic strategy one needs to design a confidence set for  $\theta_*$ ; to this end, [Filippi et al. \(2010\)](#) leverage the properties of  $\hat{\theta}_t$  along with a linearization of the link function  $\mu$ . Upon differentiation of the log-loss, direct computation yields that:

$$\begin{aligned} \sum_{s=1}^{t-1} \left[ r_{s+1} - \mu(a_s^\top \theta_*) \right] a_s - \lambda \theta_* &= \sum_{s=1}^{t-1} \left[ \mu(a_s^\top \hat{\theta}_t) - \mu(a_s^\top \theta_*) \right] a_s + \lambda (\hat{\theta}_t - \theta_*), \\ &\geq \left( \sum_{s=1}^{t-1} \dot{\mu}(z_s) a_s a_s^\top + \lambda \mathbf{I}_d \right) (\hat{\theta}_t - \theta_*), \quad (z_s \in [a_s^\top \hat{\theta}_t, a_s^\top \theta_*]) \end{aligned}$$

thanks to an exact first-order Taylor expansion. To mirror the LB approach, one would like to introduce the design matrix  $\mathbf{V}_t = \sum_{s=1}^{t-1} a_s a_s^\top + \lambda' \mathbf{I}_d$  in the above inequality. This is possible by using a lower-bound of  $\dot{\mu}$  over the set  $\{a^\top \theta, a \in \mathcal{A}, \theta \in \Theta\}$ ; a.k.a, the minimum sensitivity  $\bar{\ell}_\mu$ . Up to the correct choice of the regularization parameter  $\lambda'$  this provides the following ordering:

$$\sum_{s=1}^{t-1} \dot{\mu}(z_s) a_s a_s^\top + \lambda \mathbf{I}_d \succeq \bar{\ell}_\mu \mathbf{V}_t.$$

Straight-forward algebraic manipulations yield that:

$$\left\| \hat{\theta}_t - \theta_* \right\|_{\mathbf{V}_t} \leq (1/\bar{\ell}_\mu) \left\| \sum_{s=1}^{t-1} \left[ r_{s+1} - \mu(a_s^\top \theta_*) \right] a_s \right\|_{\mathbf{V}_t^{-1}} + \frac{\lambda S}{\sqrt{\lambda' \bar{\ell}_\mu}}.$$

Noticing that  $r_{s+1} - \mu(a_s^\top \theta_*)$  is a zero mean noise (conditioned on  $\mathcal{F}_s$ ) and applying the tail inequality of ([Abbasi-Yadkori et al., 2011](#), Theorem 1) yields a confidence set that highly resembles the LB's (see [Theorem 1.3.1](#) below). Note this approach *critically* relies on lower-bounding  $\mu$  by a linear function:

$$|\mu(a_s^\top \hat{\theta}_t) - \mu(a_s^\top \theta_*)| \geq \bar{\ell}_\mu |a_s^\top (\hat{\theta}_t - \theta_*)|.$$

By definition of  $\bar{\ell}_\mu$  this requires  $\{\hat{\theta}_t \in \Theta\}$  to hold which can not always be ensured; [Filippi et al. \(2010\)](#) replace  $\hat{\theta}_t$  by its ‘‘projection’’  $\tilde{\theta}_t$  on  $\Theta$ :

$$\tilde{\theta}_t = \arg \min_{\theta \in \Theta} \left\| \sum_{s=1}^{t-1} \left[ \mu(a_s^\top \theta) - \mu(a_s^\top \hat{\theta}_t) \right] a_s + \lambda (\theta - \hat{\theta}_t) \right\|_{\mathbf{V}_t^{-1}}. \quad (1.14)$$

**Theorem 1.3.1.** [*GLB confidence set, [Filippi et al. \(2010\)](#)*] Let  $\delta \in (0, 1]$ . For all  $t \geq 1$  let  $\mathbf{V}_t = \sum_{s=1}^{t-1} a_s a_s^\top + (\lambda/\bar{\ell}_\mu) \mathbf{I}_d$ . The set:

$$\mathcal{E}_t(\delta) := \left\{ \theta \in \Theta, \left\| \theta - \tilde{\theta}_t \right\|_{\mathbf{V}_t} \leq \frac{2}{\bar{\ell}_\mu} \left( \sqrt{\lambda \bar{\ell}_\mu} S + \sigma \sqrt{d \log \left( \frac{1 + t \bar{\ell}_\mu / \lambda}{\delta} \right)} \right) \right\},$$

is an anytime confidence set for  $\theta_*$  at level at least  $1 - \delta$  i.e  $\mathbb{P}(\forall t \geq 1, \theta_* \in \mathcal{E}_t(\delta)) \geq 1 - \delta$ .

**Algorithm 3** GLM-UCB (Filippi et al., 2010)

---

**input:** Arm set  $\mathcal{A}$ , regularization coefficient  $\lambda$ , failure level  $\delta$ , admissible parameter set  $\Theta$ .  
 Compute the reward sensitivity constants  $\bar{\ell}_\mu$ ,  $\bar{L}_\mu$  and  $\bar{\kappa}_\mu$ .  $\triangleright$  initialization  
 Set  $\mathbf{V}_1 \leftarrow (\lambda/\bar{\ell}_\mu)\mathbf{I}_d$ ,  $\hat{\theta}_1 \leftarrow 0_d$  and  $\tilde{\theta}_1 \leftarrow 0_d$ .  
**for**  $t \in [1, T]$  **do**  
   Compute the exploration bonuses  $\{\varepsilon_t(a)\}_{a \in \mathcal{A}}$  according to Eq. (1.15).  
   Play the arm  $a_t$  according to Eq. (1.16).  $\triangleright$  planning  
   Observe reward  $r_{t+1}$ .  
   Update the estimator  $\hat{\theta}_{t+1}$  and the design matrix  $\mathbf{V}_{t+1}$ .  $\triangleright$  learning  
   Compute the  $\tilde{\theta}_{t+1}$  the projection of the MLE on  $\Theta$  (cf. Theorem 1.3.1).  
**end for**

---

The radius of this confidence set is  $\mathcal{O}(\bar{\ell}_\mu^{-1}\sqrt{d\log(t/\delta)})$ ; it is essentially the same as in the LB case but *inflated* by a factor  $1/\bar{\ell}_\mu$  (which is typically *very* large). Theorem 1.3.1 does not directly appear as it in (Filippi et al., 2010); it actually brings a slight improvement to their results. It can however be trivially extracted from their different proofs combined with (Abbasi-Yadkori et al., 2011, Theorem 1). We provide the proof in Appendix 1.A for the sake of completeness.

**Planning.** In their seminal paper Filippi et al. (2010) advocate for enforcing optimism through exploration bonuses. As we discussed in the LB case and as we shall shortly see for GLBs, a sound design for exploration bonuses is obtained by upper-bounding the prediction error  $\Delta_t(a) := |\mu(a^\top \theta_\star) - \mu(a^\top \tilde{\theta}_t)|$  (the parameter  $\tilde{\theta}_t$  is the projected version of  $\hat{\theta}_t$  on  $\Theta$  defined in Eq. (1.14)). To this end Filippi et al. (2010) again resort to linearization, this time “by above”:

$$\begin{aligned} \Delta_t(a) &= |\mu(a^\top \theta_\star) - \mu(a^\top \tilde{\theta}_t)| \\ &\leq \bar{L}_\mu |a^\top (\theta_\star - \tilde{\theta}_t)| \\ &\leq \bar{L}_\mu \|a\|_{\mathbf{V}_t^{-1}} \|\theta_\star - \tilde{\theta}_t\|_{\mathbf{V}_t} . \end{aligned}$$

where the first inequality consists in using that fact that  $\dot{\mu}$  is bounded by  $L_\mu$  on the interval  $[a^\top \theta_\star, a^\top \tilde{\theta}_t]$ , and the second is obtained thanks to Cauchy-Schwarz inequality. Leveraging the confidence set from Theorem 1.3.1 directly gives a confident upper-bound on the prediction error, and therefore yields a sound bonus function. Formally, for  $\delta \in (0, 1]$  define the following exploration bonus function:

$$\varepsilon_t(a) = 2\bar{\kappa}_\mu \|a\|_{\mathbf{V}_t^{-1}} \left( \sqrt{\lambda \bar{\ell}_\mu S} + \sigma \sqrt{d \log \left( (1 + t\bar{\ell}_\mu/\lambda)/\delta \right)} \right) \quad (1.15)$$

The GLM-UCB algorithm of Filippi et al. (2010) follows the strategy:

$$\text{play } a_t = \arg \max_{a \in \mathcal{A}} \mu(a^\top \tilde{\theta}_t) + \varepsilon_t(a) . \quad (1.16)$$

Pseudo-code for GLM-UCB is provided in Algorithm 3. Notice the resemblance of the exploration bonus function with its linear counterpart (see Remark 1.2.4). This GLB bonus essentially inflates the LB one by a factor  $\bar{\kappa}_\mu$ ; it is significantly larger for highly non-linear reward models. This is a direct consequence of the linearization approach followed by Filippi et al. (2010).

**Regret bound.** The proof for the regret upper-bound of GLM-UCB closely follows the LB analysis and yields a similar result. The following bound can easily be obtained by coupling the proof of (Filippi et al., 2010, Theorem 2) with Theorem 1.3.1. It shaves of a  $\sqrt{\log(T)}$  multiplicative term from the initial regret guarantee of GLM-UCB.

**Theorem 1.3.2** (Regret of GLM-UCB, Theorem 2 of [Filippi et al. \(2010\)](#)). *With probability at least  $1 - \delta$  the regret of GLM-UCB satisfies:*

$$\text{Regret}_{\theta_*}(T) = \mathcal{O}\left(\bar{\kappa}_\mu d \sqrt{T} \log(T/\delta)\right).$$

The regret of GLM-UCB therefore grows at the same rate (w.r.t  $T$  and  $d$ ) as the regret of OFUL on the LB problem. This is expected, given the similar structure between GLBs and LB and the fact that GLM-UCB employs the same recipes as the OFUL algorithm. Notice however the presence of the problem-dependent constant  $\bar{\kappa}_\mu$  in the regret upper-bound of [Theorem 1.3.2](#); behind this seemingly innocent dependency lies the main interest of the GLB study. Indeed, the level of non-linearity is immediately present in the regret bound and allows for clear-cut interpretations regarding the effects of non-linearity on the exploration-exploitation trade-off. We will discuss this learning-theoretic interpretation for GLM-UCB in the following section.

**Similar related work.** Similar regret guarantees were proven to be achievable by randomized algorithms ([Abeille and Lazaric, 2017](#); [Kveton et al., 2020](#)). In a parallel line of work, [Zhang et al. \(2016\)](#); [Jun et al. \(2017\)](#) focused on improving the efficiency of GLM-UCB (which requires expensive batch computations to compute the maximum-likelihood estimator  $\hat{\theta}_t$ ) and managed to design fully online algorithms while retaining the same regret guarantees. [Li et al. \(2017\)](#) proposed a modified version of GLM-UCB that enjoys a smaller dependency w.r.t to the dimension  $d$  when the number of available arms at each round is finite and small. All of these approaches rely at some point on the linearization approach of [Filippi et al. \(2010\)](#); as a result, their regret upper-bounds all display the same multiplicative dependency w.r.t  $\bar{\kappa}_\mu$ .

### 1.3.4 Limitations and challenges

**Some important limitations.** The effect of non-linearity is directly perceptible in GLM-UCB’s regret bound through  $\bar{\kappa}_\mu$ , which quantifies the reward signal’s level of non-linearity. In that sense, the regret bounds depend in an “unpleasant manner on the form of the link function of the GLM, and it seems there may be significant room for improvement” ([Lattimore and Szepesvári, 2020](#), Section 19.4.5). Indeed, from a theoretical perspective the regret of GLM-UCB being *proportional* to  $\bar{\kappa}_\mu$  suggests that non-linearity makes the GLB problem *harder*: the more non-linear the reward signal, the larger  $\bar{\kappa}_\mu$  and the larger the regret. This indicates that either the current analyses fail to handle the regime where the reward function is significantly non-linear (which was the primary purpose of extending LB to GLBs) or that the problem is fundamentally hard (which could be confirmed by an adequate lower-bound). From a practical perspective, the typical scaling of  $\bar{\kappa}_\mu$  (e.g exponential for Logistic and Poisson bandits) dramatically narrows down the class of problems that existing algorithms can efficiently address. We emphasize that this conclusion is not an artifact of a potentially loose regret analysis; existing GLB algorithms are over-exploratory *by design* as their exploration bonus is explicitly inflated by  $\bar{\kappa}_\mu$ .

 On the LogB problem the regret of GLM-UCB satisfies  $\text{Regret}_{\theta_*}(T) = \tilde{\mathcal{O}}\left(e^S d \sqrt{T}\right)$ .

**Yet some hope exists.** By resorting to an asymptotic argument [Filippi et al. \(2010\)](#) conjectured that their confidence set ([Theorem 1.3.1](#)) could be deflated by a factor  $\sqrt{\bar{\kappa}_\mu}$ , which would naturally shave off the regret from the same amount. Furthermore, a strong signal suggesting that GLM-UCB is indeed sub-optimal came from the Bayesian analysis of the LogB problem by [Dong et al. \(2019\)](#). They showed that under some configurations of  $\mathcal{A}$  the Bayesian regret of the Thompson Sampling algorithm is *independent* of  $\bar{\kappa}_\mu$ . We are interested in frequentist regret

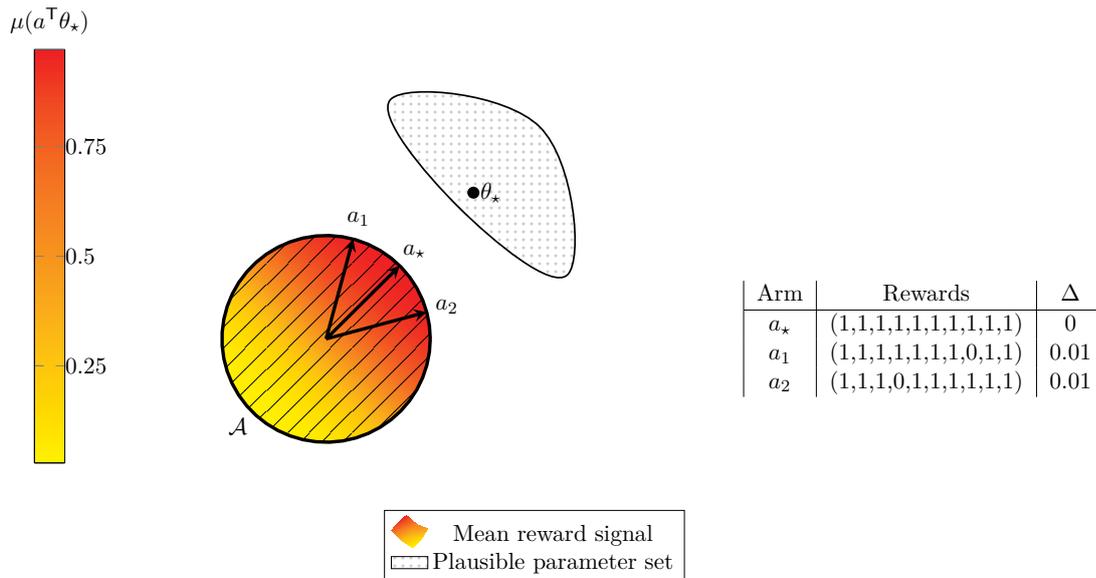


Figure 1.6: ~~(S)~~ An illustration of the information/regret tension on a LogB problem. The three arms  $a_*$ ,  $a_1$  and  $a_2$  have each been played 10 times. All lie in the flat right tail of the reward signal; they have extremely similar expected rewards and small conditional variance (cf. the second order condition of Eq. (1.11)). When pulled they almost always yield the same reward (here 1). It is particularly hard for a learning algorithm to identify with high confidence the best arm. Similarly, closely estimating  $\theta_*$  in this direction is hard since the arms hold little discriminative information about the parameters that could have generated the observed rewards. That being said, playing the arms  $a_2$  and  $a_3$  is not that harmful in terms of regret since their small sub-optimality gaps  $\Delta$  are small. This tension between information and regret is proper to GLBs and absent in the LB setting.

guarantees, which are known to be stronger than Bayesian ones - Bayesian regret upper-bounds do not imply (in general) frequentist regret upper-bounds. The result of Dong et al. (2019) nonetheless suggests that there is room for improvement in the frequentist analysis of GLBs.

**Global vs. local control of the link function.** We emphasized earlier that a salient distinction of GLBs is the shifting reward sensitivity, which varies across the arm set (e.g flat vs. exploding behaviors). Each regime of reward sensitivity is tied to its own intrinsic information/regret balance. For instance, flat areas of the reward signal are characterized by their small informative nature. Indeed the arms laying in such flat zones have extremely similar rewards; this make the estimation of  $\theta_*$  in the directions supporting these areas *hard*. This however might not be necessarily dramatic for the regret as the gap between arms in flat zones is small. We illustrate this high-level idea for the LogB in Fig. 1.6. The reverse reasoning holds for region of high sensitivity; estimation is then *easy* but small estimation errors can still contribute largely to the regret. Unfortunately a global linearization approach like the one followed to design GLM-UCB takes the worst of both cases and ignores the fact that information and regret often balance each other. Technically speaking, it uses both uniform upper ( $\bar{L}_\mu$ ) and lower ( $\bar{\ell}_\mu$ ) bounds for the derivative of the link function. Because they are not attained at the same point, at least one of them is loose. Alleviating the dependency in  $\bar{\kappa}_\mu$  thus calls for an analysis and for algorithms that better handle the non-linearity of the reward signal, switching from a global to a *local* treatment. As previously mentioned, a thorough control on the *prediction error*  $\Delta_t$  is key to the tight design of an optimistic algorithm. The challenge therefore resides in finely handling

the locality when controlling the prediction error, which also requires improved confidence sets.

## 1.4 Our approach: a self-concordant analysis for GLBs.

We bring forward a refined non-linear treatment for a wide class of GLBs known as *self-concordant*. Before discussing our main results we start by defining the self-concordance property for GLBs, which allows for a *local* treatment of the reward signal.

### 1.4.1 Setting: self-concordant GLBs

We closely follow the GLB setting laid out in [Section 1.3.1](#) but we make an additional assumption compared to previous work. We consider a slightly restricted class of GLBs known as *self-concordant*. Their inverse link function satisfies the following smoothness property.

**Assumption 1.4.1** (Generalized self-concordance). *The inverse link function is twice continuously differentiable and satisfies  $|\ddot{\mu}| \leq \dot{\mu}$ .*

This assumption is mild and the class of self-concordant GLBs is wide; it contains for instance the important Logistic and Poisson bandits. Other important instances of this class are the Binomial and Multinomial bandits. Actually, under [Assumption 1.3.1](#) all GLBs are self-concordant however with a different self-concordance constant:  $|\ddot{\mu}| \leq a\dot{\mu}$ . The constant  $a$  impacts our regret bounds only linearly so we restrict  $a = 1$  to reduce clutter.

**Remark 1.4.1** (On the self-concordant property). *An alternative definition for a self-concordant GLB requires that its log-loss is a generalized self-concordant function in the sense of [Bach \(2010\)](#). This is equivalent to the derivative control presented in [Assumption 1.4.1](#).*

In addition to the self-concordance property we will also restrict our study to GLBs that stray a little closer to exponential families. More precisely we ask for both the first and second moment conditions of [Eq. \(1.10\)](#) to be checked - previous work only required the first moment condition to be satisfied. Formally, the reward  $r_{t+1}$  obtained after playing  $a_t$  is such that:

$$\mathbb{E}[r_{t+1}|\mathcal{F}_t] = \mu(a_t^\top \theta_\star), \quad (1.17)$$

$$\text{and } \text{Var}[r_{t+1}|\mathcal{F}_t] = \dot{\mu}(a_t^\top \theta_\star). \quad (1.18)$$

This additional property is of great importance to the design of our confidence sets as it ties the variance of the rewards to the non-linearity of the link function. It is naturally checked by GLBs that are directly derived from an exponential family, which is the case for most GLBs of interest and of all the GLBs that we have listed so far (Logistic, Poisson, Multinomial, ..).

### 1.4.2 Brief summary of contributions

Here we present a very brief summary of our contributions on the GLB problem. The ambition is not to provide precise results but rather to give the high-level flavor of the different chapters. In [Chapter 2](#) we present our main technical contribution: a new Bernstein-like tail-inequality for self-normalized martingales which yields improved confidence sets for any GLB satisfying [Eq. \(1.18\)](#). A salient feature of those confidence sets is their sensitivity to the effective level of non-linearity encoded by  $\mu$ ,  $\theta_\star$  and  $\mathcal{A}$ . They improve over the LB-inspired set presented above and as a by-product, give the first formal proof of the conjecture of [Filippi et al. \(2010\)](#) in the finite-time, adaptive-design case. In [Chapter 3](#) we apply the improved confidence sets to the design of new optimistic algorithms for self-concordant GLBs. The associated regret bounds show refined problem-dependent scalings which tells a much more nuanced story about

Approach	Regret Upper-Bound	Regret Lower-Bound
Linearization analysis (Filippi et al., 2010), (Abeille and Lazaric, 2017), (Li et al., 2017), ..	$\tilde{\mathcal{O}}\left(e^S \sqrt{T}\right)$	$\Omega\left(e^{-\ \theta_\star\ /2} \sqrt{T}\right)$ (Chapter 3)
Self-concordant analysis (Chapter 3)	$\tilde{\mathcal{O}}\left(e^{-\ \theta_\star\ /2} \sqrt{T}\right)$	

Table 1.1:  Illustration of our results on the Logistic Bandit with action set  $\mathcal{A} = \mathcal{B}_d(0, 1)$ . Recall that  $\|\theta_\star\| \leq S$ . Our analysis yields tight regret bounds with improved problem-dependent rates and exponential acceleration over previous approach. Furthermore, it shows that for the LogB the larger  $\|\theta_\star\|$  (i.e the more non-linear the reward signal) the smaller the regret.

the effects of non-linearity. The most striking difference occurs for the LogB where our analysis shows that some highly non-linear problems are in fact *easier* to solve than their linear counterparts. Learning-theoretic considerations apart, our algorithms enjoy regret bounds that display critically reduced dependency w.r.t the problem-dependent constant  $\bar{\kappa}_\mu$ . For the Logistic and Poisson bandit, this lead to an exponential acceleration over GLM-UCB and related algorithms - we illustrate this in Table 1.1 for the LogB. Our algorithms also come with additional desirable qualities such as tractability (they do not require non-convex optimization routine, unlike previous algorithms). Finally, for the Logistic Bandit we show that our bounds are *tight* (w.r.t the problem-dependent quantities of interest) thanks to the first problem-dependent lower-bound in a GLB setting. We extend our findings in Chapter 4 to non-stationary settings, which comes with additional challenges for an appropriate treatment of non-linearity.

### 1.4.3 Notations and first technical results

We use this section to introduce additional notation that will be used throughout the manuscript and to provide a few technical results that are inherited from the self-concordance property.

**Learning.** As in previous work we will use the MLE  $\hat{\theta}_t$ ; a minor difference is that we will sometimes resort to time-varying regularization - this will be indicated by the time indexing of the regularization parameter ( $\lambda_t$  instead of  $\lambda$ ). Recall the definition of the MLE; for  $t \geq 1$ :

$$\hat{\theta}_t := \arg \min_{\theta \in \mathbb{R}^d} \left\{ \mathcal{L}_t(\theta) := \sum_{s=1}^{t-1} \left[ b(a_s^\top \theta) - r_{s+1} a_s^\top \theta \right] + \lambda \|\theta\|^2 / 2 \right\} .$$

Recall that  $b$  is a primitive of the inverse link function  $\mu$ . It is also convenient to define:

$$g_t(\theta) := \sum_{s=1}^{t-1} \mu(a_s^\top \theta) a_s + \lambda \theta ,$$

which is such that  $\nabla \mathcal{L}_t|_\theta = g_t(\theta) - \sum_{s=1}^{t-1} r_{s+1} a_s$ . Since by definition of the MLE we have  $\nabla \mathcal{L}_t|_{\hat{\theta}_t} = 0$  this yields the concise characterization:

$$g_t(\hat{\theta}_t) = \sum_{s=1}^{t-1} r_{s+1} a_s .$$

Conceptually the map  $g_t$  allows to characterize the *on-trajectory* directional prediction discrepancy between two parameters since for any  $\theta_1, \theta_2 \in \mathbb{R}^d$  (ignoring regularization terms):

$$g_t(\theta_1) - g_t(\theta_2) \approx \sum_{s=1}^{t-1} \left[ \mu(a_s^\top \theta_1) - \mu(a_s^\top \theta_2) \right] a_s .$$

Another important quantity is the Hessian of the log-loss:

$$\mathbf{H}_t(\theta) := \sum_{s=1}^{t-1} \dot{\mu}(a_s^\top \theta) a_s a_s^\top + \lambda \mathbf{I}_d \succ 0 .$$

**Taylor expansions.** A central idea when analyzing GLBs is to tightly link *estimation* errors (e.g between  $\hat{\theta}_t$  and  $\theta_*$ ) to *prediction* errors (e.g between  $\mu(a^\top \hat{\theta}_t)$  and  $\mu(a^\top \theta_*)$ ). Exact Taylor expansion is a powerful tool to achieve this; we will use it abundantly and in the following lines we introduce useful notations to this end. Specifically, for any  $a \in \mathcal{A}$  and  $\theta_1, \theta_2 \in \mathbb{R}^d$  define:

$$\alpha(a, \theta_1, \theta_2) := \int_{v=0}^1 \dot{\mu} \left( a^\top \theta_1 + v a^\top (\theta_2 - \theta_1) \right) dv , \quad (1.19)$$

$$\tilde{\alpha}(a, \theta_1, \theta_2) := \int_{v=0}^1 (1-v) \dot{\mu} \left( a^\top \theta_1 + v a^\top (\theta_2 - \theta_1) \right) dv , \quad (1.20)$$

so that we have the following identities (recall that  $\dot{b} = \mu$ ):

$$\begin{aligned} \mu(a^\top \theta_2) - \mu(a^\top \theta_1) &= \alpha(a, \theta_1, \theta_2) a^\top (\theta_2 - \theta_1) , \\ b(a^\top \theta_2) - b(a^\top \theta_1) &= \mu(a^\top \theta_1) a^\top (\theta_2 - \theta_1) + \tilde{\alpha}(a, \theta_1, \theta_2) (a^\top (\theta_2 - \theta_1))^2 . \end{aligned}$$

This allows use to link estimation errors to measurable on-trajectory metrics such as the map  $g_t$  or the log-loss itself. Indeed denoting:

$$\begin{aligned} \mathbf{G}_t(\theta_1, \theta_2) &:= \sum_{s=1}^{t-1} \alpha(a_s, \theta_1, \theta_2) a_s a_s^\top + \lambda \mathbf{I}_d \succ 0 , \\ \tilde{\mathbf{G}}_t(\theta_1, \theta_2) &:= \sum_{s=1}^{t-1} \tilde{\alpha}(a_s, \theta_1, \theta_2) a_s a_s^\top + \lambda \mathbf{I}_d \succ 0 , \end{aligned}$$

simple Taylor expansions allow to translate deviation in  $g_t$  or  $\mathcal{L}_t$  to deviation in parameters, under suitable metrics:

$$g_t(\theta_1) - g_t(\theta_2) = \mathbf{G}_t(\theta_1, \theta_2) (\theta_1 - \theta_2) , \quad (1.21)$$

$$\mathcal{L}_t(\theta_1) - \mathcal{L}_t(\theta_2) = \nabla \mathcal{L}_t|_{\theta_2}^\top (\theta_1 - \theta_2) + (\theta_1 - \theta_2)^\top \tilde{\mathbf{G}}_t(\theta_1, \theta_2) (\theta_1 - \theta_2) . \quad (1.22)$$

**Important inequalities and self-concordant control.** The matrix  $\mathbf{G}_t$  (resp.  $\tilde{\mathbf{G}}_t$ ) therefore allows to seamlessly switch from parameter deviations to measurable on-trajectory errors through  $g_t$  (resp.  $\mathcal{L}_t$ ) and vice-versa. In particular,  $\mathbf{G}_t(\theta, \theta_*)$  is the metric produced when measuring “distance” between  $\theta$  and  $\theta_*$  through the cumulative prediction error  $g_t(\theta) - g_t(\theta_*)$ . It induces a particular geometry of the parameter space; precisely characterizing it is central to switch from deviations in parameter to prediction errors. This geometry is however unknown (it depends on  $\theta_*$ ) and particularly cumbersome to describe, so we must find a way to approximate it. Conceptually the linearization approach of [Filippi et al. \(2010\)](#) “flattens” the manifold generated by  $\mathbf{G}_t(\theta_*, \theta)$  as it links the latter to the design matrix  $\mathbf{V}_t = \sum_{s=1}^{t-1} a_s a_s^\top + (\lambda/\bar{\ell}_\mu) \mathbf{I}_d$  - which no

longer depends on  $\theta$  nor  $\theta_*$ . More precisely, the linearization approach resorts to the following set of inequalities. For all  $a \in \mathcal{A}$  and  $\theta_1, \theta_2 \in \Theta$ :

$$\alpha(a, \theta_1, \theta_2) \geq \bar{\ell}_\mu ,$$

which translates into the following matrix inequality:

$$\mathbf{G}_t(\theta_1, \theta_2) \succeq \bar{\ell}_\mu \mathbf{V}_t \quad \text{for all } \theta_1, \theta_2 \in \Theta . \quad (1.23)$$

The self-concordance property allows for a finer approximation of the prediction error geometry, aware of the effective sensitivity of the reward signal. We provide here some of the central inequalities that will be used throughout the manuscript. They are inspired from the study of Newton's method for logistic regression by [Bach \(2010\)](#). The technical details of the proof are deferred to [Appendix 1.B](#).

**Proposition 1.4.1** (Self-concordance control). *Under [Assumption 1.4.1](#) we have the following list of inequalities. Let  $a \in \mathcal{A}$  and  $\theta_1, \theta_2 \in \mathbb{R}^d$ . Then:*

$$\alpha(a, \theta_1, \theta_2) \geq \frac{\dot{\mu}(a^\top \theta)}{1 + |a^\top(\theta_1 - \theta_2)|} \quad \text{for any } \theta \in \{\theta_1, \theta_2\} , \quad (1.24)$$

$$\geq \frac{\dot{\mu}(a^\top \theta)}{1 + 2S} \quad \text{when } \theta_1, \theta_2 \in \Theta . \quad (1.25)$$

Similar results holds for  $\tilde{\alpha}$ :

$$\tilde{\alpha}(a, \theta_1, \theta_2) \geq \frac{\dot{\mu}(a^\top \theta_1)}{2 + |a^\top(\theta_1 - \theta_2)|} , \quad (1.26)$$

$$\geq \frac{\dot{\mu}(a^\top \theta)}{2 + 2S} \quad \text{when } \theta_1, \theta_2 \in \Theta . \quad (1.27)$$

This results into the following matrix inequalities; for any  $\theta_1, \theta_2 \in \Theta$ :

$$\mathbf{G}_t(\theta_1, \theta_2) \succeq (1 + 2S)^{-1} \mathbf{H}_t(\theta) \quad \text{for any } \theta \in \{\theta_1, \theta_2\} , \quad (1.28)$$

$$\tilde{\mathbf{G}}_t(\theta_1, \theta_2) \succeq (2 + 2S)^{-1} \mathbf{H}_t(\theta_1) \quad (1.29)$$

At first sight it might not be clear already why the above inequalities are stronger than the uniform bound of [Eq. \(1.23\)](#). Without going into too much details it namely allows (for instance) to link the cumulative prediction error  $g_t(\hat{\theta}_t) - g_t(\theta_*)$  and the deviation in parameter space through the Hessian matrix  $\mathbf{H}_t(\theta_*)$ . The metric  $\|\cdot\|_{\mathbf{H}_t(\theta_*)}$  directly depends on to the effective reward sensitivity and is the right *concentration* metric in GLBs. Proving this last statement is the whole point of the following chapter.

# Appendix

## Appendix 1.A LB-inspired confidence set

In this section we provide a proof for [Theorem 1.3.1](#) since this result does not appear as it in [Filippi et al. \(2010\)](#). It is obtained by a straightforward application of ([Abbasi-Yadkori et al., 2011](#), Theorem 1) combined with the bounding strategy of ([Filippi et al., 2010](#)). We use notations from [Section 1.4.3](#).

**Theorem 1.3.1.** [*GLB confidence set, [Filippi et al. \(2010\)](#)*] Let  $\delta \in (0, 1]$ . For all  $t \geq 1$  let  $\mathbf{V}_t = \sum_{s=1}^{t-1} a_s a_s^\top + (\lambda/\bar{\ell}_\mu) \mathbf{I}_d$ . The set:

$$\mathcal{E}_t(\delta) := \left\{ \theta \in \Theta, \|\theta - \tilde{\theta}_t\|_{\mathbf{V}_t} \leq \frac{2}{\bar{\ell}_\mu} \left( \sqrt{\lambda \bar{\ell}_\mu} S + \sigma \sqrt{d \log \left( \frac{1 + t \bar{\ell}_\mu / \lambda}{\delta} \right)} \right) \right\},$$

is an anytime confidence set for  $\theta_\star$  at level at least  $1 - \delta$  i.e  $\mathbb{P}(\forall t \geq 1, \theta_\star \in \mathcal{E}_t(\delta)) \geq 1 - \delta$ .

*Proof.* By [Eq. \(1.21\)](#):

$$\|g_t(\tilde{\theta}_t) - g_t(\theta_\star)\|_{\mathbf{G}_t^{-1}} = \|\tilde{\theta}_t - \theta_\star\|_{\mathbf{G}_t}. \quad (1.30)$$

Recall that  $\theta_\star \in \Theta$  by [Assumption 1.3.1](#) and  $\tilde{\theta}_t \in \Theta$  by its definition in [Eq. \(1.14\)](#). Therefore [Eq. \(1.23\)](#) holds and we have the following chain of inequalities:

$$\begin{aligned} \|\tilde{\theta}_t - \theta_\star\|_{\mathbf{V}_t} &\leq \bar{\ell}_\mu^{-1/2} \|\tilde{\theta}_t - \theta_\star\|_{\mathbf{G}_t} && \text{(Eq. (1.23))} \\ &= \bar{\ell}_\mu^{-1/2} \|g_t(\tilde{\theta}_t) - g_t(\theta_\star)\|_{\mathbf{G}_t^{-1}} && \text{(Eq. (1.30))} \\ &\leq \bar{\ell}_\mu^{-1} \|g_t(\tilde{\theta}_t) - g_t(\theta_\star)\|_{\mathbf{V}_t^{-1}} && \text{(Eq. (1.23))} \\ &\leq \bar{\ell}_\mu^{-1} \left( \|g_t(\theta_\star) - g_t(\hat{\theta}_t)\|_{\mathbf{V}_t^{-1}} + \|g_t(\tilde{\theta}_t) - g_t(\hat{\theta}_t)\|_{\mathbf{V}_t^{-1}} \right) && \text{(triangle inequality)} \\ &\leq 2\bar{\ell}_\mu^{-1} \|g_t(\theta_\star) - g_t(\hat{\theta}_t)\|_{\mathbf{V}_t^{-1}}. && \text{(def. of } \tilde{\theta}_t \text{) (1.31)} \end{aligned}$$

where in the last inequality we used that by definition of  $\tilde{\theta}_t$  and given  $\theta_\star \in \Theta$  we have:

$$\|g_t(\theta_\star) - g_t(\hat{\theta}_t)\|_{\mathbf{V}_t^{-1}} \geq \|g_t(\tilde{\theta}_t) - g_t(\hat{\theta}_t)\|_{\mathbf{V}_t^{-1}}.$$

It remains to bound the l.h.s of [Eq. \(1.31\)](#). By the optimality condition of the MLE ( $\nabla \mathcal{L}_t|_{\hat{\theta}_t} = 0_d$ ) we have that  $g_t(\hat{\theta}_t) = \sum_{s=1}^{t-1} r_{s+1} a_s$ . Therefore by direct computation:

$$\begin{aligned} g_t(\hat{\theta}_t) - g_t(\theta_\star) &= \sum_{s=1}^{t-1} \left[ r_{s+1} - \mu(a_s^\top \theta_\star) \right] a_s - \lambda \theta_\star, \\ &= \sum_{s=1}^{t-1} \eta_{s+1} a_s - \lambda \theta_\star, \end{aligned}$$

where  $\eta_{s+1} := r_{s+1} - \mu(a_s^\top \theta_\star)$ . As a result:

$$\|g_t(\hat{\theta}_t) - g_t(\theta_\star)\|_{\mathbf{V}_t^{-1}} \leq \left\| \sum_{s=1}^{t-1} \eta_{s+1} a_s \right\|_{\mathbf{V}_t^{-1}} + \sqrt{\lambda \bar{\ell}_\mu} S$$

since  $\mathbf{V}_t \succeq (\lambda/\bar{\ell}_\mu)\mathbf{I}_d$  and  $\|\theta_\star\| \leq S$ . Finally, note that by Eq. (1.11) along with Assumption 1.3.2 we have that for all  $s \geq 1$ ,  $\eta_{s+1}$  is  $\sigma$ -sub-Gaussian conditionally on  $\mathcal{F}_s$ . Therefore a direct application of Theorem 1 of Abbasi-Yadkori et al. (2011) yields:

$$\left\| \sum_{s=1}^{t-1} \eta_{s+1} a_s \right\|_{\mathbf{V}_t^{-1}} \leq \sigma \sqrt{2 \log \left( \frac{\det(\mathbf{V}_t) \bar{\ell}_\mu^{d/2}}{\delta \lambda^{d/2}} \right)}. \quad (1.32)$$

Assembling Eq. (1.32) with Eq. (1.31) along with an application of the determinant-trace inequality (see Lemma B.2) yields the announced result.  $\blacksquare$

## Appendix 1.B Proof of self-concordance results

We start by stating and proving the following lemmas, obtained by following the line of proof from (Bach, 2010, Lemma 1).

**Lemma 1.B.1.** *Let  $f$  be a strictly increasing, twice differentiable function such that  $|\ddot{f}| \leq \dot{f}$ , and let  $\mathcal{Z}$  be any bounded interval of  $\mathbb{R}$ . Then, for all  $z_1, z_2 \in \mathcal{Z}$ :*

$$\int_{v=0}^1 \dot{f}(z_1 + v(z_2 - z_1)) dv \geq \frac{\dot{f}(z)}{1 + |z_1 - z_2|} \quad \text{for } z \in \{z_1, z_2\}.$$

*Proof.* The function  $f$  being strictly increasing, we have that  $\dot{f}(z) > 0$  for any  $z \in \mathcal{Z}$ . Therefore:

$$\begin{aligned} -1 &\leq \frac{\ddot{f}(z)}{\dot{f}(z)} \leq 1 \\ \Rightarrow & -|z_1 - z_0| \leq \int_{z_1 \wedge z_0}^{z_1 \vee z_0} \frac{\ddot{f}(z)}{\dot{f}(z)} dz \leq |z_1 - z_0| && \text{for any } z_0 \in \mathcal{Z} \\ \Leftrightarrow & -|z_1 - z_0| \leq \log \left( \dot{f}(z_1 \vee z_0) / \dot{f}(z_1 \wedge z_0) \right) \leq |z_1 - z_0| \\ \Leftrightarrow & \dot{f}(z_1 \wedge z_0) \exp(-|z_1 - z_0|) \leq \dot{f}(z_1 \vee z_0) \leq \dot{f}(z_1 \wedge z_0) \exp(|z_1 - z_0|). \end{aligned} \quad (1.33)$$

Assume for now that  $z_2 \geq z_1$ , let  $v \geq 0$  and set  $z_0 = z_1 + v(z_2 - z_1)$ , which is such that  $z_0 \geq z_1$ . Using this definition with the l.h.s inequality of Eq. (1.33) we easily get:

$$\begin{aligned} \dot{f}(z_1 + v(z_2 - z_1)) &\geq \dot{f}(z_1) \exp(-v|z_2 - z_1|) \\ \Rightarrow \int_{v=0}^1 \dot{f}(z_1 + v(z_2 - z_1)) dv &\geq \dot{f}(z_1) \frac{1 - \exp(-|z_1 - z_2|)}{|z_1 - z_2|} \\ &\geq \dot{f}(z_1) (1 + |z_1 - z_2|)^{-1}. \end{aligned}$$

where the last inequality is easily obtained by using  $\exp(x) \geq 1 + x$  for all  $x \in \mathbb{R}$ . The same inequality can be proven when  $z_2 \leq z_1$  by using the r.h.s inequality of Eq. (1.33) instead. We have therefore proven the announced result, but only for  $z = z_1$ . The proof is concluded by realizing that  $z_1$  and  $z_2$  play a symmetric role in the problem (for instance, perform the change of variable  $u \leftarrow (1 - v)$  in the integral that we wish to lower-bound).  $\blacksquare$

We now state a second result, which proof closely follows the one of Lemma 1.B.1.

**Lemma 1.B.2.** *Let  $f$  be a strictly increasing function such that  $|\ddot{f}| \leq \dot{f}$ , and let  $\mathcal{Z}$  be any bounded interval of  $\mathbb{R}$ . Then, for all  $z_1, z_2 \in \mathcal{Z}$ :*

$$\int_{v=0}^1 (1 - v) \dot{f}(z_1 + v(z_2 - z_1)) dv \geq \frac{\dot{f}(z_1)}{2 + |z_1 - z_2|}.$$

*Proof.* From Eq. (1.33) it can easily be extracted that for all  $v \geq 0$ :

$$\dot{f}(z_1 + v(z_2 - z_1)) \geq \dot{f}(z_1) \exp(-v|z_1 - z_2|).$$

Integrating between  $v \in [0, 1]$  and subsequently integrating by part, we obtain:

$$\begin{aligned} \int_{v=0}^1 (1-v) \dot{f}(z_1 + v(z_2 - z_1)) dv &\geq \dot{f}(z_1) \left( \frac{1}{|z_1 - z_2|} + \frac{\exp(-|z_1 - z_2|) - 1}{|z_1 - z_2|^2} \right) \\ &= \dot{f}(z_1) g(|z_1 - z_2|). \end{aligned}$$

where we defined:

$$g(x) := \frac{1}{x} \left( 1 + \frac{\exp(-x) - 1}{x} \right).$$

Finally, we use Lemma B.1 which guarantees that  $g(x) \geq (2+x)^{-1}$  for all  $x \geq 0$  to prove the claimed result.  $\blacksquare$

We are now ready to prove Proposition 1.4.1.

**Proposition 1.4.1** (Self-concordance control). *Under Assumption 1.4.1 we have the following list of inequalities. Let  $a \in \mathcal{A}$  and  $\theta_1, \theta_2 \in \mathbb{R}^d$ . Then:*

$$\alpha(a, \theta_1, \theta_2) \geq \frac{\dot{\mu}(a^\top \theta)}{1 + |a^\top (\theta_1 - \theta_2)|} \quad \text{for any } \theta \in \{\theta_1, \theta_2\}, \quad (1.24)$$

$$\geq \frac{\dot{\mu}(a^\top \theta)}{1 + 2S} \quad \text{when } \theta_1, \theta_2 \in \Theta. \quad (1.25)$$

Similar results holds for  $\tilde{\alpha}$ :

$$\tilde{\alpha}(a, \theta_1, \theta_2) \geq \frac{\dot{\mu}(a^\top \theta_1)}{2 + |a^\top (\theta_1 - \theta_2)|}, \quad (1.26)$$

$$\geq \frac{\dot{\mu}(a^\top \theta)}{2 + 2S} \quad \text{when } \theta_1, \theta_2 \in \Theta. \quad (1.27)$$

This results into the following matrix inequalities; for any  $\theta_1, \theta_2 \in \Theta$ :

$$\mathbf{G}_t(\theta_1, \theta_2) \succeq (1 + 2S)^{-1} \mathbf{H}_t(\theta) \quad \text{for any } \theta \in \{\theta_1, \theta_2\}, \quad (1.28)$$

$$\tilde{\mathbf{G}}_t(\theta_1, \theta_2) \succeq (2 + 2S)^{-1} \mathbf{H}_t(\theta_1) \quad (1.29)$$

*Proof.* Eq. (1.24) is a direct consequence of Lemma 1.B.1 and Eq. (1.25) is obtained thanks to the bound  $|a^\top (\theta_1 - \theta_2)| \leq 2S$  whenever  $\theta_1, \theta_2 \in \Theta$  (recall that  $\|a\| \leq 1$  by Assumption 1.3.1). Eq. (1.26) and Eq. (1.27) are obtained similarly thanks to Lemma 1.B.2. We can now prove Eq. (1.28); for all  $\theta_1, \theta_2 \in \Theta$  and  $\theta \in \{\theta_1, \theta_2\}$ :

$$\mathbf{G}_t(\theta_1, \theta_2) := \sum_{s=1}^{t-1} \alpha(a_s, \theta_1, \theta_2) a_s a_s^\top + \lambda \mathbf{I}_d \quad (\text{def.})$$

$$\succeq \sum_{s=1}^{t-1} \frac{\dot{\mu}(a_s^\top \theta)}{1 + 2S} a_s a_s^\top + \lambda \mathbf{I}_d \quad (\text{Eq. (1.25)})$$

$$= \frac{1}{1 + 2S} \left( \sum_{s=1}^{t-1} \dot{\mu}(a_s^\top \theta) a_s a_s^\top + (1 + 2S) \lambda \mathbf{I}_d \right)$$

$$\succeq \frac{1}{1 + 2S} \left( \sum_{s=1}^{t-1} \dot{\mu}(a_s^\top \theta) a_s a_s^\top + \lambda \mathbf{I}_d \right) \quad (1 + 2S \geq 1)$$

$$= \frac{1}{1 + 2S} \mathbf{H}_t(\theta) \quad (\text{def. of } \mathbf{H}_t)$$

The proof Eq. (1.29) follows the same reasoning.  $\blacksquare$

We will need one last technical result obtained from the self-concordance property. Its proof can be extracted from [Eq. \(1.33\)](#) in the proof of [Lemma 1.B.1](#).

**Lemma 1.B.3.** *Let  $f$  be a strictly increasing function such that  $|\ddot{f}| \leq \dot{f}$ , and let  $\mathcal{Z}$  be any bounded interval of  $\mathbb{R}$ . Then, for all  $z_1, z_2 \in \mathcal{Z}$ :*

$$\dot{f}(z_2) \exp(-|z_2 - z_1|) \leq f(z_1) \leq \dot{f}(z_2) \exp(|z_2 - z_1|)$$

# Variance-Aware Confidence Sets for Generalized Linear Bandits

---

The goal of this chapter is to derive improved confidence sets for GLBs. We first provide some intuition on a “candidate” set that fits our requirements, obtained by an asymptotical analysis and in a random-design setting. To prove its validity in the general bandit setting we provide a new concentration result based on the theory of self-normalized process: a Bernstein-like tail-inequality for self-normalized martingales. We review its ties, similarities and differences with the concentration inequality of [Abbasi-Yadkori et al. \(2011\)](#) before applying it to the design of an improved confidence set for GLBs. The main feature of this confidence set resides in its local variance sensitivity which captures the effective level of non-linearity in the environment. This feature is central to the rest of our contributions as it allows for a refined local treatment of the non-linearity. We provide several illustrations in the Logistic Bandit setting and present several variants (*e.g* a convex relaxation) as well as an extension to a weighted martingale version which will be used for non-stationary environments.

## Outline

---

2.1	Towards improved confidence sets . . . . .	49
2.2	Bernstein-like tail-inequality for self-normalized martingales . . . . .	50
2.2.1	Result and discussion . . . . .	50
2.2.2	Proof of the main theorem . . . . .	51
2.3	Application to the design of confidence-sets for GLBs . . . . .	54
2.3.1	Confidence set . . . . .	54
2.3.2	A convex relaxation . . . . .	56
2.4	An extension to weighted self-normalized martingales . . . . .	57

---

## 2.1 Towards improved confidence sets

In the previous chapter we introduced the confidence set of [Filippi et al. \(2010\)](#):

$$\mathcal{E}_t(\delta) := \left\{ \theta \in \Theta, \left\| \theta - \tilde{\theta}_t \right\|_{\mathbf{V}_t} \leq \frac{2}{\bar{\ell}_\mu} \left( \sqrt{\lambda \bar{\ell}_\mu} S + \sigma \sqrt{d \log \left( (1 + t \bar{\ell}_\mu / \lambda) / \delta \right)} \right) \right\},$$

which as in the LB case is an ellipsoid, but is now inflated by a factor  $1/\bar{\ell}_\mu$ . As we already mentioned this comes with two important drawbacks; **(1)** the confidence set is prohibitively large and leads to over-explorative algorithms and **(2)** it is not sensitive to the effective non-linearity of the reward signal. The goal of this section is to build some intuition on what a better alternative that fixes those drawbacks could look like.

**An asymptotic argument.** We start by an asymptotic reasoning under a random design, inspired by ([Filippi et al., 2010](#), Section 4.2). We assume that the arms are randomly drawn according to a fixed distribution (*e.g.*  $a \sim \mathcal{N}(0, 1)$ ) and consider a GLB tied to an exponential family - that is such that the reward distribution has a density (w.r.t a given reference measure)  $p(r|a, \theta_\star) \propto \exp(ra^\top \theta_\star - b(a^\top \theta_\star))$ . The Fisher information matrix of this model writes:

$$\mathbf{F}_{\theta_\star} := \mathbb{E}_a \mathbb{E}_r \left[ -\nabla^2 \log p(r|a, \theta) \Big|_{\theta_\star} \right] = \mathbb{E}_a \left[ \dot{\mu}(a^\top \theta_\star) a a^\top \right].$$

Existing results regarding the asymptotic normality of the maximum-likelihood estimator ([Van der Vaart, 2000](#), Section 5.3) yields that ( $\xrightarrow{d}$  indicates convergence in distribution):

$$\sqrt{t} \mathbf{F}_{\theta_\star}^{1/2} \left( \hat{\theta}_t - \theta_\star \right) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d).$$

Recall that  $\mathbf{H}_t(\theta_\star) = \sum_{s=1}^t \dot{\mu}(a_s^\top \theta_\star) a_s a_s^\top + \lambda \mathbf{I}_d$ ; by the law of large numbers we have that  $t^{-1} \mathbf{H}_t(\theta_\star) \xrightarrow{a.s.} \mathbf{F}_{\theta_\star}$ . Henceforth by a direct application of Slutsky's lemma ([Van der Vaart, 2000](#), Lemma 2.8) and the continuous mapping theorem ([Van der Vaart, 2000](#), Theorem 2.3):

$$\left( \hat{\theta}_t - \theta_\star \right)^\top \mathbf{H}_t(\theta_\star) \left( \hat{\theta}_t - \theta_\star \right) \xrightarrow{d} \chi_d^2.$$

Leveraging known tail bounds for Chi-Square random variable ([Laurent and Massart, 2000](#)) this suggests the following asymptotic confidence set at confidence level  $1 - \delta$ :

$$\mathcal{C}_t^\infty(\delta) = \left\{ \theta \in \Theta, \left\| \theta - \hat{\theta}_t \right\|_{\mathbf{H}_t(\theta)} \leq 2\sqrt{d \ln(1/\delta)} \right\}.$$

The matrix  $\mathbf{H}_t(\theta)$  therefore appears as the right metric to measure distance between parameters. This asymptotic confidence interval indeed addresses the two issues we highlighted for  $\mathcal{E}_t(\delta)$ ; **(1)** by using the bound  $\mathbf{H}_t(\theta) \succeq \bar{\ell}_\mu \mathbf{V}_t$  when  $\theta \in \Theta$  one can show that  $\mathcal{C}_t^\infty(\delta)$  is smaller than  $\mathcal{E}_t(\delta)$  by a factor at least  $\bar{\ell}_\mu^{1/2}$ . Further **(2)** this confidence set is sensitive to the effective non-linearity of the reward signal since the metric  $\mathbf{H}_t(\theta)$  accounts for the varying reward sensitivity through the derivative  $\dot{\mu}$  of the link function.

**Towards the general case.** We will focus now on proving a finite-time alternative to  $\mathcal{C}_t^\infty(\delta)$  that is valid under adaptive design - the arms  $\{a_s\}_{s=1}^T$  are far from being independent when generated by a bandit algorithm. To do so we need to resort to adequate concentration inequalities as the design of the confidence set is linked to the control of a sum of random variables. This

link is made explicit by the following set of inequalities (to simplify the exposition we assume for now that  $\hat{\theta}_t \in \Theta$ ):

$$\left\| \hat{\theta}_t - \theta_\star \right\|_{\mathbf{H}_t(\theta_\star)} \leq \sqrt{1 + 2S} \left\| \hat{\theta}_t - \theta_\star \right\|_{\mathbf{G}_t(\theta_\star, \hat{\theta}_t)} \quad (\text{Eq. (1.28)})$$

$$= \sqrt{1 + 2S} \left\| g_t(\hat{\theta}_t) - g_t(\theta_\star) \right\|_{\mathbf{G}_t^{-1}(\theta_\star, \hat{\theta}_t)} \quad (\text{Eq. (1.21)})$$

$$\leq (1 + 2S) \left\| g_t(\hat{\theta}_t) - g_t(\theta_\star) \right\|_{\mathbf{H}_t^{-1}(\theta_\star)} \quad (\text{Eq. (1.28)})$$

$$\leq (1 + 2S) \left( \sqrt{\lambda} S + \left\| \sum_{s=1}^{t-1} \eta_{s+1} a_s \right\|_{\mathbf{H}_t^{-1}(\theta_\star)} \right),$$

where we last used the optimality condition for  $\hat{\theta}_t$  and defined  $\eta_{s+1} := r_{s+1} - \mu(a^\top \theta_\star)$ . Our goal is therefore to control  $\left\| \sum_{s=1}^{t-1} \eta_{s+1} a_s \right\|_{\mathbf{H}_t^{-1}(\theta_\star)}$  without resorting to bounding strategies that involves  $\bar{\ell}_\mu$ . It is useful to gain a bit of intuition on why this should be achievable. The residual noise  $\eta_{s+1}$  is zero-mean by Eq. (1.17) and its conditional variance is such that  $\text{Var}(\eta_{s+1} | \mathcal{F}_s) = \dot{\mu}(a_s^\top \theta_\star)$  by Eq. (1.18). Therefore in the directions  $a \in \mathcal{A}$  such that the following quantity:

$$a^\top \mathbf{H}_t^{-1}(\theta_\star) a \approx \frac{1}{\dot{\mu}(a^\top \theta_\star)} \left( \sum_{s=1}^{t-1} \mathbf{1}(a_s = a) \right)^{-1}$$

is *large*, the conditional variance of the associated residuals is *small* and they concentrate *fast*. Intuitively this argument suggests that  $\left\| \sum_{s=1}^{t-1} \eta_{s+1} a_s \right\|_{\mathbf{H}_t^{-1}(\theta_\star)}$  indeed scales independently of the smallest eigenvalues of  $\mathbf{H}_t(\theta_\star)$  and in particular independently of  $\bar{\ell}_\mu$ . This variance sensitivity argument justifies the name for the formal result presented in the next section as it brings a similar conclusion than the Bernstein concentration inequality (see Lemma A.5).

## 2.2 Bernstein-like tail-inequality for self-normalized martingales

### 2.2.1 Result and discussion

We present here an important technical result at the core of improved confidence sets for GLBs. It extends known results on self-normalized martingales (de la Pena et al., 2004). Its main novelty compared to the concentration inequality from (Abbasi-Yadkori et al., 2011, Theorem 1) resides in considering martingale increments that satisfy a Bernstein-like condition instead of a sub-Gaussian condition. This allows to derive tail-inequalities for martingales “re-normalized” by their quadratic variation.

**Theorem 2.2.1** (Bernstein-like tail-inequality for self-normalized martingales). *Let  $\{\mathcal{F}_t\}_{t=1}^\infty$  be a filtration. Let  $\{a_t\}_{t=1}^\infty$  be a stochastic process in  $\mathcal{B}_d(0, 1)$  such that  $a_t$  is  $\mathcal{F}_t$ -measurable. Let  $\{\eta_t\}_{t=2}^\infty$  be a martingale difference sequence such that  $\eta_{t+1}$  is  $\mathcal{F}_{t+1}$ -measurable. Furthermore, assume that conditionally on  $\mathcal{F}_t$  we have  $|\eta_{t+1}| \leq \sigma$  almost surely and denote  $v_t^2 := \mathbb{E}[\eta_{t+1}^2 | \mathcal{F}_t]$ . Let  $\lambda > 0$  and for any  $t \geq 1$  define:*

$$\mathbf{H}_t := \sum_{s=1}^{t-1} v_s^2 a_s a_s^\top + \lambda \mathbf{I}_d, \quad S_t := \sum_{s=1}^{t-1} \eta_{s+1} a_s.$$

Then for any  $\delta \in (0, 1]$ :

$$\mathbb{P} \left( \forall t \geq 1, \|S_t\|_{\mathbf{H}_t^{-1}} \leq \frac{\sqrt{\lambda}}{2\sigma} + \frac{2\sigma}{\sqrt{\lambda}} \log \left( 2^d \det(\mathbf{H}_t)^{\frac{1}{2}} \lambda^{-\frac{d}{2}} / \delta \right) \right) \geq 1 - \delta.$$

A direct application of the trace-determinant inequality (see [Lemma B.2](#)) along simple manipulations yield a slightly degraded confidence upper-bound.

**Corollary 2.2.1.** *Under the conditions of [Theorem 2.2.1](#) with probability at least  $1 - \delta$ :*

$$\forall t \geq 1, \|S_t\|_{\mathbf{H}_t^{-1}} \leq \frac{\sqrt{\lambda}}{2\sigma} + \frac{\sigma d}{\sqrt{\lambda}} \log\left(4(1 + \sigma^2 t / (d\lambda)) / \delta\right).$$

[Theorem 2.2.1](#) also holds when the regularization is time-dependent.

**Corollary 2.2.2** (Time-varying regularization). *The confidence bound of [Theorem 2.2.1](#) and [Corollary 2.2.1](#) are preserved with time-varying regularization  $\lambda_t$  if the sequence of regularization coefficients  $\{\lambda_t\}_{t=1}^\infty$  is deterministic.*

**Discussion.** The closest inequality of this type was derived by [Abbasi-Yadkori et al. \(2011\)](#) yet [Theorem 2.2.1](#) cannot be recovered by their bound. Indeed introducing  $\omega^2 := \inf_{s \geq 1} v_s^2$ , it can be extracted from their Theorem 1 that with probability at least  $1 - \delta$  for all  $t \geq 1$ :

$$\|S_t\|_{\mathbf{V}_t^{-1}} \leq \sigma \sqrt{2d \log((1 + \omega^2 t / (\lambda d)) / \delta)},$$

where  $\mathbf{V}_t = \sum_{s=1}^{t-1} a_s a_s^\top + (\lambda / \omega^2) \mathbf{I}_d$ . This result can be used to derive a high-probability bound on  $\|S_t\|_{\mathbf{H}_t^{-1}}$ . Noticing that  $\mathbf{H}_t \succeq \omega^2 \mathbf{V}_t$  yields that with probability at least  $1 - \delta$  for all  $t \geq 1$ :

$$\|S_t\|_{\mathbf{H}_t^{-1}} = \mathcal{O}\left(\omega^{-1} \sqrt{d \log(t/\delta)}\right). \quad (2.1)$$

In contrast the bound of [Corollary 2.2.1](#) gives that with probability at least  $1 - \delta$  for all  $t \geq 1$ :

$$\|S_t\|_{\mathbf{H}_t^{-1}} = \mathcal{O}(d \log(t/\delta)).$$

This shaves a multiplicative factor  $1/\omega$  which is large if some  $\eta_s$  have small conditional variance. It is however lagging by a  $\sqrt{d \log(t)}$  factor behind the bound provided in [Eq. \(2.1\)](#). This issue can be fixed by adjusting the regularization parameter. Indeed by setting  $\lambda_t = d \log(2 + t/\delta)$  [Corollary 2.2.2](#) ensures that with probability at least  $1 - \delta$  for all  $t \geq 1$ :

$$\|S_t\|_{\mathbf{H}_t^{-1}} = \mathcal{O}\left(\sqrt{d \log(t/\delta)}\right).$$

In this case [Theorem 2.2.1](#) brings a strict improvement over [Eq. \(2.1\)](#) which involves the minimum conditional variance  $\omega$ .

## 2.2.2 Proof of Theorem 2.2.1

First notice that by normalization it is enough to prove the result for  $\sigma = 1$  and the general result follows simply by replacing  $\lambda$  by  $\lambda/\sigma^2$ . In the following we therefore consider  $\sigma = 1$ . The proof follows the steps of the pseudo-maximization principle introduced in [de la Pena et al. \(2004\)](#), used by [Abbasi-Yadkori et al. \(2011\)](#) for the linear bandit and thoroughly detailed in Chapter 20 of [Lattimore and Szepesvári \(2020\)](#). The main idea is to realize that  $\|S_t\|_{\mathbf{H}_t^{-1}}^2 = 4 \max_{\xi \in \mathbb{R}^d} \xi^\top S_t - \|\xi\|_{\mathbf{H}_t}^2$ . We will show that the exponential of this r.h.s term is (almost) a super-martingale. Unfortunately,  $\exp(\max_{\xi \in \mathbb{R}^d} \xi^\top S_t - \|\xi\|_{\mathbf{H}_t}^2)$  cannot be directly controlled; it can however be approximated by integration over  $\xi$  - a technique known as the Laplace trick or pseudo-maximization. For readability concerns define  $\beta = \sqrt{2\lambda}$  and write:

$$\mathbf{H}_t = \sum_{s=1}^{t-1} v_s^2 a_s a_s^\top + \frac{\beta^2}{2} \mathbf{I}_d = \bar{\mathbf{H}}_t + \frac{\beta^2}{2} \mathbf{I}_d.$$

where  $\bar{\mathbf{H}}_t := \sum_{s=1}^{t-1} v_s^2 a_s a_s^\top$ . For all  $\xi \in \mathbb{R}^d$  let  $M_0(\xi) = 1$  and for  $t \geq 1$  define:

$$M_t(\xi) := \exp\left(\xi^\top S_t - \|\xi\|_{\bar{\mathbf{H}}_t}^2\right).$$

We start the proof by the following intermediary result.

**Lemma 2.2.1** (Extracted from Proposition 3.5 of [Freedman \(1975\)](#)). *Let  $\varepsilon$  be a centered random variable of variance  $v^2$  and such that  $|\varepsilon| \leq 1$  almost surely. Then for all  $u \in [-1, 1]$ :*

$$\mathbb{E}[\exp(u\varepsilon)] \leq 1 + u^2 v^2.$$

This is useful to prove that  $M_t(\xi)$  is a super-martingale for some values of  $\xi$ .

**Lemma 2.2.2.** *For all  $\xi \in \mathcal{B}_2(0, 1)$ ,  $\{M_t(\xi)\}_{t=1}^\infty$  is a  $\{\mathcal{F}_t\}_{t=1}^\infty$ -adapted non-negative super-martingale.*

*Proof.* For all  $t \geq 1$  we have that:

$$\mathbb{E}\left[\exp(\xi^\top S_t) \middle| \mathcal{F}_{t-1}\right] = \exp(\xi^\top S_{t-1}) \mathbb{E}\left[\exp(\xi^\top a_{t-1} \eta_t) \middle| \mathcal{F}_{t-1}\right].$$

By Cauchy-Schwarz  $|\xi^\top a_{t-1}| \leq \|\xi\| \|a_{t-1}\| \leq 1$  and therefore by [Lemma 2.2.1](#):

$$\begin{aligned} \mathbb{E}\left[\exp(\xi^\top S_t) \middle| \mathcal{F}_{t-1}\right] &\leq \exp(\xi^\top S_{t-1}) (1 + v_{t-1}^2 (a_{t-1}^\top \xi)^2), \\ &\leq \exp(\xi^\top S_{t-1} + v_{t-1}^2 (a_{t-1}^\top \xi)^2). \end{aligned} \quad (1 + x \leq e^x)$$

Therefore:

$$\begin{aligned} \mathbb{E}[M_t(\xi) \middle| \mathcal{F}_{t-1}] &= \mathbb{E}\left[\exp\left(\xi^\top S_t - \|\xi\|_{\bar{\mathbf{H}}_t}^2\right) \middle| \mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\left[\exp\left(\xi^\top S_t\right) \middle| \mathcal{F}_{t-1}\right] \exp\left(-\sum_{s=1}^{t-1} v_s^2 (a_s^\top \xi)^2\right) \\ &\leq \exp\left(\xi^\top S_{t-1} + v_{t-1}^2 (a_{t-1}^\top \xi)^2 - \sum_{s=1}^{t-1} v_s^2 (a_s^\top \xi)^2\right) \\ &= M_{t-1}(\xi) \end{aligned}$$

yielding the announced result. ■

Note that  $M_t(\xi)$  is a super-martingale only for  $\xi \in \mathcal{B}_d(0, 1)$ . This is a difference with the approach of [Abbasi-Yadkori et al. \(2011\)](#) (their counterpart for  $M_t(\xi)$  is a super-martingale for any  $\xi \in \mathbb{R}^d$ ) and calls for some refinements when using the Laplace trick to provide a high-probability bound on the maximum of  $\log M_t(\xi)$ .

Let  $h(\xi)$  be a probability density function with support on  $\mathcal{B}_d(0, 1)$  (to be defined later). For  $t \geq 0$  let:

$$\bar{M}_t := \int_{\xi} M_t(\xi) dh(\xi)$$

By [Lemma 20.3 of Lattimore and Szepesvári \(2020\)](#)  $\bar{M}_t$  is also a non-negative super-martingale, and  $\mathbb{E}[\bar{M}_0] = 1$ . Let  $\tau$  be a stopping time with respect to the filtration  $\{F_t\}_{t=0}^\infty$ . We can follow the proof of [Lemma 8 in Abbasi-Yadkori et al. \(2011\)](#) to justify that  $\bar{M}_\tau$  is well-defined (independently of whether  $\tau < \infty$  holds or not) and that  $\mathbb{E}[\bar{M}_\tau] \leq 1$ . Therefore, with  $\delta \in (0, 1)$  and thanks to the maximal inequality:

$$\mathbb{P}\left(\log(\bar{M}_\tau) \geq \log\left(\frac{1}{\delta}\right)\right) = \mathbb{P}\left(\bar{M}_\tau \geq \frac{1}{\delta}\right) \leq \delta \quad (2.2)$$

We can now proceed to compute  $\bar{M}_t$  (more precisely a lower bound on  $\bar{M}_t$ ). Set  $h$  to be the density of a centered normal distribution of isotropic precision  $\beta^2$  truncated on  $\mathcal{B}_d(0, 1)$ . We will denote  $N(h)$  its normalization constant. Simple computations show that:

$$\bar{M}_t = \frac{1}{N(h)} \int_{\mathcal{B}_d(0,1)} \exp\left(\xi^\top S_t - \|\xi\|_{\mathbf{H}_t}^2\right) d\xi$$

To ease notations, let  $f(\xi) := \xi^\top S_t - \|\xi\|_{\mathbf{H}_t}^2$  and  $\xi_* = \arg \max_{\|\xi\| \leq 1/2} f(\xi)$ . Because:

$$f(\xi) = f(\xi_*) + (\xi - \xi_*)^\top \nabla f(\xi_*) - (\xi - \xi_*)^\top \mathbf{H}_t (\xi - \xi_*)$$

we obtain that:

$$\begin{aligned} \bar{M}_t &= \frac{e^{f(\xi_*)}}{N(h)} \int_{\mathbb{R}^d} \mathbf{1}_{\|\xi\| \leq 1} \exp\left((\xi - \xi_*)^\top \nabla f(\xi_*) - (\xi - \xi_*)^\top \mathbf{H}_t (\xi - \xi_*)\right) d\xi \\ &= \frac{e^{f(\xi_*)}}{N(h)} \int_{\mathbb{R}^d} \mathbf{1}_{\|\xi + \xi_*\| \leq 1} \exp\left(\xi^\top \nabla f(\xi_*) - \xi^\top \mathbf{H}_t \xi\right) d\xi && \text{(change of variable } \xi + \xi_*) \\ &\geq \frac{e^{f(\xi_*)}}{N(h)} \int_{\mathbb{R}^d} \mathbf{1}_{\|\xi\| \leq 1/2} \exp\left(\xi^\top \nabla f(\xi_*) - \xi^\top \mathbf{H}_t \xi\right) d\xi && \text{(as } \|\xi_*\| \leq 1/2) \\ &= \frac{e^{f(\xi_*)}}{N(h)} \int_{\mathbb{R}^d} \mathbf{1}_{\|\xi\| \leq 1/2} \exp\left(\xi^\top \nabla f(\xi_*)\right) \exp\left(-\frac{1}{2} \xi^\top (2\mathbf{H}_t) \xi\right) d\xi . \end{aligned}$$

By defining  $g$  the density of the centered normal distribution of precision  $2\mathbf{H}_t$  truncated on the ball  $\{\xi \in \mathbb{R}^d, \|\xi\| \leq 1/2\}$  and noting  $N(g)$  its normalizing constant, we can rewrite:

$$\begin{aligned} \bar{M}_t &\geq \exp(f(\xi_*)) \frac{N(g)}{N(h)} \mathbb{E}_g \left[ \exp\left(\xi^\top \nabla f(\xi_*)\right) \right] \\ &\geq \exp(f(\xi_*)) \frac{N(g)}{N(h)} \exp\left(\mathbb{E}_g \left[\xi^\top \nabla f(\xi_*)\right]\right) && \text{(Jensen's inequality)} \\ &\geq \exp(f(\xi_*)) \frac{N(g)}{N(h)} && \text{(as } \mathbb{E}_g [\xi] = 0). \end{aligned} \quad (2.3)$$

Unpacking this results and assembling (2.2) and (2.3), we obtain that:

$$\begin{aligned} \mathbb{P}\left(\bar{M}_t \geq \frac{1}{\delta}\right) &\geq \mathbb{P}\left(\exp(f(\xi_*)) \frac{N(g)}{N(h)} \geq 1/\delta\right) \\ &= \mathbb{P}\left(\log\left(\exp(f(\xi_*)) \frac{N(g)}{N(h)}\right) \geq \log(1/\delta)\right) \\ &= \mathbb{P}\left(f(\xi_*) \geq \log(1/\delta) + \log\left(\frac{N(h)}{N(g)}\right)\right) \\ &= \mathbb{P}\left(\max_{\|\xi\| \leq 1/2} \xi^\top S_t - \|\xi\|_{\mathbf{H}_t}^2 \geq \log(1/\delta) + \log\left(\frac{N(h)}{N(g)}\right)\right) \\ &\geq \mathbb{P}\left(\xi_0^\top S_t - \|\xi_0\|_{\mathbf{H}_t}^2 \geq \log(1/\delta) + \log\left(\frac{N(h)}{N(g)}\right)\right), \end{aligned} \quad (2.4)$$

for any  $\xi_0$  such that  $\|\xi_0\| \leq 1/2$ . In particular, we can use:

$$\xi_0 := \frac{\mathbf{H}_t^{-1} S_t}{\|S_t\|_{\mathbf{H}_t^{-1}}} \frac{\beta}{2\sqrt{2}},$$

since

$$\|\xi_0\| \leq \frac{\beta}{2\sqrt{2}} \left( \lambda_{\min}(\bar{\mathbf{H}}_t) + \frac{\beta^2}{2} \right)^{-1/2} \leq 1/2.$$

Using this value of  $\xi_0$  in Equation (2.4) yields:

$$\mathbb{P} \left( \|S_t\|_{\mathbf{H}_t^{-1}} \geq \frac{\beta}{2\sqrt{2}} + \frac{2\sqrt{2}}{\beta} \log \left( \frac{N(h)}{\delta N(g)} \right) \right) \leq \mathbb{P} \left( \bar{M}_t \geq \frac{1}{\delta} \right).$$

To finish the proof we have left to explicit the quantities  $N(h)$  and  $N(g)$ . Lemma 2.2.3 provides an upper-bound for the log of their ratio. Its proof is given by straight-forward computations and deferred to Section 2.A.

**Lemma 2.2.3.** *The following inequality holds:*

$$\log \left( \frac{N(h)}{N(g)} \right) \leq \log \left( \frac{2^{d/2} \det(\mathbf{H}_t)^{1/2}}{\beta^d} \right) + d \log(2).$$

Therefore with probability at least  $1 - \delta$  and by using Equation (2.2):

$$\|S_\tau\|_{\mathbf{H}_\tau^{-1}} \leq \frac{\beta}{2\sqrt{2}} + \frac{2\sqrt{2}}{\beta} \log(1/\delta) + \frac{2\sqrt{2}}{\beta} \log \left( \frac{2^{d/2} \det(\mathbf{H}_\tau)^{1/2}}{\beta^d} \right) + \frac{2\sqrt{2}}{\beta} d \log(2).$$

Directly following the stopping time construction argument in the proof of Theorem 1 of Abbasi-Yadkori et al. (2011) we obtain that with probability at least  $1 - \delta$ , for all  $t \in \mathbb{N}$ :

$$\|S_t\|_{\mathbf{H}_t^{-1}} \leq \frac{\beta}{2\sqrt{2}} + \frac{2\sqrt{2}}{\beta} \log \left( \frac{2^{d/2} \det(\mathbf{H}_t)^{1/2}}{\beta^d \delta} \right) + \frac{2\sqrt{2}}{\beta} d \log(2).$$

Finally, recalling that  $\beta = \sqrt{2\lambda}$  and straight-forward factorization provide the announced result. The proofs for Corollary 2.2.1 and Corollary 2.2.2 are deferred to Appendix 2.B.

## 2.3 Application to the design of confidence-sets for GLBs

We now go back to the design of confidence sets for GLBs. To ease exposition we use the slightly looser form of the concentration inequality presented in Corollary 2.2.1, however when used in an algorithm it is always desirable to use the log-determinant form of Theorem 2.2.1.

### 2.3.1 Confidence set

We have seen that time-varying regularization allows to obtain improved dependencies w.r.t  $t$  and  $d$ ; we will use it to obtain tight confidence sets and order-optimal algorithms. In what follows for  $\delta \in (0, 1]$  we set  $\lambda_t = d \log(2 + t/\delta)$  for all  $t \geq 1$  and define:

$$\gamma_t(\delta) = \sqrt{\lambda_t} (S + (2\sigma)^{-1}) + \frac{\sigma d}{\sqrt{\lambda_t}} \log(4(1 + \sigma^2 t / (d\lambda_t)) / \delta) = \mathcal{O} \left( \sqrt{d \log(t/\delta)} \right).$$

A direct applications of our concentration inequality yields the following confidence set, valid for all GLBs satisfying the second-moment condition of Eq. (1.18) (i.e not only self-concordant).

**Theorem 2.3.1** (Variance-sensitive confidence set). *Let  $\delta \in (0, 1]$ . The set:*

$$\mathcal{C}_t(\delta) = \left\{ \theta \in \Theta, \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{H}_t^{-1}(\theta)} \leq \gamma_t(\delta) \right\},$$

*is an anytime confidence set for  $\theta_\star$  at level at least  $1 - \delta$ ;*

$$\mathbb{P}(\forall t \geq 1, \theta_\star \in \mathcal{C}_t(\delta)) \geq 1 - \delta.$$

*Proof.* By using the optimality of  $\hat{\theta}_t$  and defining  $\eta_{s+1} := r_{s+1} - \mu(a^\top \theta_\star)$  for all  $s \geq 1$  simple upper-bounding yields:

$$\left\| g_t(\hat{\theta}_t) - g_t(\theta_\star) \right\|_{\mathbf{H}_t^{-1}(\theta_\star)} \leq \sqrt{\lambda_t} S + \left\| \sum_{s=1}^{t-1} \eta_{s+1} a_s \right\|_{\mathbf{H}_t^{-1}(\theta_\star)}$$

By Eqs. (1.17) and (1.18) and Assumption 1.3.2 the most r.h.s term satisfy all conditions of Theorem 2.2.1. Applying Corollary 2.2.2 yields that with probability at least  $1 - \delta$ ;

$$\forall t \geq 1, \left\| g_t(\hat{\theta}_t) - g_t(\theta_\star) \right\|_{\mathbf{H}_t^{-1}(\theta_\star)} \leq \gamma_t(\delta).$$

which proves the announced result. ■

The confidence set  $\mathcal{C}_t(\delta)$  is not the easiest to manipulate and visualize because it involves the map  $g_t$ . We introduce in Corollary 2.3.1 a marginally modified version for self-concordant GLBs, more intuitive and easier to compare with previous confidence sets. We will replace  $\hat{\theta}_t$  by its adequate “projection” on  $\Theta$ :

$$\tilde{\theta}_t := \arg \min_{\theta \in \Theta} \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{H}_t^{-1}(\theta)}. \quad (2.5)$$

This projection mirrors the one of Eq. (1.14) necessary for the design of the confidence set of (Filippi et al., 2010). Similarly it is made trivial when  $\hat{\theta}_t \in \Theta$ . It is the center of the following confidence region, which allows for a neat comparison with the one from Theorem 1.3.1.

**Corollary 2.3.1.** *Let  $\delta \in (0, 1]$ . Under Assumption 1.4.1 the set:*

$$\mathcal{C}'_t(\delta) = \left\{ \theta \in \Theta, \left\| \theta - \tilde{\theta}_t \right\|_{\mathbf{H}_t(\theta)} \leq 2(1 + 2S)\gamma_t(\delta) \right\}, \quad (2.6)$$

*is such that  $\mathcal{C}_t(\delta) \subseteq \mathcal{C}'_t(\delta)$  and therefore is also an anytime confidence set for  $\theta_\star$ .*

The proof is already sketched in Section 2.1 and deferred it to Appendix 2.C. Corollary 2.3.1 brings a positive answer to our goal from Section 2.1: derive a finite-time, adaptive-design version of the asymptotic confidence set  $\mathcal{C}_t^\infty(\delta)$ . It brings a significant improvement over the confidence set  $\mathcal{E}_t(\delta)$  of Filippi et al. (2010). Indeed, in the *worst-case* (w.r.t to the sequence of arms played) for which  $\mathbf{H}_t(\theta) = \bar{\ell}_\mu \mathbf{V}_t$ , the sets  $\mathcal{E}_t(\delta)$  and  $\mathcal{C}'_t(\delta)$  have the same shape but the radius of the former is smaller than the latter’s by a factor  $\approx \bar{\ell}_\mu^{1/2}$  which is typically *exponentially* small w.r.t  $S$ . This is illustrated in Fig. 2.1 in the LogB setting.

**Remark 2.3.1** (Proving the conjecture of Filippi et al. (2010)). *In their section 4.2 Filippi et al. (2010) relied on an asymptotical argument to motivate a heuristic algorithm which exploration bonus is deflated by a factor  $\bar{\ell}_\mu^{1/2}$  compared to GLM-UCB. More precisely, they argue that under a*

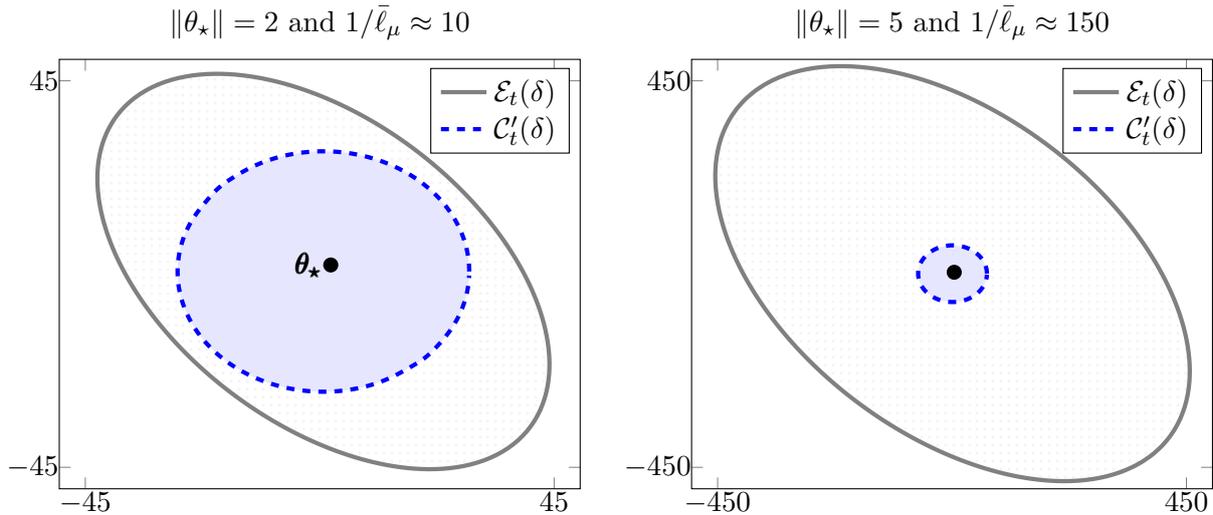


Figure 2.1: (✂) Visualization of two-dimensional Logistic bandit confidence sets. The confidence sets are generated by the same trajectory  $\{a_s\}_{s=1}^{200}$  but on different environments. On the first one (*left*)  $\|\theta_\star\|$  is moderate and the level of non-linearity is small; the confidence set  $\mathcal{E}_t(\delta)$  and  $\mathcal{C}'_t(\delta)$  are comparable. On the second one (*right*)  $\|\theta_\star\|$  is large and the reward signal is highly non-linear. In this configuration  $\mathcal{C}'_t(\delta)$  is much tighter than  $\mathcal{E}_t(\delta)$ . In the LogB case the difference in diameter between the two confidence sets grows exponentially with  $\|\theta_\star\|$ .

random design the prediction error  $\Delta_t(a)$  can be asymptotically bounded by  $\tilde{\mathcal{O}}(\bar{L}_\mu/\sqrt{\bar{\ell}_\mu})$ . *Corollary 2.3.1* shows that this still holds non-asymptotically and under an adaptive design. This gives the first formal theoretical justification for the heuristic algorithm of *Filippi et al. (2010)*. Indeed by linearization and the Cauchy-Schwarz inequality;

$$\begin{aligned} \Delta_t(a) &= |\mu(a^\top \tilde{\theta}_t) - \mu(a^\top \theta_\star)| \leq \bar{L}_\mu \|a\|_{\mathbf{H}_t^{-1}(\theta_\star)} \left\| \tilde{\theta}_t - \theta_\star \right\|_{\mathbf{H}_t^{-1}(\theta_\star)}, \\ &\leq 2(1 + 2S) \bar{L}_\mu \bar{\ell}_\mu^{-1/2} \gamma_t(\delta) \|a\|_{\mathbf{V}_t^{-1}}, \end{aligned}$$

where we last used  $\mathbf{H}_t(\theta_\star) \succeq \bar{\ell}_\mu \mathbf{V}_t$  and *Corollary 2.3.1*. It turns out that this is actually a degradation of our confidence set's properties and that improved prediction bounds are obtainable - this is the main point of the following chapter.

### 2.3.2 A convex relaxation

The confidence sets  $\mathcal{C}_t(\delta)$  and  $\mathcal{C}'_t(\delta)$  from *Theorem 2.3.1* improve over  $\mathcal{E}_t(\delta)$  but can be challenging to incorporate in one's algorithmic design as they are in all generality *non-convex* sets (see *Fig. 2.2*). The following theorem shows that for self-concordant GLBs they can be relaxed into a convex confidence set, only at the cost of a minor radius inflation. This result will be important to derive tractable algorithms. Define for all  $t \geq 1$ :

$$\beta_t(\delta) := \gamma_t(\delta) + \gamma_t(\delta)^2/\sqrt{\lambda_t},$$

which is a  $\mathcal{O}(\sqrt{d \log(t/\delta)})$  since  $\lambda_t = d \log(2 + t/\delta)$ .

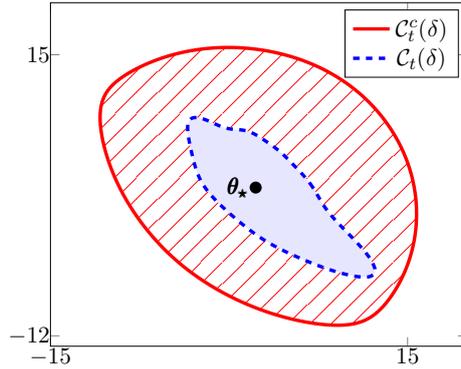


Figure 2.2:  The confidence set  $\mathcal{C}_t(\delta)$  and its convex relaxation  $\mathcal{C}_t^c(\delta)$  on a LogB problem with  $\|\theta_\star\| = 3$ . The sequence of arm  $\{a_s\}_{s=1}^{1000}$  is chosen to highlight the non-convexity of  $\mathcal{C}_t(\delta)$ .

**Corollary 2.3.2** (Convex relaxation). *Let  $\delta \in (0, 1]$ . For all  $t \geq 1$  the set:*

$$\mathcal{C}_t^c(\delta) := \left\{ \theta \in \Theta, \mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t) \leq \beta_t(\delta)^2 \right\},$$

is convex and under [Assumption 1.4.1](#) satisfies:

- (1)  $\mathcal{C}_t(\delta) \subseteq \mathcal{C}_t^c(\delta)$  i.e  $\mathcal{C}_t^c(\delta)$  is an anytime confidence set for  $\theta_\star$  at level  $1 - \delta$ .
- (2) With probability at least  $1 - \delta$  and  $\forall \theta \in \mathcal{C}_t^c(\delta)$ :

$$\|\theta - \theta_\star\|_{\mathbf{H}_t(\theta_\star)} \leq (2 + 2S)\gamma_t(\delta) + 2\sqrt{1 + S}\beta_t(\delta).$$

This result essentially allows to replace  $\mathcal{C}_t(\delta)$  by its convex counterpart  $\mathcal{C}_t^c(\delta)$  while conserving the similar guarantees. Indeed  $\mathcal{C}_t^c(\delta)$  is a confidence set with a similar diameter since by (2) it measures a similar deviation w.r.t  $\theta_\star$  as  $\mathcal{C}_t(\delta)$  since with high probability:

$$\forall \theta \in \mathcal{C}_t^c(\delta), \|\theta - \theta_\star\|_{\mathbf{H}_t(\theta_\star)} = \mathcal{O}\left(\sqrt{d \log(t/\delta)}\right).$$

The proof mostly involves self-concordance inequalities and is deferred to [Appendix 2.D](#).

## 2.4 An extension to weighted self-normalized martingales

In this section we present an extension of [Theorem 2.2.1](#) to *weighted* self-normalized martingales. Slightly anticipating on [Chapter 4](#) it will be useful in non-stationary settings where it is important to re-weight samples according to their freshness. This allows to “forget” old interactions which signal might have become meaningless given the non-stationary nature of the environment. Formally, consider a sequence of weights  $\{w_s\}_{s=1}^T$ . We make the following assumption on this sequence, which for instance fit exponential weights  $w_s = \gamma^{T-s}$  for  $\gamma > 0$ .

**Assumption 2.4.1** (Admissible weights). *The weights  $\{w_s\}_{s=1}^T$  are deterministic, strictly positive and non-decreasing:*

$$0 < w_1 \leq w_t \leq w_{t+1} \leq w_T \quad \text{for all } 1 \leq t \leq T - 1.$$

[Theorem 2.4.1](#) to follow provides a high-confidence bound for the weighted martingale  $S_t = \sum_{s=1}^{t-1} w_s \eta_{s+1} a_s$  re-normalized by its quadratic variation.

**Theorem 2.4.1** (Bernstein-like tail-inequality for weighted self-normalized martingales). *Let  $T$  be a known integer and  $\{\mathcal{F}_t\}_{t=1}^T$  a filtration. Let  $\{a_t\}_{t=1}^T$  be a stochastic process on  $\mathcal{B}_2(0, 1)$  such that  $a_t$  is  $\mathcal{F}_t$ -measurable. Let  $\{\eta_t\}_{t=2}^T$  be a martingale difference sequence such that  $\eta_{t+1}$  is  $\mathcal{F}_{t+1}$ -measurable. Let  $\{w_t\}_{t=1}^T$  a sequence of weights satisfying [Assumption 2.4.1](#) and  $\{\lambda_t\}_{t=1}^T$  be a deterministic, strictly positive sequence of regularization terms. Furthermore, assume that conditionally on  $\mathcal{F}_t$  we have  $|\eta_{t+1}| \leq \sigma$  a.s. and denote  $v_t^2 = \mathbb{E}[\eta_{t+1}^2 | \mathcal{F}_t]$ . For all  $t \in [T]$  define:*

$$\tilde{\mathbf{H}}_t = \sum_{s=1}^{t-1} w_s^2 v_s^2 a_s a_s^\top + \lambda_t \mathbf{I}_d \quad \text{and} \quad S_t = \sum_{s=1}^{t-1} w_s \eta_{s+1} a_s .$$

Then for any  $\delta \in (0, 1]$  and any fixed  $t \in [T]$ :

$$\mathbb{P} \left( \|S_t\|_{\tilde{\mathbf{H}}_t^{-1}} \leq \frac{\sqrt{\lambda_t}}{2\sigma w_{t-1}} + \frac{2\sigma w_{t-1}}{\sqrt{\lambda_t}} \log \left( \frac{2^d \det(\tilde{\mathbf{H}}_t)^{1/2}}{\delta \lambda_t^{d/2}} \right) \right) \geq 1 - \delta .$$

A few remarks are in order. First note that the high-confident bound of [Theorem 2.4.1](#) is not anytime but holds only for a fixed  $t$ . A union bound is required to ensure it holds simultaneously over all  $t \in [T]$ . It could appear at first that [Theorem 2.4.1](#) is a direct corollary of [Theorem 2.2.1](#) obtained by redefining  $\tilde{\eta}_{s+1} = w_s \eta_{s+1}$  and using the normalization property of the self-normalized martingale. This approach ultimately yields a much looser bound that will replace  $w_{t-1}$  in [Theorem 2.4.1](#) by  $\min_{s \in [T]} w_{s-1}$ . This is problematic when the latter is much smaller than the former - e.g with exponential weights, this leads an exponentially degraded bound. With this technical difficulty in mind the proof of [Theorem 2.2.1](#) can still be re-used; the main “trick” is to study the super-martingale:

$$M_s^t(\xi) := \exp \left( (mw_{t-1})^{-1} \xi^\top S_s - (mw_{t-1})^{-2} \xi^\top \left( \sum_{u=1}^{s-1} w_u^2 v_u^2 a_u a_u^\top \right) \xi \right)$$

where  $t$  is fixed. This however rules out the stopping time argument of [Abbasi-Yadkori et al. \(2011\)](#) hence the loss of the anytime behavior of the resulting confidence bound. The formal proof is deferred to [Appendix 2.E](#). A simple union bound yields the following anytime bound.

**Corollary 2.4.1.** *Under the conditions of [Theorem 2.4.1](#), with probability at least  $1 - \delta$ :*

$$\forall t \in [T], \quad \|S_t\|_{\tilde{\mathbf{H}}_t^{-1}} \leq \frac{\sqrt{\lambda_t}}{2\sigma w_{t-1}} + \frac{2\sigma w_{t-1}}{\sqrt{\lambda_t}} \log \left( \frac{2^d T \det(\tilde{\mathbf{H}}_t)^{1/2}}{\delta \lambda_t^{d/2}} \right)$$

# Appendix

## Appendix 2.A Proof of Lemma 2.2.3

**Lemma 2.2.3.** *The following inequality holds:*

$$\log \left( \frac{N(h)}{N(g)} \right) \leq \log \left( \frac{2^{d/2} \det(\mathbf{H}_t)^{1/2}}{\beta^d} \right) + d \log(2).$$

*Proof.* Recall that  $h$  is the density of a  $d$ -dimensional, centered normal distribution with isotropic precision  $\beta^2$  and truncated on  $\mathcal{B}_d(0, 1)$ . The quantity  $N(h)$  is its normalization constant and thanks to a change of variable:

$$\begin{aligned} N(h) &= \int_{\mathbb{R}^d} \mathbf{1} [\|\xi\|_2 \leq 1] \exp \left( -\frac{1}{2} \beta^2 \|\xi\|_2^2 \right) d\xi \\ &= \beta^{-d} \int_{\mathbb{R}^d} \mathbf{1} [\|\xi\|_2 \leq \beta] \exp \left( -\frac{1}{2} \|\xi\|_2^2 \right) d\xi \end{aligned}$$

On the other hand  $g$  is the density of the centered normal distribution with precision matrix  $2\mathbf{H}_t$  truncated on  $\mathcal{B}_d(0, 1/2)$ . Also by a change of variable

$$\begin{aligned} N(g) &= \int_{\mathbb{R}^d} \mathbf{1} [\|\xi\|_2 \leq 1/2] \exp \left( -\frac{1}{2} \xi^\top (2\mathbf{H}_t) \xi \right) d\xi \\ &= \det(\mathbf{H}_t)^{-1/2} 2^{-d/2} \int_{\mathbb{R}^d} \mathbf{1} [\|2^{-1/2} \mathbf{H}_t^{-1/2} \xi\|_2 \leq 1/2] \exp \left( -\frac{1}{2} \|\xi\|_2^2 \right) \\ &= \det(\mathbf{H}_t)^{-1/2} 2^{-d/2} \int_{\mathbb{R}^d} \mathbf{1} \left[ \left\| \left( \bar{\mathbf{H}}_t + \frac{\beta^2}{2} \mathbf{I}_d \right)^{-1/2} \xi \right\|_2 \leq 1/\sqrt{2} \right] \exp \left( -\frac{1}{2} \|\xi\|_2^2 \right) \\ &\geq \det(\mathbf{H}_t)^{-1/2} 2^{-d/2} \int_{\mathbb{R}^d} \mathbf{1} [\|\xi\|_2 \leq \beta/2] \exp \left( -\frac{1}{2} \|\xi\|_2^2 \right) \end{aligned}$$

We obtain the following upper-bound on the ratio  $N(h)/N(g)$ :

$$\frac{N(h)}{N(g)} \leq \beta^{-d} \det(\mathbf{H}_t)^{1/2} 2^{d/2} \frac{\int_{\mathbb{R}^d} \mathbf{1} [\|\xi\|_2 \leq \beta] \exp \left( -\frac{1}{2} \|\xi\|_2^2 \right) d\xi}{\int_{\mathbb{R}^d} \mathbf{1} [\|\xi\|_2 \leq \beta/2] \exp \left( -\frac{1}{2} \|\xi\|_2^2 \right) d\xi} \quad (2.7)$$

Note that:

$$\begin{aligned} \frac{\int_{\mathbb{R}^d} \mathbf{1} [\|\xi\|_2 \leq \beta] \exp \left( -\frac{1}{2} \|\xi\|_2^2 \right) d\xi}{\int_{\mathbb{R}^d} \mathbf{1} [\|\xi\|_2 \leq \beta/2] \exp \left( -\frac{1}{2} \|\xi\|_2^2 \right) d\xi} &= 1 + \frac{\int_{\mathbb{R}^d} \mathbf{1} [\beta/2 \leq \|\xi\|_2 \leq \beta] \exp \left( -\frac{1}{2} \|\xi\|_2^2 \right) d\xi}{\int_{\mathbb{R}^d} \mathbf{1} [\|\xi\|_2 \leq \beta/2] \exp \left( -\frac{1}{2} \|\xi\|_2^2 \right) d\xi} \\ &\leq 1 + \frac{\exp \left( -\frac{1}{8} \beta^2 \right)}{\exp \left( -\frac{1}{8} \beta^2 \right)} \cdot \frac{\int_{\mathbb{R}^d} \mathbf{1} [\beta/2 \leq \|\xi\|_2 \leq \beta] d\xi}{\int_{\mathbb{R}^d} \mathbf{1} [\|\xi\|_2 \leq \beta/2] d\xi} \\ &= 1 + \frac{\mathcal{V}_d(\beta) - \mathcal{V}_d(\beta/2)}{\mathcal{V}_d(\beta/2)} \\ &= 2^d \end{aligned}$$

where  $\mathcal{V}_d(\beta) \propto \beta^d$  denotes the volume of the  $d$ -dimensional ball of radius  $\beta$ . Plugging this result in Equation (2.7) and taking the logarithm yields the announced result:

$$\log \left( \frac{N(h)}{N(g)} \right) \leq \log \left( \frac{2^{d/2} \det(\mathbf{H}_t)^{1/2}}{\beta^d} \right) + d \log(2).$$

■

## Appendix 2.B Proof of corollaries of Theorem 2.2.1

### 2.B.1 Proof of Corollary 2.2.1

**Corollary 2.2.1.** *Under the conditions of [Theorem 2.2.1](#) with probability at least  $1 - \delta$ :*

$$\forall t \geq 1, \|S_t\|_{\mathbf{H}_t^{-1}} \leq \frac{\sqrt{\lambda}}{2\sigma} + \frac{\sigma d}{\sqrt{\lambda}} \log\left(4(1 + \sigma^2 t / (d\lambda)) / \delta\right).$$

*Proof.* A simple application of the trace-determinant inequality (see [Lemma B.2](#)) yields that:

$$\log \det(\mathbf{H}_t) \leq d \log(\lambda + t\sigma^2/d)$$

by using the fact that  $v_t^2 \leq \sigma^2$  for any  $t \geq 2$ . Rearranging in [Theorem 2.2.1](#) along with simple upper-bounding yields the result.  $\blacksquare$

### 2.B.2 Proof of Corollary 2.2.2

We state and prove below the formalization of [Corollary 2.2.2](#). The result is obtained by following an idea from ([Russac et al., 2019](#), Proposition 1).

**Proposition 2.B.1** (Formalization of [Corollary 2.2.2](#)). *Let  $\{\mathcal{F}_t\}_{t=1}^\infty$  be a filtration. Let  $\{a_t\}_{t=1}^\infty$  be a stochastic process in  $\mathcal{B}_d(0, 1)$  such that  $a_t$  is  $\mathcal{F}_t$ -measurable. Let  $\{\eta_t\}_{t=2}^\infty$  be a martingale difference sequence such that  $\eta_{t+1}$  is  $\mathcal{F}_{t+1}$ -measurable. Furthermore, assume that conditionally on  $\mathcal{F}_t$  we have  $|\eta_{t+1}| \leq \sigma$  almost surely and denote  $v_t^2 := \mathbb{E}[\eta_{t+1}^2 | \mathcal{F}_t]$ . Let  $\{\lambda_t\}_{t=1}^\infty$  be a predictable sequence of non-negative scalars and for any  $t \geq 1$  define:*

$$\mathbf{H}_t := \sum_{s=1}^{t-1} v_s^2 a_s a_s^\top + \lambda_t \mathbf{I}_d, \quad S_t := \sum_{s=1}^{t-1} \eta_{s+1} a_s.$$

Then for any  $\delta \in (0, 1]$ :

$$\mathbb{P}\left(\forall t \geq 1, \|S_t\|_{\mathbf{H}_t^{-1}} \leq \frac{\sqrt{\lambda_t}}{2\sigma} + \frac{2\sigma}{\sqrt{\lambda_t}} \log\left(2^d \det(\mathbf{H}_t)^{\frac{1}{2}} \lambda_t^{-\frac{d}{2}} / \delta\right)\right) \geq 1 - \delta.$$

*Proof.* The proof essentially follows the proof of [Theorem 2.2.1](#) up to a minor modification to allow for a time-varying regularization. Re-using notations from [Section 2.2.2](#):

$$M_0(\xi) = 1 \quad \text{and} \quad M_t(\xi) := \exp\left(\xi^\top S_t - \|\xi\|_{\mathbf{H}_t}^2\right) \quad \forall t \geq 1.$$

Recall that  $M_t(\xi)$  is a super-martingale and hence checks  $\mathbb{E}[M_t(\xi)] \leq 1$  for all  $\xi \in \mathcal{B}_d(0, 1)$ . Further, let  $g_t(\xi)$  be the density of the normal distribution of precision  $2\mathbf{H}_t$  truncated on the ball  $\mathcal{B}_d(0, 1/2)$  and let:

$$\bar{M}_t = \int M_t(\xi) g_t(\xi) d\xi.$$

Note that  $\bar{M}_t$  is not (in all generality) a super-martingale - this is where the analysis changes. This however doesn't hurt the final result as one can still apply an appropriate stopping time construction. Let  $\tau$  be a stopping time with respect to  $\{\mathcal{F}_t\}_{t=1}^\infty$ . One can easily check (see for instance the proof of Theorem 1 in [Abbasi-Yadkori et al. \(2011\)](#)) that  $M_\tau(\xi)$  is well-defined and  $\mathbb{E}[M_\tau(\xi)] \leq 1$  for all  $\xi \in \mathcal{B}_d(0, 1/2)$ . Clearly we have:

$$\mathbb{E}[\bar{M}_\tau] = \int \mathbb{E}[M_\tau(\xi)] g_\tau(\xi) d\xi \leq 1.$$

Computing  $\bar{M}_\tau$  following the proof of [Theorem 2.2.1](#) eventually leads us to:

$$\mathbb{P} \left( \|S_\tau\|_{\mathbf{H}_\tau} \leq \frac{\sqrt{\lambda_\tau}}{2} + \frac{2}{\sqrt{\lambda_\tau}} \log \left( \frac{2^d \det(\mathbf{H}_\tau)^{1/2}}{\delta \lambda_\tau^{d/2}} \right) \right) \geq 1 - \delta .$$

From there, directly following the stopping time construction in the proof of [Theorem 1](#) in [Abbasi-Yadkori et al. \(2011\)](#) yields the announced result. ■

## Appendix 2.C Proof of Corollary 2.3.1

**Corollary 2.3.1.** *Let  $\delta \in (0, 1]$ . Under [Assumption 1.4.1](#) the set:*

$$\mathcal{C}'_t(\delta) = \left\{ \theta \in \Theta, \left\| \theta - \tilde{\theta}_t \right\|_{\mathbf{H}_t(\theta)} \leq 2(1 + 2S)\gamma_t(\delta) \right\} , \quad (2.6)$$

is such that  $\mathcal{C}_t(\delta) \subseteq \mathcal{C}'_t(\delta)$  and therefore is also an anytime confidence set for  $\theta_\star$ .

*Proof.* Let  $\theta \in \Theta \cap \mathcal{C}_t(\delta)$ . Then for all  $t \geq 1$ :

$$\begin{aligned} \left\| \theta - \tilde{\theta}_t \right\|_{\mathbf{H}_t(\theta)} &\leq \sqrt{1 + 2S} \left\| \theta - \tilde{\theta}_t \right\|_{\mathbf{G}_t(\theta, \tilde{\theta}_t)} && \text{(Eq. (1.28))} \\ &= \sqrt{1 + 2S} \left\| g_t(\theta) - g_t(\tilde{\theta}_t) \right\|_{\mathbf{G}_t(\theta, \tilde{\theta}_t)} && \text{(Eq. (1.21))} \\ &= \sqrt{1 + 2S} \left( \left\| g_t(\tilde{\theta}_t) - g_t(\hat{\theta}_t) \right\|_{\mathbf{G}_t(\theta, \tilde{\theta}_t)} + \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{G}_t(\theta, \tilde{\theta}_t)} \right) \\ &\leq (1 + 2S) \left( \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{H}_t^{-1}(\theta)} + \left\| g_t(\tilde{\theta}_t) - g_t(\hat{\theta}_t) \right\|_{\mathbf{H}_t^{-1}(\tilde{\theta}_t)} \right) && \text{(Eq. (1.28))} \\ &\leq 2(1 + 2S) \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{H}_t^{-1}(\theta)} && \text{(Eq. (2.5))} \\ &\leq 2(1 + 2S)\gamma_t(\delta) && (\theta \in \mathcal{C}_t(\delta)) \end{aligned}$$

which proves that  $\theta \in \mathcal{C}'_t(\delta)$  and the announced result. ■

## Appendix 2.D Proof of Corollary 2.3.2

**Corollary 2.3.2** (Convex relaxation). *Let  $\delta \in (0, 1]$ . For all  $t \geq 1$  the set:*

$$\mathcal{C}_t^c(\delta) := \left\{ \theta \in \Theta, \mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t) \leq \beta_t(\delta)^2 \right\} ,$$

is convex and under [Assumption 1.4.1](#) satisfies:

- (1)  $\mathcal{C}_t(\delta) \subseteq \mathcal{C}_t^c(\delta)$  i.e  $\mathcal{C}_t^c(\delta)$  is an anytime confidence set for  $\theta_\star$  at level  $1 - \delta$ .
- (2) With probability at least  $1 - \delta$  and  $\forall \theta \in \mathcal{C}_t^c(\delta)$ :

$$\left\| \theta - \theta_\star \right\|_{\mathbf{H}_t(\theta_\star)} \leq (2 + 2S)\gamma_t(\delta) + 2\sqrt{1 + S}\beta_t(\delta) .$$

*Proof.* We start by proving that  $\mathcal{C}_t(\delta) \subseteq \mathcal{C}_t^c(\delta)$ . First, we claim [Lemma 2.D.1](#), which proof is deferred to [Section 2.D.1](#).

**Lemma 2.D.1.** *Let  $\delta \in (0, 1]$ . For all  $\theta \in \mathcal{C}_t(\delta)$ :*

$$\left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{G}_t^{-1}(\theta, \hat{\theta}_t)} \leq \frac{\gamma_t^2(\delta)}{\sqrt{\lambda_t}} + \gamma_t(\delta) .$$

By an exact second-order Taylor expansion of the log-loss (see Eq. (1.22)) for all  $\theta \in \mathbb{R}^d$ :

$$\mathcal{L}_t(\theta) = \mathcal{L}_t(\hat{\theta}_t) + \nabla \mathcal{L}_t|_{\hat{\theta}_t}^\top (\theta - \hat{\theta}_t) + (\theta - \hat{\theta}_t)^\top \tilde{\mathbf{G}}_t(\theta, \hat{\theta}_t)(\theta - \hat{\theta}_t)$$

By definition of  $\hat{\theta}_t$  we have that  $\nabla \mathcal{L}_t|_{\hat{\theta}_t} = 0$  and therefore:

$$\begin{aligned} \mathcal{L}_t(\theta) &= \mathcal{L}_t(\hat{\theta}_t) + \left\| \theta - \hat{\theta}_t \right\|_{\tilde{\mathbf{G}}_t(\hat{\theta}_t, \theta)}^2 \\ &\leq \mathcal{L}_t(\hat{\theta}_t) + \left\| \theta - \hat{\theta}_t \right\|_{\mathbf{G}_t(\hat{\theta}_t, \theta)}^2 && (\tilde{\mathbf{G}}_t \leq \mathbf{G}_t) \\ &= \mathcal{L}_t(\hat{\theta}_t) + \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{G}_t^{-1}(\hat{\theta}_t, \theta)}^2 && (\text{Equation (1.21)}) . \\ &= \mathcal{L}_t(\hat{\theta}_t) + \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{G}_t^{-1}(\theta, \hat{\theta}_t)}^2 && (\mathbf{G}_t(\hat{\theta}_t, \theta) = \mathbf{G}_t(\theta, \hat{\theta}_t)) . \end{aligned}$$

Therefore for any  $\theta \in \mathcal{C}_t(\delta)$ :

$$\begin{aligned} \mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t) &\leq \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{G}_t^{-1}(\theta, \hat{\theta}_t)}^2 \\ &\leq \left( \frac{\gamma_t^2(\delta)}{\sqrt{\lambda_t}} + \gamma_t(\delta) \right)^2 = \beta_t(\delta)^2 && (\text{Lemma 2.D.1}) . \end{aligned}$$

proving that  $\theta \in \mathcal{C}_t(\delta) \Rightarrow \theta \in \mathcal{C}_t^c(\delta)$  and therefore  $\mathcal{C}_t(\delta) \subseteq \mathcal{C}_t^c(\delta)$ . We now prove the second part of Corollary 2.3.2. We will assume that  $\{\theta_\star \in \mathcal{C}_t(\delta)\}$  which happens with probability at least  $1 - \delta$ . We rely on the following second-order Taylor expansion. For all  $\theta \in \mathcal{C}_t^c(\delta)$ :

$$\mathcal{L}_t(\theta) = \mathcal{L}_t(\theta_\star) + (\theta - \theta_\star)^\top \nabla \mathcal{L}_t(\theta_\star) + \left\| \theta - \theta_\star \right\|_{\tilde{\mathbf{G}}_t(\theta_\star, \theta)}^2$$

Therefore:

$$\begin{aligned} \mathcal{L}_t(\theta) - \mathcal{L}_t(\theta_\star) - (\theta - \theta_\star)^\top \nabla \mathcal{L}_t(\theta_\star) &= \left\| \theta - \theta_\star \right\|_{\tilde{\mathbf{G}}_t(\theta_\star, \theta)}^2 \\ &\geq (2 + 2S)^{-1} \left\| \theta - \theta_\star \right\|_{\mathbf{H}_t(\theta_\star)}^2 && (\text{Eq. (1.29)}) \end{aligned}$$

which can be rewritten as:

$$\begin{aligned} \left\| \theta - \theta_\star \right\|_{\mathbf{H}_t(\theta_\star)}^2 &\leq (2 + 2S) |\mathcal{L}_t(\theta) - \mathcal{L}_t(\theta_\star)| + (2 + 2S) \left| (\theta - \theta_\star)^\top \nabla \mathcal{L}_t(\theta_\star) \right| \\ &\leq 2(2 + 2S) \beta_t(\delta)^2 + (2 + 2S) \left| (\theta - \theta_\star)^\top \nabla \mathcal{L}_t(\theta_\star) \right| && (\theta, \theta_\star \in \mathcal{C}_t^c(\delta)) \\ &\leq 2(2 + 2S) \beta_t(\delta)^2 + (2 + 2S) \left\| \theta - \theta_\star \right\|_{\mathbf{H}_t(\theta_\star)} \left\| \nabla \mathcal{L}_t(\theta_\star) \right\|_{\mathbf{H}_t^{-1}(\theta_\star)} \\ &\leq 2(2 + 2S) \beta_t(\delta)^2 + (2 + 2S) \gamma_t(\delta) \left\| \theta - \theta_\star \right\|_{\mathbf{H}_t(\theta_\star)} \end{aligned}$$

where we last used:

$$\begin{aligned} \left\| \nabla \mathcal{L}_t(\theta_\star) \right\|_{\mathbf{H}_t^{-1}(\theta_\star)} &= \left\| g_t(\theta_\star) - \sum_{s=1}^{t-1} r_{s+1} a_s \right\|_{\mathbf{H}_t^{-1}(\theta_\star)} \\ &= \left\| g_t(\theta_\star) - g_t(\hat{\theta}_t) \right\|_{\mathbf{H}_t^{-1}(\theta_\star)} \\ &\leq \gamma_t(\delta) . && (\theta_\star \in \mathcal{C}_t(\delta)) \end{aligned}$$

To sum-up, we have the following polynomial inequality on  $\|\theta - \theta_\star\|_{\mathbf{H}_t(\theta_\star)}$ :

$$\|\theta - \theta_\star\|_{\mathbf{H}_t(\theta_\star)}^2 \leq 2(2 + 2S)\beta_t(\delta)^2 + (2 + 2S)\gamma_t(\delta) \|\theta - \theta_\star\|_{\mathbf{H}_t(\theta_\star)} .$$

Solving it (cf. [Proposition B.1](#)) yields:

$$\|\theta - \theta_\star\|_{\mathbf{H}_t(\theta_\star)} \leq (2 + 2S)\gamma_t(\delta) + 2\sqrt{1 + S}\beta_t(\delta) .$$

Finally recall  $\lambda_t = d \log(2 + t/\delta)$  we obtain the following scalings:

$$\begin{aligned} \gamma_t(\delta) &= \mathcal{O}(\sqrt{d \log(t/\delta)}) , \\ \beta_t(\delta) &= \gamma_t(\delta) + \gamma_t^2(\delta)/\sqrt{\lambda_t} = \mathcal{O}(\sqrt{d \log(t/\delta)}) , \end{aligned}$$

and therefore we obtain that  $\forall \theta \in \mathcal{C}_t^c(\delta)$ :

$$\|\theta - \theta_\star\|_{\mathbf{H}_t(\theta_\star)} = \mathcal{O}\left(\sqrt{d \log(t/\delta)}\right) .$$

■

### 2.D.1 Proof of [Lemma 2.D.1](#)

**Lemma 2.D.1.** *Let  $\delta \in (0, 1]$ . For all  $\theta \in \mathcal{C}_t(\delta)$ :*

$$\left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{G}_t^{-1}(\theta, \hat{\theta}_t)} \leq \frac{\gamma_t^2(\delta)}{\sqrt{\lambda_t}} + \gamma_t(\delta) .$$

*Proof.* Note that thanks to [Eq. \(1.26\)](#) we have:

$$\begin{aligned} \mathbf{G}_t(\theta, \hat{\theta}_t) &= \sum_{s=1}^{t-1} \alpha(a_s, \theta, \hat{\theta}_t) a_s a_s^\top + \lambda_t \mathbf{I}_d \\ &\succeq \sum_{s=1}^{t-1} \left(1 + |a_s^\top(\theta - \hat{\theta}_t)|\right)^{-1} \dot{\mu}(a_s^\top \theta) a_s a_s^\top + \lambda_t \mathbf{I}_d && \text{(Lemma 1.B.1)} \\ &\succeq \sum_{s=1}^{t-1} \left(1 + \|a_s\|_{\mathbf{G}_t^{-1}(\theta, \hat{\theta}_t)} \left\| \theta - \hat{\theta}_t \right\|_{\mathbf{G}_t(\theta, \hat{\theta}_t)}\right)^{-1} \dot{\mu}(a_s^\top \theta) a_s a_s^\top + \lambda_t \mathbf{I}_d && \text{(Cauchy-Schwarz)} \\ &\succeq \left(1 + \lambda_t^{-1/2} \left\| \theta - \hat{\theta}_t \right\|_{\mathbf{G}_t(\theta, \hat{\theta}_t)}\right)^{-1} \sum_{s=1}^{t-1} \dot{\mu}(a_s^\top \theta) a_s a_s^\top + \lambda_t \mathbf{I}_d && (\mathbf{G}_t(\theta, \hat{\theta}_t) \geq \lambda_t \mathbf{I}_d) \\ &\succeq \left(1 + \lambda_t^{-1/2} \left\| \theta - \hat{\theta}_t \right\|_{\mathbf{G}_t(\theta, \hat{\theta}_t)}\right)^{-1} \left( \sum_{s=1}^{t-1} \dot{\mu}(a_s^\top \theta) a_s a_s^\top + \lambda_t \mathbf{I}_d \right) \\ &= \left(1 + \lambda_t^{-1/2} \left\| \theta - \hat{\theta}_t \right\|_{\mathbf{G}_t(\theta, \hat{\theta}_t)}\right)^{-1} \mathbf{H}_t(\theta) \\ &= \left(1 + \lambda_t^{-1/2} \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{G}_t^{-1}(\theta, \hat{\theta}_t)}\right)^{-1} \mathbf{H}_t(\theta) && \text{(Eq. (1.21))} \end{aligned}$$

Using this inequality, we therefore obtain that:

$$\begin{aligned} \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{G}_t^{-1}(\theta, \hat{\theta}_t)}^2 &\leq \left(1 + \lambda_t^{-1/2} \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{G}_t^{-1}(\theta, \hat{\theta}_t)}\right) \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{H}_t^{-1}(\theta)}^2 \\ &\leq \lambda_t^{-1/2} \gamma_t^2(\delta) \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{G}_t^{-1}(\theta, \hat{\theta}_t)} + \gamma_t^2(\delta) && (\theta \in \mathcal{C}_t(\delta)) \end{aligned}$$

Solving this polynomial inequality in  $\|g_t(\theta) - g_t(\hat{\theta}_t)\|_{\mathbf{G}_t^{-1}(\theta, \hat{\theta}_t)}$  (cf. [Proposition B.1](#)) yields :

$$\|g_t(\theta) - g_t(\hat{\theta}_t)\|_{\mathbf{G}_t^{-1}(\theta, \hat{\theta}_t)} \leq \gamma_t(\delta)^2 / \sqrt{\lambda_t} + \gamma_t(\delta)$$

which proves the announced result. ■

## Appendix 2.E Proof of Theorem 2.4.1

**Theorem 2.4.1** (Bernstein-like tail-inequality for weighted self-normalized martingales). *Let  $T$  be a known integer and  $\{\mathcal{F}_t\}_{t=1}^T$  a filtration. Let  $\{a_t\}_{t=1}^T$  be a stochastic process on  $\mathcal{B}_2(0, 1)$  such that  $a_t$  is  $\mathcal{F}_t$ -measurable. Let  $\{\eta_t\}_{t=2}^T$  be a martingale difference sequence such that  $\eta_{t+1}$  is  $\mathcal{F}_{t+1}$ -measurable. Let  $\{w_t\}_{t=1}^T$  a sequence of weights satisfying [Assumption 2.4.1](#) and  $\{\lambda_t\}_{t=1}^T$  be a deterministic, strictly positive sequence of regularization terms. Furthermore, assume that conditionally on  $\mathcal{F}_t$  we have  $|\eta_{t+1}| \leq \sigma$  a.s. and denote  $v_t^2 = \mathbb{E}[\eta_{t+1}^2 | \mathcal{F}_t]$ . For all  $t \in [T]$  define:*

$$\tilde{\mathbf{H}}_t = \sum_{s=1}^{t-1} w_s^2 v_s^2 a_s a_s^\top + \lambda_t \mathbf{I}_d \quad \text{and} \quad S_t = \sum_{s=1}^{t-1} w_s \eta_{s+1} a_s .$$

Then for any  $\delta \in (0, 1]$  and any fixed  $t \in [T]$ :

$$\mathbb{P} \left( \|S_t\|_{\tilde{\mathbf{H}}_t^{-1}} \leq \frac{\sqrt{\lambda_t}}{2\sigma w_{t-1}} + \frac{2\sigma w_{t-1}}{\sqrt{\lambda_t}} \log \left( \frac{2^d \det(\tilde{\mathbf{H}}_t)^{1/2}}{\delta \lambda_t^{d/2}} \right) \right) \geq 1 - \delta .$$

*Proof.* Let  $t$  be a fixed round and let  $M_s^t(\xi)$  for  $\xi \in \mathbb{R}^d$  and  $1 \leq s \leq t$  be defined as

$$M_s^t(\xi) = \exp \left( \frac{1}{\sigma w_{t-1}} \xi^\top S_s - \frac{1}{\sigma^2 w_{t-1}^2} \xi^\top \bar{\mathbf{H}}_s \xi \right) ,$$

with  $S_s = \sum_{u=1}^{s-1} w_u \eta_{u+1} a_u$  and  $\bar{\mathbf{H}}_s = \sum_{u=1}^{s-1} w_u^2 v_{u+1}^2 a_u a_u^\top$  where  $v_u^2 = \mathbb{E}[\eta_u^2 | \mathcal{F}_s]$ . When  $s = t$ , we will use the notation  $M_t$  for  $M_t^t$ . A direct application of [Lemma 2.2.1](#) yields that for all  $\xi \in \mathcal{B}_d(0, 1)$  and  $2 \leq s \leq t$  and under [Assumption 2.4.1](#);

$$\mathbb{E} \left[ M_s^t(\xi) | \mathcal{F}_{s-1} \right] \leq M_{s-1}^t(\xi) \quad \text{a.s.}$$

Hence, for all  $1 \leq s \leq t$  and  $\xi \in \mathcal{B}_d(0, 1)$  we have  $\mathbb{E}[M_t(\xi)] \leq \mathbb{E}[M_s^t(\xi)] \leq \mathbb{E}[M_1^t(\xi)] \leq 1$ . Further for  $1 \leq s \leq t$  define

$$\bar{M}_s^t = \int_{\xi} M_s^t(\xi) dh_s(\xi) ,$$

where  $h_s$  is the density of an isotropic normal distribution of precision  $\frac{2\lambda_s}{\sigma^2 w_{t-1}^2}$  truncated on  $\mathcal{B}_d(0, 1)$ . Following a similar reasoning as in the proof of [Corollary 2.2.2](#) one easily obtains that  $\mathbb{E}[\bar{M}_t^s] \leq 1$  and in particular for  $s = t$ . From there applying the maximal inequality and following the proof of [Theorem 2.2.1](#) to compute  $\bar{M}_t$  yields the announced result. ■

# Locality-Sensitive Algorithms for GLBs

---

In this chapter we apply our new confidence set from Chapter 2 to the design of improved self-concordant GLB algorithms. We introduce two algorithms which both rely on this enhanced confidence set but differ in how they ensure optimism (either through exploration bonus or parameter-search). For both algorithms we prove regret upper-bounds that tell a much more nuanced story about the effects of non-linearity. Such effects are indeed deferred to a second-order term of the regret, tied to a *transitory* regime during which the algorithms search for highly rewarding areas of the action set. The regret suffered during this phase is still negatively impacted by the non-linearity but becomes *negligible* for large horizons as the algorithms enter a *permanent* regime. Non-linearity then no longer plays a role; only the reward sensitivity around the optimal action does. In addition to such a contrasting conclusion, our algorithms display a clear improvement over previous approaches as they enjoy regret upper-bounds that are *exponentially smaller* w.r.t problem-dependent constants. The end of the chapter is dedicated to the Logistic Bandit setting, for which we identify configurations where non-linearity does not impact the transitory phase. This ultimately removes its detrimental effects from the regret bounds, even for short horizons. Finally, we derive a problem-dependent lower bound for the Logistic Bandit, proving that in the permanent regime our algorithm are *minimax-optimal* w.r.t the dimension  $d$ , the horizon  $T$  and the constant  $\kappa_\mu$  that embodies the effects of non-linearity. We conclude this chapter with some numerical experiments illustrating our theoretical findings.

## Outline

---

3.1	Tighter and local exploration bonuses . . . . .	66
3.1.1	Algorithm and regret upper-bound . . . . .	66
3.1.2	Discussion . . . . .	67
3.1.3	Proof of the regret bound . . . . .	68
3.2	The parameter-search alternative . . . . .	71
3.2.1	Algorithm and regret upper-bound . . . . .	71
3.2.2	Discussion . . . . .	72
3.2.3	Proof of the regret bound . . . . .	72
3.2.4	A tractable algorithm: convex relaxation . . . . .	73
3.3	Non-linearity and transitory regret . . . . .	74
3.3.1	Transitory regret and detrimental arms . . . . .	75
3.3.2	Non-linearity in LogB: a blessing? . . . . .	77
3.4	Optimality of the permanent regret . . . . .	79

3.4.1	Regret lower-bound . . . . .	79
3.4.2	Proof of the lower-bound . . . . .	80
3.5	Numerical simulations . . . . .	84

In this chapter we only consider self-concordant GLBs - *i.e* for which [Assumption 1.4.1](#) holds.

### 3.1 Tighter and local exploration bonuses

In this section we present GLM-UCB+, an improved algorithm for self-concordant GLBs. It mimics the approach of GLM-UCB and resorts to exploration bonuses to enforce optimism and drive exploration. It leverages our improved confidence set from [Theorem 2.3.1](#):

$$\mathcal{C}_t(\delta) = \left\{ \theta \in \Theta, \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{H}_t^{-1}(\theta)} \leq \gamma_t(\delta) \right\},$$

to design smaller exploration bonuses. It therefore enjoys improved theoretical guarantees with finer dependencies w.r.t to the problem-dependent constants characterizing the level of non-linearity. Recall that to enjoy a reduced scaling of  $\gamma_t(\delta)$  (w.r.t  $d$  and  $t$ ) we set the regularization parameters to  $\lambda_t = d \log(2 + t)$  for  $t \geq 1$ .

#### 3.1.1 Algorithm and regret upper-bound

**Algorithm.** Similarly to GLM-UCB, the algorithm GLM-UCB+ relies on a “projected” version of the MLE estimator  $\hat{\theta}_t$ . We define the set  $\mathcal{W}_t$  of *information-preserving* parameters:

$$\mathcal{W}_t = \left\{ \theta \in \Theta, \dot{\mu}(a_s^\top \theta) \geq \min_{\theta' \in \mathcal{C}_s(\delta)} \dot{\mu}(a_s^\top \theta') \text{ for all } s \in [t] \right\}, \quad (3.1)$$

and project  $\hat{\theta}_t$  back to this set with an map preserving its learning guarantees. Formally:

$$\tilde{\theta}_t \in \arg \min_{\theta \in \mathcal{W}_t} \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{H}_t^{-1}(\theta)}. \quad (3.2)$$

This program is well-defined by the continuity of the objective and the non-emptiness of  $\mathcal{W}_t$ . Define the exploration bonus:

$$\varepsilon_t(a) := (2 + 4S) \dot{\mu}(a^\top \tilde{\theta}_t) \gamma_t(\delta) \|a\|_{\mathbf{H}_t^{-1}(\tilde{\theta}_t)} + (4 + 8S) \bar{\kappa}_\mu \gamma_t^2(\delta) \|a\|_{\mathbf{V}_t^{-1}}, \quad (3.3)$$

which unlike the exploration function of GLM-UCB does depend on the precise location of the estimator  $\tilde{\theta}_t$ . GLM-UCB+ prescribes the following strategy:

$$\text{play } a_t \in \arg \max \mu(a^\top \tilde{\theta}_t) + \varepsilon_t(a). \quad (3.4)$$

The pseudo-code is provided in [Algorithm 4](#).

**Theoretical guarantees.** The following theorem (informally stated, see [Section 3.1.3](#) for the formal statement) ensures that GLM-UCB+ enjoys a sub-linear regret with improved problem-dependent dependencies.

**Theorem 3.1.1** (Regret of GLM-UCB+, informal). *The regret of GLM-UCB+ satisfies:*

$$\text{Regret}_{\theta_\star}(T) = \tilde{\mathcal{O}} \left( \sqrt{\dot{\mu}(a_\star^\top \theta_\star) T} + \bar{\kappa}_\mu \right) \text{ with high probability.}$$

**Algorithm 4** GLM-UCB+

**input:** Arm set  $\mathcal{A}$ , regularization coefficients  $\{\lambda_t\}_t$ , failure level  $\delta$ , admissible parameter set  $\Theta$ .

---

Compute the reward sensitivity constants  $\bar{\ell}_\mu$ ,  $\bar{L}_\mu$  and  $\bar{\kappa}_\mu$ .  $\triangleright$  *initialization*  
 Set  $\mathbf{V}_1 \leftarrow (\lambda_1/\bar{\ell}_\mu)\mathbf{I}_d$ ,  $\mathbf{H}_1 \leftarrow \lambda_1\mathbf{I}_d$ ,  $\hat{\theta}_1 \leftarrow 0_d$  and  $\tilde{\theta}_1 \leftarrow 0_d$ .  
**for**  $t \in [1, T]$  **do**  
   Compute the exploration bonuses  $\{\varepsilon_t(a)\}_{a \in \mathcal{A}}$  according to Eq. (3.3).  
   Play the arm  $a_t$  according to Eq. (3.4).  $\triangleright$  *planning*  
   Observe reward  $r_{t+1}$ .  
   Update the estimator  $\hat{\theta}_{t+1}$  and design matrix  $\mathbf{V}_{t+1}$ .  $\triangleright$  *learning*  
   Compute  $\tilde{\theta}_{t+1}$  according to Eq. (3.2) and compute  $\mathbf{H}_{t+1} \leftarrow \mathbf{H}_{t+1}(\tilde{\theta}_{t+1})$ .  
**end for**

---

The regret therefore decomposes in two terms; **(1)** a first-order term which dominates when the horizon  $T$  is large and **(2)** a second order term which dominates at the beginning of the experiment. The first-order term still scales as  $\sqrt{T}$  but shows refined problem-dependent dependencies as it involves the slope of the *effective* reward signal at the optimal action  $a_*$ . This therefore comes as a dramatic improvement over previous works which regret scale with  $\bar{\kappa}_\mu \gg \dot{\mu}(a_*^\top \theta_*)$ . This is particularly well-illustrated in the LogB setting as demonstrated below. The dependency in  $\bar{\kappa}_\mu$  is deferred to the second order-term and therefore is only additive, not multiplicative.

 Consider the Logistic Bandit problem where  $\mathcal{A} = \mathcal{B}_d(0, 1)$ . In this configuration we have  $\dot{\mu}(a_*^\top \theta_*) = \ell_\mu \approx \exp(-\|\theta_*\|)$ ; the regret of GLM-UCB+ therefore satisfies for  $T$  large enough:

$$\text{Regret}_{\theta_*} = \tilde{O}\left(\exp(-\|\theta_*\|/2)\sqrt{T}\right) \quad \text{w.h.p.}$$

In other words GLM-UCB+ brings an *exponential* acceleration over previous work which regret bounds in the same problem scale as  $\tilde{O}(\exp(S)\sqrt{T})$  where  $S \geq \|\theta_*\|$ .

**Remark 3.1.1** (Another illustration: the Poisson Bandit case). *The conclusion remains the same for the Poisson bandit; the regret of GLM-UCB scales with  $\exp(2S)$  while the regret of GLM-UCB+ scales with  $\exp(\|\theta_*\|/2)$  for  $T$  large enough.*

### 3.1.2 Discussion

**Impact of non-linearity.** The regret bound of [Theorem 3.1.1](#) tells a much more nuanced story about the impact of non-linearity. It indeed proves that in the long-term regime the non-linearity of the reward signal does not impact the performance. What matters is only the sensitivity of the reward function around the best action  $a_*$ . Intuitively this makes a lot of sense; any algorithm that enjoys a sub-linear regret (w.r.t the horizon  $T$ ) must eventually play most of its actions “close” to  $a_*$ . In this regime the observed reward signal therefore behaves like a linear bandit with slope  $\dot{\mu}(a_*^\top \theta_*)$ ; if this quantity is small (resp. large) then the reward function is flat (resp. peaked) and playing sub-optimal arms results in a small (resp. large) instantaneous regret. The effects of non-linearity mostly show up in the second-order term which still scales with  $\bar{\kappa}_\mu$ . This suggests that in a highly non-linear environment, discovering (approximately)  $a_*$  is tedious and reaching the long-term regime requires a long *burn-in* phase.

**Information-preserving projection.** A singular feature of GLM-UCB+ is the involvement of the set of information preserving estimators  $\mathcal{W}_t$ . It is tightly linked to the form of the

exploration bonus  $\varepsilon_t(\cdot)$ . Note that the first term of the latter (cf. Eq. (3.3)) directly depends on the estimator  $\tilde{\theta}_t$ . As we mentioned before, the exploration bonus for each action  $a \in \mathcal{A}$  has to smoothly vanish if one expects to enjoy sub-linear regret bounds. Indeed as information about  $a$  is gathered, the uncertainty about its reward diminishes and exploration “around” this arm should vanish. Now, the first term of the exploration bonus from Eq. (3.3) evolves with  $\dot{\mu}(a^\top \tilde{\theta}_t) a^\top \mathbf{H}_t^{-1}(\tilde{\theta}_t) a$ . If  $\tilde{\theta}_t$  was unconstrained this quantity could *increase*; updating  $\tilde{\theta}_t$  could decrease some eigenvalues of  $\mathbf{H}_t(\tilde{\theta}_t)$ , “destroying” information that was previously gathered in the directions of the corresponding eigenvectors. Projecting  $\tilde{\theta}_t$  on  $\mathcal{W}_t$  effectively avoids such behavior as it allows to lower-bound (in the Loewner order sense)  $\mathbf{H}_t(\tilde{\theta}_t)$  by a matrix which eigenvalues are strictly increasing over time; in other words, it preserves information acquired in the past.

$$\tilde{\theta}_t \in \mathcal{W}_t \implies \mathbf{H}_t(\tilde{\theta}_t) \succeq \mathbf{L}_t := \sum_{s=1}^{t-1} \dot{\mu}(a_s^\top \bar{\theta}_s) a_s a_s^\top + \lambda_t \mathbf{I}_d,$$

where  $\bar{\theta}_s := \arg \min_{\theta \in \mathcal{C}_s} \dot{\mu}(a_s^\top \theta)$ . This guarantees diminishing bonuses or equivalently that the bonuses cumulates at a correct rate. This intuitive argument is formalized in Lemma 3.1.2.

**Remark 3.1.2** (An alternative formulation for  $\mathcal{W}_t$ ). *The expression for the information-preserving set from Eq. (3.1) is rather intuitive but complex to handle in a practical experiment. For any inverse link function whose derivative  $\dot{\mu}$  achieves its extremum in its tails (such as in the Logistic and Poisson case) it can be trivially simplified. For instance in the LogB case;*

$$\mathcal{W}_t = \left\{ \theta \in \Theta, |a_s^\top \theta| \leq \max_{\theta' \in \mathcal{C}_s} |a_s^\top \theta'| \text{ for all } s \in [t] \right\}.$$

In this case  $\mathcal{W}_t$  therefore appears as a convex set made up of  $2 \min(|\mathcal{A}|, t-1)$  linear constraints (plus the convex constraint  $\theta \in \Theta$ ).

**Tractability.** From a practical standpoint and despite  $\mathcal{W}_t$  being convex in practical cases of interest GLM-UCB+ suffers an important drawback because of the projection step of Eq. (3.2). It indeed requires running a non-convex optimization routine at every round since:

$$\theta \mapsto \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{H}_t^{-1}(\theta)},$$

is non-convex. This drawback is shared with the approach of Filippi et al. (2010) (their projection map is also non-convex) but its impact is more burdensome here. Indeed GLM-UCB can discard the projection step whenever  $\hat{\theta}_t \in \Theta$  which, in practice, happens at almost every round. For GLM-UCB+ the projection step can only be discarded when  $\hat{\theta}_t \in \mathcal{W}_t$  which is less frequent (practically this requires  $\hat{\theta}_t \in \mathcal{C}_s(\delta)$  for every  $s \in [t]$ ). The projection step will be removed in the following section where we design a parameter-search alternative to GLM-UCB+.

### 3.1.3 Proof of the regret bound

Below we state a more formal version of Theorem 3.1.1.

**Theorem 3.1.1** (Regret of GLM-UCB+, formal). *Let  $\delta \in (0, 1]$ . With probability at least  $1 - \delta$  the regret of GLM-UCB+ satisfies:*

$$\text{Regret}_{\theta_*}(T) = \mathcal{O} \left( d \log(T/\delta) \sqrt{T \dot{\mu}(a_*^\top \theta_*)} + \bar{\kappa}_\mu d^2 \log(T/\delta)^2 \right).$$

Notice that the scaling of the first-order term w.r.t the dimension  $T$  and  $d$  still match with the  $d \log(T) \sqrt{T}$  rate obtained in the linear case. We now proceed with the proof.

### Regret and prediction error.

The proof relies on the following result tying the regret to the sum of bonuses on the trajectory  $\{a_t\}_{t=1}^T$  when the bonus function upper-bounds the prediction error. This result is classical and the proof deferred to [Appendix 3.A](#).

**Proposition 3.1.1** (Regret and exploration bonus). *Recall the prediction error  $\Delta_t(a) = |\mu(a^\top \tilde{\theta}_t) - \mu(a^\top \theta_\star)|$ . If for all  $a \in \mathcal{A}$  and  $t \in [T]$  we have  $\varepsilon_t(a) \geq \Delta_t(a)$  then:*

$$\text{Regret}_{\theta_\star}(T) \leq 2 \sum_{t=1}^T \varepsilon_t(a_t).$$

The goal is therefore to design *tight* upper-bounds on the prediction error. This is achieved by leveraging our improved confidence set which yields the bonus function of [Eq. \(3.3\)](#). To keep the proof concise we defer the demonstration of this result to [Appendix 3.B](#).

**Lemma 3.1.1** (Confident prediction-error upper-bound). *Under the event  $\{\theta_\star \in \mathcal{C}_t(\delta), \forall t \geq 1\}$  for any  $a \in \mathcal{A}$  and  $t \geq 1$ :*

$$\Delta_t(a) \leq 2(1 + 2S)\dot{\mu}(a^\top \tilde{\theta}_t) \|a\|_{\mathbf{H}_t^{-1}(\tilde{\theta}_t)} \gamma_t(\delta) + 2(1 + 2S)^2 \bar{\kappa}_\mu \gamma_t^2(\delta) \|a\|_{\mathbf{V}_t^{-1}}^2 = \varepsilon_t(a).$$

Combining [Proposition 3.A.1](#) and [Lemma 3.1.1](#) yields that under the event  $\{\theta_\star \in \mathcal{C}_t(\delta), \forall t \geq 1\}$  the regret decomposes in two terms:

$$\text{Regret}_{\theta_\star}(T) \leq 4(1 + 2S)\bar{\gamma}_T(\delta) \underbrace{\sum_{t=1}^T \dot{\mu}(a_t^\top \tilde{\theta}_t) \|a_t\|_{\mathbf{H}_t^{-1}(\tilde{\theta}_t)}}_{R_1(T)} + 4(1 + 2S)^2 \bar{\kappa}_\mu \bar{\gamma}_T(\delta)^2 \underbrace{\sum_{t=1}^T \|a_t\|_{\mathbf{V}_t^{-1}}^2}_{R_2(T)}, \quad (3.5)$$

where  $\bar{\gamma}_T(\delta) = \max_{t \in [T]} \gamma_t(\delta)$ .

### Bounding $R_2(T)$ .

The second term is easily bounded thanks to the Elliptical Potential lemma (see [Lemma B.3](#)):

$$R_2(T) \leq 2d \log \left( \lambda_T + T\bar{\ell}_\mu/d \right). \quad (3.6)$$

The second term in [Eq. \(3.5\)](#) therefore scales as  $\log^2(T)$  and is a *dominated* term in the final regret bound. It shows a multiplicative dependency in  $\bar{\kappa}_\mu$  because we used a uniform bound on  $\dot{\mu}$  (which itself is dominated by  $\dot{\mu}$ ) over  $\mathcal{A} \times \Theta$ . This however allowed for  $R_1(T)$  (which will turn out to be the dominating term of the regret) to depend only on the estimator  $\tilde{\theta}_t$ .

### Bounding $R_1(T)$ .

The last step of the proof requires bounding  $R_1(T)$ . By simple manipulations one can arrive to the following inequality (the proof can be found in [Appendix 3.C](#)) where  $\bar{\theta}_t := \arg \min_{\theta \in \mathcal{C}_t} \dot{\mu}(a_t^\top \theta)$ :

$$R_1(T) \leq \sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t) \|a_t\|_{\mathbf{H}_t^{-1}(\bar{\theta}_t)} + 2\sqrt{1 + 2S\bar{\kappa}_\mu} \bar{\gamma}_T(\delta) R_2(T). \quad (3.7)$$

Let us focus on the first term. Because for all  $t \geq 1$ ,  $\tilde{\theta}_t \in \mathcal{W}_t$  we have that by definition  $\dot{\mu}(a_t^\top \tilde{\theta}_t) \geq \dot{\mu}(a_t^\top \bar{\theta}_t)$ ; this yields the following matrix lower bound on  $\mathbf{H}_t(\tilde{\theta}_t)$ :

$$\mathbf{H}_t(\tilde{\theta}_t) \succeq \mathbf{L}_t := \sum_{s=1}^{t-1} \dot{\mu}(a_s^\top \bar{\theta}_s) a_s a_s^\top + \lambda_t \mathbf{I}_d = \sum_{s=1}^{t-1} \bar{a}_s \bar{a}_s^\top + \lambda_t \mathbf{I}_d,$$

where for all  $s \in [T]$  we defined  $\bar{a}_s = \sqrt{\dot{\mu}(a_s^\top \bar{\theta}_s)} a_s$ . Therefore:

$$\sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t) \|a_t\|_{\mathbf{H}_t^{-1}(\bar{\theta}_t)} \leq \sum_{t=1}^T \sqrt{\dot{\mu}(a_t^\top \bar{\theta}_t)} \|\bar{a}_t\|_{\mathbf{L}_t^{-1}}. \quad (3.8)$$

We are almost done but are faced with a slight technical difficulty. Indeed after applying the Cauchy-Schwarz inequality we obtain:

$$\sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t) \|a_t\|_{\mathbf{H}_t^{-1}(\bar{\theta}_t)} \leq \sqrt{\sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t)} \sqrt{\sum_{t=1}^T \|\bar{a}_t\|_{\mathbf{L}_t^{-1}}^2}.$$

It is tempting to bound the most r.h.s term by a direct application of the Elliptical Potential lemma; however  $\|\bar{a}_t\|^2$  is bounded at worst by  $\bar{L}_\mu$  which with a naive application of the Elliptical Potential will appear multiplicatively in the bound on this first-order term. This is innocent for GLBs with bounded Lipschitz constant (*e.g* the LogB for which  $\bar{L}_\mu \leq 1/4$ ) but it is extremely unpleasant for some other models (*e.g* the Poisson Bandit for which  $\bar{L}_\mu \approx \exp(S)$ ) as this dependency will appear in the regret bound's leading term (which is exactly what we are trying to avoid). Fortunately this is merely an analysis issue, easily circumvented by an adapted decomposition which again defers this annoying dependency to a small additive term. We state this refined result in the lemma below, which proof is deferred to [Appendix 3.D](#).

**Lemma 3.1.2.** *The following holds:*

$$\sum_{t=1}^T \sqrt{\dot{\mu}(a_t^\top \bar{\theta}_t)} \|\bar{a}_t\|_{\mathbf{L}_t^{-1}} \leq \sqrt{2d \log(\lambda_T + T/d)} \sqrt{\sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t) + 2d \bar{L}_\mu^2 \log(\lambda_T + \bar{L}_\mu T/d)}.$$

To simplify matter and reduce clutter, in what follows we ignore the additional term scaling with  $\bar{L}_\mu^2$  (the detail-oriented reader can check that this term is even a ‘‘third-order’’ term as it is dominated by the final regret bound's second order term). We will therefore use:

$$\sum_{t=1}^T \sqrt{\dot{\mu}(a_t^\top \bar{\theta}_t)} \|\bar{a}_t\|_{\mathbf{L}_t^{-1}} \leq \sqrt{2d \log(\lambda_T + T/d)} \sqrt{\sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t)}. \quad (3.9)$$

To finish bounding  $R_1(T)$  we only have left to bound  $\sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t)$ . Note that if  $\theta_\star \in \mathcal{C}_t(\delta)$  by definition of  $\bar{\theta}_t$  we have  $\dot{\mu}(a_t^\top \bar{\theta}_t) \leq \dot{\mu}(a_t^\top \theta_\star)$ . Therefore under the event  $\{\theta_\star \in \mathcal{C}_t(\delta) \text{ for all } t \geq 1\}$ ;

$$\sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t) \leq \sum_{t=1}^T \dot{\mu}(a_t^\top \theta_\star) \leq T \dot{\mu}(a_\star^\top \theta_\star) + \text{Regret}_{\theta_\star}(T). \quad (3.10)$$

The proof for this last inequality relies on a Taylor-expansion and the self-concordance property; it is deferred to [Appendix 3.E](#). Combining [Eq. \(3.8\)](#), [Eq. \(3.9\)](#) and [Eq. \(3.10\)](#) along with the inequality  $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$  for all  $a, b > 0$  yields;

$$\sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t) \|a_t\|_{\mathbf{H}_t^{-1}(\bar{\theta}_t)} \leq \sqrt{2d \log(\lambda_T + T/d)} \left( \sqrt{T \dot{\mu}(a_\star^\top \theta_\star)} + \sqrt{\text{Regret}_{\theta_\star}(T)} \right).$$

Assembling this with [Eq. \(3.7\)](#) we obtain:

$$R_1(T) \leq \sqrt{2d \log(\lambda_T + \bar{L}_\mu T/d)} \left( \sqrt{T \dot{\mu}(a_\star^\top \theta_\star)} + \sqrt{\text{Regret}_{\theta_\star}(T)} \right) + 2\sqrt{1 + 2S \bar{\kappa}_\mu \bar{\gamma}_T(\delta)} R_2(T). \quad (3.11)$$

**Finishing the bound.**

Assembling the regret decomposition of Eq. (3.5) along with the bounds on  $R_1(T)$  and  $R_2(T)$  (respectively from Eq. (3.11) and Eq. (3.6)) along with straight-forward upper-bounding yields:

$$\text{Regret}_{\theta_\star}(T) \leq 4\sqrt{2}f(T) \left( \sqrt{T\dot{\mu}(a_\star^\top\theta_\star)} + \sqrt{\text{Regret}_{\theta_\star}(T)} \right) + 24\bar{\kappa}_\mu f(T)^2,$$

where we defined  $f(T) := (1 + 2S)\bar{\gamma}_T(\delta)\sqrt{d\log(\lambda_T + \bar{L}_\mu T/d)}$ . This is a second-order polynomial inequation with unknown  $\sqrt{\text{Regret}_{\theta_\star}(T)}$ ; solving it (see Proposition B.1) and using  $(a + b)^2 \leq 2a^2 + 2b^2$  to simplify yields;

$$\text{Regret}_{\theta_\star}(T) \leq 8\sqrt{2}f(T)\sqrt{T\dot{\mu}(a_\star^\top\theta_\star)} + 64f(T)^2 + 48\bar{\kappa}_\mu f(T)^2.$$

The claimed scaling is easily obtained by recalling that  $\gamma_t(\delta) = \mathcal{O}(\sqrt{d\log(t/\delta)})$  which implies that  $f(T) = \mathcal{O}(d\log(T/\delta))$ . Recall that this proof checks when the event  $\{\theta_\star \in \mathcal{C}_t(\delta) \text{ for all } t \geq 1\}$  which holds with probability at least  $1 - \delta$ .

**Remark 3.1.3** (The contextual case). *This analysis shows that the story is similar when arm-sets are time-varying (e.g in the contextual bandit setting). Straight-forward manipulations shows that the term  $\sqrt{T\dot{\mu}(a_\star^\top\theta_\star)}$  is then replaced by  $\sqrt{T}\sqrt{(1/T)\sum_{t=1}^T\dot{\mu}(a_{\star,t}^\top\theta_\star)}$  where  $a_{\star,t}$  is the best action at round  $t$ . In other words the local slope around the best action is replaced by the “averaged” local slope.*

## 3.2 The parameter-search alternative

In the previous section we introduced GLM-UCB+, which leverages the enhanced confidence set  $\mathcal{C}_t(\delta)$  but mimics the original GLM-UCB algorithm by resorting to exploration-bonuses to enforce optimism. We now study its parameter-search counterpart which directly finds an optimistic parameter in  $\mathcal{C}_t(\delta)$  and plays greedily w.r.t to this parameter. We coin this algorithm OFU-GLB for its resemblance to the OFUL algorithm of Abbasi-Yadkori et al. (2011). Note that in the linear case, the exploration-bonus and parameter-search approaches are *strictly* equivalent; as illustrated throughout this section this is no longer true with non-linear reward signals.

### 3.2.1 Algorithm and regret upper-bound

**Algorithm.** OFU-GLB plays according to the following simple strategy:

$$\text{play } a_t \in \arg \max_{a \in \mathcal{A}} \max_{\theta \in \mathcal{C}_t(\delta)} a^\top \theta.$$

The pseudo-code is provided in Algorithm 5; note that this algorithm does not requires to compute the reward sensitivity constants but only needs knowledge of  $S$ .

**Theoretical guarantees.** Again we start with an informal statement to ease discussions.

**Theorem 3.2.1** (Regret of OFU-GLB, informal). *The regret of OFU-GLB satisfies:*

$$\text{Regret}_{\theta_\star}(T) = \tilde{\mathcal{O}}(\sqrt{\dot{\mu}(a_\star^\top\theta_\star)T} + \bar{L}_\mu/\ell_\mu).$$

Compared to GLM-UCB+ the regret’s second-order term is reduced; it scales as  $\bar{L}_\mu/\ell_\mu$  - recall that  $\ell_\mu = \max_{a \in \mathcal{A}} \dot{\mu}(a^\top\theta_\star)$  measures the *effective* minimum reward sensitivity (not the worst-case over  $\Theta$ ). This bound already hints at the refined problem-dependent behavior of the parameter search approach. Of course  $\ell_\mu \geq \bar{\ell}_\mu$  so this second order term can be upper-bounded by  $\tilde{\mathcal{O}}(\bar{\kappa}_\mu)$  to retrieve the regret upper-bound of GLM-UCB+.

**Algorithm 5** OFU-GLB

**input:** Arm set  $\mathcal{A}$ , regularization coefficients  $\{\lambda_t\}_t$ , failure level  $\delta$ , admissible parameter set  $\Theta$ .

Set  $\mathbf{H}_1 \leftarrow \lambda_1 \mathbf{I}_d$ ,  $\hat{\theta}_1 \leftarrow 0_d$ .

**for**  $t \in [1, T]$  **do**

Solve  $a_t \in \arg \max_{\mathcal{A}} \max_{\theta \in \mathcal{C}_t(\delta)} a^\top \theta$ .  $\triangleright$  *planning*

Play the arm  $a_t$  and observe reward  $r_{t+1}$ .

Update the estimator  $\hat{\theta}_{t+1}$  and the confidence interval  $\mathcal{C}_t(\delta)$ .  $\triangleright$  *learning*

**end for**

 The above comment is well illustrated in the LogB case. Let  $\mathcal{A} = \mathcal{B}_2(0, 1)$ ; then the second-order term of GLM-UCB+'s regret upper bound scale with  $\exp(S)$  but “only”  $\exp(\|\theta_\star\|)$  for OFU-GLB, which can be much smaller. Note that this is not a by-product of a loose analysis as for GLM-UCB+ this scaling is already hard-coded in the exploration bonus.

### 3.2.2 Discussion

**Impact of non-linearity.** The conclusions brought by this regret bound are essentially the same as for GLM-UCB+; the effect of non-linearity is pushed in a second-order term of the regret, conceptually corresponding to a burn-in phase. In [Section 3.3](#) we will leverage OFU-GLB to study more in details the nature of this second-order term.

**Practical advantages.** OFU-GLB has several practical advantages over GLM-UCB+; it does not require computing the reward sensitivity constants  $\bar{L}_\mu$ ,  $\bar{\ell}_\mu$  and  $\bar{\kappa}_\mu$  and more importantly does not require the expensive maintaining of the information-preserving set  $\mathcal{W}_t$ . We argued in [Section 3.1.2](#) that with an exploration-bonus approach projecting on such a set is necessary. This is no longer the case with a parameter-search approach which allows to quantify the acquired information through the metric  $\mathbf{H}_t(\theta_\star)$  which is naturally increasing with  $t$ . There is still an important drawback to OFU-GLB as the constraint  $\theta \in \mathcal{C}_t(\delta)$  in the planning objective is not convex. This will be circumvented with minor impact on the regret bound in [Section 3.2.4](#).

### 3.2.3 Proof of the regret bound

**Theorem 3.2.2** (Regret of OFU-GLB, formal). *Let  $\delta \in (0, 1]$ . With probability at least  $1 - \delta$  the regret of OFU-GLB satisfies:*

$$\text{Regret}_{\theta_\star}(T) = \mathcal{O} \left( d \log(T/\delta) \sqrt{T \dot{\mu}(a_\star^\top \theta_\star)} + (\bar{L}_\mu / \bar{\ell}_\mu) d^2 \log(T/\delta)^2 \right) .$$

The demonstration is naturally similar to the proof of [Theorem 3.1.1](#) laid out in [Section 3.1.3](#); it is however much neater. Throughout this whole proof we work under the event  $\{\theta_\star \in \mathcal{C}_t(\delta) \text{ for all } t \geq 1\}$  which holds with probability at least  $1 - \delta$ . Define the optimistic parameter  $\theta_t$  as follows:

$$(a_t, \theta_t) := \arg \max_{a \in \mathcal{A}, \theta \in \mathcal{C}_t(\delta)} a^\top \theta .$$

Since  $\theta_\star \in \mathcal{A}$  we therefore have by optimism that  $a_\star^\top \theta_\star \leq a_t^\top \theta_t$ . Using this along with a second

order Taylor expansion yields the following set of inequalities:

$$\begin{aligned}
\text{Regret}_{\theta_*}(T) &\leq \sum_{t=1}^T \mu(a_t^\top \theta_t) - \mu(a_t^\top \theta_*) && (\mu \nearrow) \\
&\leq \sum_{t=1}^T \dot{\mu}(a_t^\top \theta_*) a_t^\top (\theta_t - \theta_*) + \bar{L}_\mu [a_t^\top (\theta_t - \theta_*)]^2 / 2 && (|\dot{\mu}| \leq \dot{\mu} \leq \bar{L}_\mu) \\
&\leq \sum_{t=1}^T \dot{\mu}(a_t^\top \theta_*) \|a_t\|_{\mathbf{H}_t^{-1}(\theta_*)} \|\theta_t - \theta_*\|_{\mathbf{H}_t(\theta_*)} && (\text{Cauchy-Schwarz}) \\
&\quad + (\bar{L}_\mu / 2) \sum_{t=1}^T \|a_t\|_{\mathbf{H}_t^{-1}(\theta_*)}^2 \|\theta_t - \theta_*\|_{\mathbf{H}_t(\theta_*)}^2
\end{aligned}$$

Easy computations involving the self-concordance property yield the following deviation bound (for completeness the proof is deferred to [Appendix 3.F](#)):

$$\|\theta_t - \theta_*\|_{\mathbf{H}_t(\theta_*)} \leq 2(1 + 2S)\gamma_t(\delta). \quad (3.12)$$

Plugging this in the above regret bound yields;

$$\text{Regret}_{\theta_*}(T) \leq 2(1 + 2S)\bar{\gamma}_T(\delta) \sum_{t=1}^T \dot{\mu}(a_t^\top \theta_*) \|a_t\|_{\mathbf{H}_t^{-1}(\theta_*)} + 2(1 + 2S)^2 \bar{L}_\mu \bar{\gamma}_T(\delta)^2 \sum_{t=1}^T \|a_t\|_{\mathbf{H}_t^{-1}(\theta_*)}^2.$$

The second term is easily bounded by using  $\mathbf{H}_t(\theta_*) \succeq \ell_\mu \mathbf{V}_t$  and the Elliptical Lemma. The first-order term can be bounded thanks to [Lemma 3.1.2](#) by defining this time  $\bar{a}_t = \sqrt{\mu(a_t^\top \theta_*)} a_t$  which yields (again, we will omit the third-order term in the remaining of this proof to reduce clutter):

$$\sum_{t=1}^T \dot{\mu}(a_t^\top \theta_*) \|a_t\|_{\mathbf{H}_t^{-1}(\theta_*)} \leq \sqrt{2d \log(\lambda_T + T/d)} \sqrt{\sum_{t=1}^T \dot{\mu}(a_t^\top \theta_*)}.$$

The remaining term is bounded as in the proof of [Theorem 3.1.1](#) by showing:

$$\sum_{t=1}^T \dot{\mu}(a_t^\top \theta_*) \leq T \dot{\mu}(a_*^\top \theta_*) + \text{Regret}_{\theta_*}(T).$$

This yields the following second-order polynomial inequation involving the regret itself;

$$\text{Regret}_{\theta_*}(T) \leq 2\sqrt{2}f(T)(\sqrt{T\mu(a_*^\top \theta_*)} + \sqrt{\text{Regret}_{\theta_*}(T)}) + 4(\bar{L}_\mu/\ell_\mu)f(T)^2.$$

Solving the above and using  $f(T) = \mathcal{O}(d \log(T/\delta))$  leads to the announced result.

### 3.2.4 A tractable algorithm: convex relaxation

We finish this section on OFU-GLB by introducing a tractable alternative in the finite arm-set case. As mentioned earlier the planning phase of OFU-GLB involves solving for some given  $a$ :

$$\max_{\theta \in \mathcal{C}_i(\delta)} a^\top \theta.$$

**Algorithm 6** OFU-GLB-r

---

**input:** Finite arm-set  $\mathcal{A}$ , regularization coefficients  $\{\lambda_t\}_t$ , failure level  $\delta$ , admissible parameter set  $\Theta$ .  
Set  $\mathbf{H}_1 \leftarrow \lambda_1 \mathbf{I}_d$ ,  $\hat{\theta}_1 \leftarrow 0_d$ .  
**for**  $t \in [1, T]$  **do**  
  **for**  $a \in \mathcal{A}$  **do**  
    Solve  $\theta_a \leftarrow \arg \max_{\theta \in \mathcal{C}_t^c(\delta)} a^\top \theta$ .  
  **end for**  
  Play  $a_t \in \arg \max_{\mathcal{A}} a^\top \theta_a$ .  $\triangleright$  *planning*  
  Play the arm  $a_t$  and observe reward  $r_{t+1}$ .  
  Update the estimator  $\hat{\theta}_{t+1}$ .  $\triangleright$  *learning*  
**end for**

---

Since  $\mathcal{C}_t(\delta)$  is in all generality non-convex there exist no standard approach for provably approximately solving this program. This is easily circumvented by replacing  $\mathcal{C}_t(\delta)$  by its convex relaxation  $\mathcal{C}_t^c(\delta)$  provided in [Corollary 2.3.2](#):

$$\mathcal{C}_t^c(\delta) = \left\{ \theta \in \Theta, \mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t) \leq \beta_t(\delta) = \gamma_t(\delta) + \gamma_t(\delta)^2 / \sqrt{\lambda_t} \right\}.$$

The planning of the resulting algorithm which we coin **OFU-GLB-r** writes:

$$\text{play } a_t \in \arg \max_{a \in \mathcal{A}} \max_{\theta \in \mathcal{C}_t^c(\delta)} a^\top \theta.$$

This can be solved efficiently when  $|\mathcal{A}|$  is finite by exhaustive search over  $\mathcal{A}$  since  $\max_{\theta \in \mathcal{C}_t^c(\delta)} a^\top \theta$  is a convex program ( $\theta \in \mathcal{C}_t^c(\delta)$  is a convex constraint) which can be solved to arbitrary precision. The pseudo-code for this algorithm is provided in [Algorithm 6](#). This convex relaxation comes at virtually no cost on the regret bound; indeed thanks to [Corollary 2.3.2](#) we know that **(1)**  $\mathcal{C}_t^c(\delta)$  is an anytime confidence set for  $\theta_*$  at level at least  $1 - \delta$ , which ensures that with high probability **OFU-GLB-r** plays according to an optimistic couple  $(a_t, \theta_t)$ . Furthermore [Corollary 2.3.2](#) provides a direct alternative to [Eq. \(3.12\)](#) since for any  $\theta \in \mathcal{C}_t^c(\delta)$  we have **(2)**:

$$\|\theta - \theta_*\|_{\mathbf{H}_t(\theta_*)} \leq (2 + 2S)\gamma_t(\delta) + 2\sqrt{1 + 2S}\beta_t(\delta) = \mathcal{O}\left(\sqrt{d \log(t)}\right).$$

Those two points are the main building blocks behind the proof of **OFU-GLB**'s regret bound. It can be exactly reproduced for bounding the regret of **OFU-GLB-r**, which enjoys a similar regret bound where  $(1 + 2S)\gamma_t(\delta)$  is replaced by  $(2 + 2S)\gamma_t(\delta) + 2\sqrt{1 + 2S}\beta_t(\delta)$  - which has the same scaling w.r.t to  $d$  and  $\log(T)$ .

### 3.3 Non-linearity and transitory regret in LogB

**Non-linearity and the transitory regime.** The regret analysis of **GLM-UCB+** and **OFU-GLB** bring forward the same conclusions on the effects on non-linearity. In particular they highlight that in a *permanent* regime ( $T$  large enough) the non-linearity no longer plays a role in the exploration-exploitation trade-off; what matters is only the local sensitivity of the reward function around the best arm. This long-term regime conceptually corresponds to the phase the algorithm enters once the optimal action is approximately identified. The non-linearity affects a dominated term in the regret, scaling linearly with the reward sensitivity ratio  $\bar{\kappa}_\mu$ . It is appealing to think of this regret term as tied to a *transitory* regime during which the algorithm

performs a somewhat uniform exploration with the goal of discovering the most rewarding area of the arm set. The first goal of this section is to formalize this intuitive characterization in the LogB case. The presence of  $\bar{\kappa}_\mu$  in the regret's second-order term implies that the more non-linear the reward signal the longer this takes and the harder approximately locating  $a_\star$  is. The second goal of this section is to show that in the LogB case this conclusion is extremely worst-case and that for many instances the regret incurred during the transitory phase can be bounded independently of  $\bar{\kappa}_\mu$  - which ultimately removes any trace of the non-linearity in the regret bound.

**On the Bayesian regret lower-bound of (Dong et al, 2019).** The authors of (Dong et al., 2019) consider LogB instances for which there is only one “good” arm ( $a^\top \theta_\star > 0$ ) and many arms in the left-tail of the reward signal ( $a^\top \theta_\star < 0$ ). Their lower-bound is essentially built on the idea that for such an arm-set and given any horizon  $T$ , for any policy it exists a sufficiently non-linear LogB problem ( $\kappa_\mu$  large enough) such that the average number of rounds before the algorithm plays the only good arm is  $\Omega(T)$ . This construction corroborates our interpretation of the transitory regime described hereinbefore as it effectively links the level of non-linearity with a *burn-in* phase during which the agent has to learn to discard low-rewarding arms. In such worst-case instances the problem's structure is of little help because all low-rewarding arms lead with high probability to the same null reward. A direct implication of this construction is that for any policy there exists an arbitrarily non-linear LogB problem such that its Bayesian regret is  $\Omega(T)$ . This suggests that in all generality there is no hope of removing  $\kappa_\mu$  from the regret's second-order term as for the two bounds to be coherent, this term must diverge when  $\kappa_\mu \rightarrow \infty$ . This conclusion is however particularly worst-case; the effects of non-linearity being highly problem-dependent it is natural to wonder if it still holds for arm-sets that evade the construction of Dong et al. (2019). In the rest of this section we answer this question by the negative; for many “reasonable” arm-sets the second-order term of the regret scales *independently* of  $\kappa_\mu$ .

### 3.3.1 Transitory regret and detrimental arms

We introduce below the notion of *detrimental* arms; conceptually they are arms with large sub-optimality gaps that carry little information.

**Definition 3.3.1** (Detrimental arms).

$$\mathcal{A}_- := \begin{cases} \left\{ a \in \mathcal{A}, a^\top \theta_\star \leq -1 \right\} & \text{if } a_\star^\top \theta_\star > 0, \\ \left\{ a \in \mathcal{A}, \dot{\mu}(a^\top \theta_\star) \leq \dot{\mu}(a_\star^\top \theta_\star)/2 \right\} & \text{otherwise.} \end{cases}$$

$\mathcal{A}_-$  contains arms  $a$  such that  $\mu(a^\top \theta_\star) \ll \mu(a_\star^\top \theta_\star)$  (large gap) and  $\dot{\mu}(a^\top \theta_\star) \approx 0$  (small conditional variance). They lay in the far left-tail of the logistic function: their associated reward realization are almost always 0. We provide an illustration of  $\mathcal{A}_-$  in Fig. 3.1.

**Remark 3.3.1** (On the definition of  $\mathcal{A}_-$ ). We use two alternative definitions for  $\mathcal{A}_-$  depending on the sign of  $a_\star^\top \theta_\star$ . This is linked to the two regimes of the logistic function: convex on  $\mathbb{R}^-$  and concave on  $\mathbb{R}^+$ . Detrimental arms suffer from the same negative properties irrespectively of the considered case. For the case  $a_\star^\top \theta_\star > 0$  the reference value  $a^\top \theta_\star \leq -1$  is rather arbitrary; any value  $a^\top \theta_\star \leq -c$  where  $c \ll \|\theta_\star\|$  works similarly.

The following theorem provides a new regret bound for OFU-GLB with a refined second order term which highlights the role of the detrimental arms.

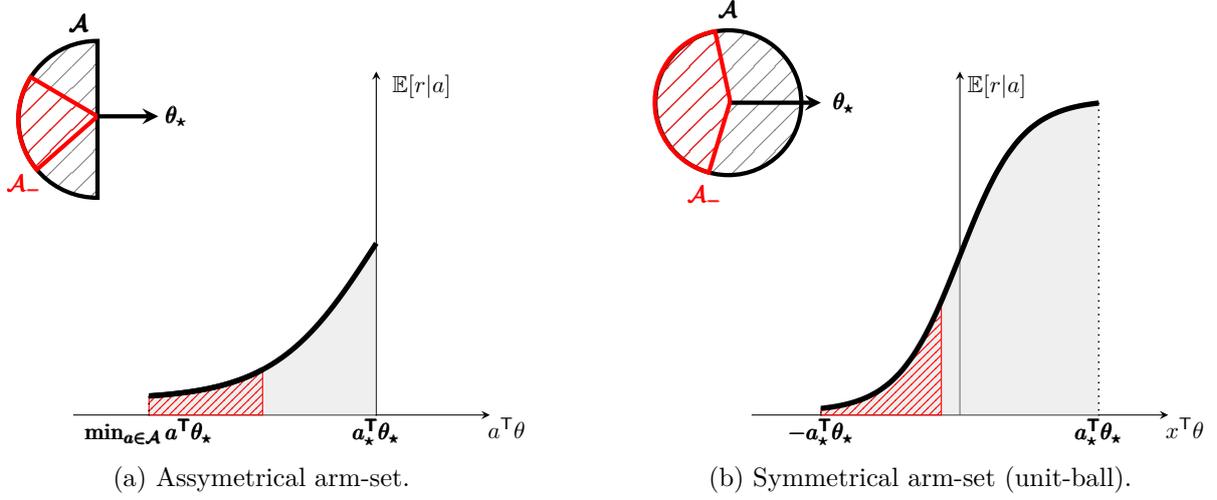


Figure 3.1: Graphical illustration of the detrimental arms  $\mathcal{A}_-$ .

**Theorem 3.3.1.** *On any LogB problem the regret of OFU-GLB satisfies with probability at least  $1 - \delta$ :*

$$\text{Regret}_{\theta_*}(T) = \tilde{\mathcal{O}} \left( \sqrt{\dot{\mu}(a_*^\top \theta_*) T} + \left[ \kappa_\mu \wedge \mu(a_*^\top \theta_*) \sum_{t=1}^T \mathbf{1}(a_t \in \mathcal{A}_-) \right] \right).$$

From [Theorem 3.3.1](#) we can easily retrieve the bound of [Theorem 3.2.2](#). It however leaves room for improvement by stressing that the second order term is significantly smaller when detrimental arms  $\mathcal{A}_-$  are discarded fast enough. This enforces the idea that it is fundamentally tied to an initial burn-in phase after which the detrimental arms are no longer played.

### Sketch of proof

We give below a sketch of proof for [Theorem 3.3.1](#). The entire demonstration is deferred to [Appendix 3.G](#). By an exact second-order Taylor expansion of the regret;

$$\text{Regret}_{\theta_*}(T) = \sum_{t=1}^T \dot{\mu}(a_t^\top \theta_*) (a_* - a_t)^\top \theta_* + \sum_{t=1}^T \tilde{\vartheta}_t \left( (a_* - a_t)^\top \theta_* \right)^2,$$

where we defined  $\tilde{\vartheta}_t = \int_{v=0}^1 (1-v) \ddot{\mu}(a_t^\top \theta_* + v(a_* - a_t)^\top \theta_*) dv$ . The main idea of the proof is to show that when  $a_t \notin \mathcal{A}_-$  then  $\tilde{\vartheta}_t$  is small and therefore so is the resulting second order term. For rounds where  $a_t \in \mathcal{A}_-$  we use a brutal bound to show that:

$$\tilde{\vartheta}_t \left( (a_* - a_t)^\top \theta_* \right)^2 \mathbf{1}(a_t \in \mathcal{A}_-) \leq S \mu(a_*^\top \theta_*) \mathbf{1}(a_t \in \mathcal{A}_-).$$

We now turn to the case  $a_t \in \mathcal{A}_+ = \mathcal{A} \setminus \mathcal{A}_-$  and in this sketch of proof we restrict ourselves to the case  $a_*^\top \theta_* > 0$  so that  $\mathcal{A}_+ = \{a \in \mathcal{A}, a^\top \theta_* \geq -1\}$ . Thanks to the self-concordance property and the concavity of  $\mu$  on  $\mathbb{R}_+$  (i.e.  $\ddot{\mu}(z) < 0$  for  $z > 0$ ) we can show the following bound:

$$\begin{aligned} \tilde{\vartheta}_t \left( (a_* - a_t)^\top \theta_* \right)^2 \mathbf{1}(a_t \in \mathcal{A}_+) &\leq e^1 \dot{\mu}(a_t^\top \theta_*) \left( (a_* - a_t)^\top \theta_* \right)^2 \\ &\leq e^1 \gamma_t(\delta) \dot{\mu}(a_t^\top \theta_*) \|a_t\|_{\mathbf{H}_t^{-1}(\theta_*)}^2. \end{aligned}$$

Note that by the Elliptical Potential lemma  $\sum_{t=1}^T \dot{\mu}(a_t^\top \theta_\star) \|a_t\|_{\mathbf{H}_t^{-1}(\theta_\star)}^2 = \mathcal{O}(d^2 \log(T)^2)$  which is independent of  $\kappa_\mu$ . This yields the following bound on the regret's second-order term:

$$\sum_{t=1}^T \tilde{\vartheta}_t \left( (a_\star - a_t)^\top \theta_\star \right)^2 = \mathcal{O} \left( \sum_{t=1}^T \mathbf{1}(a_t \in \mathcal{A}_-) + d^2 \log(T) \right).$$

Following the proof of [Theorem 3.2.2](#) to bound the first-order term eventually yields:

$$\text{Regret}_{\theta_\star}(T) = \tilde{\mathcal{O}} \left( \sqrt{\dot{\mu}(a_\star^\top \theta_\star) T} + \mu(a_\star^\top \theta_\star) \sum_{t=1}^T \mathbf{1}(a_t \in \mathcal{A}_-) \right).$$

Taking the minimum between this bound and [Theorem 3.2.2](#) provides the announced result.

### 3.3.2 Non-linearity in LogB: a blessing?

**Length of the transitory phase.** In worst-case configuration of the arm-set we cannot guarantee that the regret incurred during the transitory phase is smaller than  $\kappa_\mu$ ; however for more “reasonable” arm-sets it is safe to expect that the permanent regime is reached much faster. We formalize this intuition in the following proposition by exhibiting arm-set structures for which the transitory regime is short.

**Proposition 3.3.1** (Length of transitory regime). *Let  $\{a_1, \dots, a_T\}$  be generated by OFU-GLB on a LogB problem with arm-set  $\mathcal{A}$ . The following holds with high probability:*

$$\mu(a_\star^\top \theta_\star) \sum_{t=1}^T \mathbf{1}(a_t \in \mathcal{A}_-) = \tilde{\mathcal{O}}(d^2 + dK) \quad \text{if } |\mathcal{A}_-| \leq K, \quad (3.13)$$

$$\mu(a_\star^\top \theta_\star) \sum_{t=1}^T \mathbf{1}(a_t \in \mathcal{A}_-) = \tilde{\mathcal{O}}(d^3) \quad \text{if } \mathcal{A} = \mathcal{B}_d(0, 1). \quad (3.14)$$

This result formalizes that OFU-GLB quickly discards detrimental arms when [\(3.13\)](#) there are only a few or [\(3.14\)](#) the problem's structure is symmetric. In such case the final regret bound is oblivious to non-linearity;

$$\text{Regret}_{\theta_\star}(T) = \tilde{\mathcal{O}} \left( \sqrt{\dot{\mu}(a_\star^\top \theta_\star) T} \right) \quad \text{w.h.p.}$$

While [Proposition 3.3.1](#) only identifies two such case where the non-linearity does not affect the length of the transitory regime, we expect this enjoyable property to hold for many “reasonable” arm set structures - it breaks for fairly peculiar arm-set such as [Dong et al. \(2019\)](#)'s counter-example or the arm-set of [Fig. 3.1a](#). The intuition is that as soon as the detrimental arms are few in number or in proportion OFU-GLB quickly discover “good” arms - which potentially carry little information (small conditional variance) but which sub-optimality gaps are small (they lie in the positive flat tail of the reward signal).

**Remark 3.3.2** (On the superiority of the parameter-search approach in LogB). *The results presented in this section for OFU-GLB easily extends to OFU-GLB-r. They could also be extended to GLM-UCB+ but at the price of a modified algorithmic design. Indeed the regret bound of GLM-UCB+ is tied to the amount of exploration which is hard-coded in its exploration bonus. Therefore to obtain tight regret bounds this bonus must be re-computed for every arm-set  $\mathcal{A}$  by re-deriving a tight prediction error bound adapted to the arm-set's geometry. This is not the case for the parameter search approach as this step is pushed to analysis time and does not impact the effective performance of the algorithm. In other words in the LogB setting (and virtually for all GLBs) the parameter-search approach automatically adapts to the effective complexity of the problem encoded in the arm-set's geometry. This is not the case for its bonus-based counterpart.*

**The unit ball case.** The following result is obtained by merging [Theorem 3.3.1](#) with [Proposition 3.3.1](#) and embodies the improvement over previous work obtained with our approach.

**Corollary 3.3.1** (LogB unit ball regret bound). *Consider a LogB problem with  $\mathcal{A} = \mathcal{B}_d(0, 1)$ . Then  $\mu(a_\star^\top \theta_\star) \leq \exp(-\|\theta_\star\|)$  and the regret of OFU-GLB on this problem satisfies w.h.p:*

$$\text{Regret}_{\theta_\star}(T) = \tilde{\mathcal{O}}\left(\exp(-\|\theta_\star\|/2)\sqrt{T}\right).$$

This is an exponential improvement over the performance of GLM-UCB which regret bound on the same problem is  $\mathcal{O}(\exp(\|\theta_\star\|)\sqrt{T})$ . Furthermore it tells a much different story about the effects of non-linearity in LogB. Indeed in this particular configuration the level of non linearity is directly tied to  $\|\theta_\star\|$  as  $\kappa_\mu \geq 4\exp(\|\theta_\star\|)$ . Therefore the larger  $\|\theta_\star\|$ , the higher the level of non-linearity of the reward signal, and by [Corollary 3.3.1](#) the *smaller* the regret. In this case the non-linearity is therefore *beneficial* for the regret minimization task which turns out to be even easier than for its LB counterpart problem.

**Remark 3.3.3.** *The conclusion on the effects of non-linearity detailed above has to be handled with care as it holds only for the Logistic Bandit - the analysis conducted in the section heavily leverages the specific properties of this model. In the Poisson Bandit setting it is likely that the regret's second-order term still scales linearly with the reward sensitivity ratio. Further, the first order term scales with  $\mu(a_\star^\top \theta_\star)$  which for the Poisson Bandit in the unit ball case evolves as  $\exp(\|\theta_\star\|)$ ; the “exploding” behavior of the reward signal around the optimum still negatively impacts the performance of our algorithms (although in a less dramatic fashion than for GLM-UCB). This remark emphasizes the problem-dependent impacts of non-linearity; a thorough understanding for a given problem requires a precise problem-dependent analysis (such as the one conducted above for the LogB). In other words not all GLBs are equal in face of non-linearity and to be completely described, virtually each different GLB requires a dedicated analysis.*

### Sketch of proof

We provide here a sketch of proof for [Eq. \(3.13\)](#) from [Proposition 3.3.1](#); the complete demonstration is deferred to [Appendix 3.H](#). As usual we work under  $\{\forall t \geq 1, \theta_\star \in \mathcal{C}_t(\delta)\}$ ; this happens with probability at least  $1 - \delta$ . We restrict this sketch of proof to the case  $a_\star^\top \theta_\star \geq 0$ . In this case detrimental arms have a large constant sub-optimality gap as for any  $a \in \mathcal{A}_-$ :

$$\mu(a_\star^\top \theta_\star) - \mu(a^\top \theta_\star) \geq \mu(a_\star^\top \theta_\star) - \mu(-1) \geq 1/2 - \mu(-1) \geq 1/5.$$

We can use this result to show that OFU-GLB plays detrimental arms only logarithmically often. For any  $a \in \mathcal{A}_-$  let  $\tau_a$  be the last time-step when  $a$  is played and  $N_a$  the number of time  $a$  was played over  $[T]$ . By using the lower-bound on the sub-optimality gap for any  $a \in \mathcal{A}_-$ :

$$\begin{aligned} 1/5 &\leq \mu\left(a_\star^\top \theta_\star\right) - \mu\left(a_{\tau_a}^\top \theta_\star\right) \\ &\leq \mu\left(a_{\tau_a}^\top \theta_{\tau_a}\right) - \mu\left(a_{\tau_a}^\top \theta_\star\right) && \text{(optimism)} \\ &\leq \alpha(a_{\tau_a}, \theta_{\tau_a}, \theta_\star) a_{\tau_a}^\top (\theta_{\tau_a} - \theta_\star) && \text{(mean-value theorem)} \\ &= \alpha(a_{\tau_a}, \theta_{\tau_a}, \theta_\star) a_{\tau_a}^\top \mathbf{G}_{\tau_a}^{-1}(\theta_{\tau_a}, \theta_\star) (g_{\tau_a}(\theta_{\tau_a}) - g_{\tau_a}(\theta_\star)) && \text{(Eq. (1.21))} \\ &\leq \alpha(a_{\tau_a}, \theta_{\tau_a}, \theta_\star) \|a_{\tau_a}\|_{\mathbf{G}_{\tau_a}^{-1}(\theta_{\tau_a}, \theta_\star)} \|g_{\tau_a}(\theta_{\tau_a}) - g_{\tau_a}(\theta_\star)\|_{\mathbf{G}_{\tau_a}^{-1}(\theta_{\tau_a}, \theta_\star)} && \text{(Cauchy-Schwarz)} \\ &\leq 2\sqrt{1 + 2S}\gamma_{\tau_a}(\delta) \alpha(a_{\tau_a}, \theta_{\tau_a}, \theta_\star) \|a_{\tau_a}\|_{\mathbf{G}_{\tau_a}^{-1}(\theta_{\tau_a}, \theta_\star)} && \text{(3.15)} \end{aligned}$$

where we last used  $\|g_t(\theta_t) - g_t(\theta_\star)\|_{\mathbf{G}_t^{-1}(\theta_t, \theta_\star)} \leq 2\sqrt{1 + 2S}\gamma_t(\delta)$ . Note also that  $\mathbf{G}_{\tau_a}(\theta_{\tau_a}, \theta_\star) \succeq N_a \alpha(a, \theta_{\tau_a}, \theta_\star) a a^\top + \lambda_{\tau_a} \mathbf{I}_d$ . It is therefore easy to show (for instance, using the Sherman-Morison

formula) that  $\|a_{\tau_a}\|_{\mathbf{G}_{\tau_a}^{-1}(\theta_{\tau_a}, \theta_*)}^2 \leq (\alpha(a_{\tau_a}, \theta_{\tau_a}, \theta_*)N_a)^{-1}$ . We therefore finally obtain by injecting this into Eq. (3.30):

$$N_a \leq 100(1 + 2S)\gamma_{\tau_a}(\delta)^2 \alpha(a_{\tau_a}, \theta_{\tau_a}, \theta_*) \leq 25(1 + 2S)\gamma_{\tau_a}(\delta)^2$$

Recall that  $\bar{\gamma}_T(\delta) = \mathcal{O}(\sqrt{d \log(T)})$ . This finishes the proof as:

$$\sum_{t=1}^T \mathbf{1}(a_t \in \mathcal{A}_-) = \sum_{a \in \mathcal{A}_-} N_a \leq 25(1 + 2S)|\mathcal{A}_-| \bar{\gamma}_T(\delta)^2 .$$

### 3.4 (✂) Optimality of the permanent regret in LogB.

We dedicated the last section to a finer understanding of the transitory regime by focusing on the LogB setting. We continue in this spirit but focus on the permanent regime during which we saw that algorithms suffer a  $\tilde{\mathcal{O}}(d\sqrt{\dot{\mu}(a_*^\top \theta_*)T})$  regret. An important question remains; is this *optimal*? We focus in this section on answering this question in the LogB case.

#### 3.4.1 Regret lower-bound

To simplify notations we introduce the following notation:

$$\kappa_*(\theta) := 1/\dot{\mu}(a_*(\theta)^\top \theta) \quad \text{for } \theta \in \Theta ,$$

and re-write the regret upper-bound of OFU-GLB and GLM-UCB+ as:

$$\text{Regret}_{\theta_*}(T) = \tilde{\mathcal{O}}\left(d\sqrt{T/\kappa_*(\theta_*)}\right) .$$

Our goal is to show that this bound is optimal by going after a *problem-dependent* lower-bound.

**Challenges.** Studying minimax-optimality w.r.t to  $d$ ,  $T$  and  $\kappa_*(\theta_*)$  altogether raises new challenges for proving lower-bounds. Because  $\kappa_*(\theta_*)$  is a problem-dependent quantity, obtaining a meaningful lower-bound requires to identify a entire set of hard problem instances which comes with a wide range of values for  $\kappa_*(\theta_*)$  (especially large values which are in the domain of interest for our study as they are tied to highly non-linear instances). Unfortunately this precludes reproducing the lower-bound strategy laid out for instance in (Lattimore and Szepesvári, 2020, Theorem 24.2) as their construction relies on problem for which  $\|\theta_*\| \approx d/\sqrt{T}$ . Such problems come with small values of  $\kappa_*(\theta_*)$  (especially when  $T$  is large); this can still lead to valid lower-bound, however in this case not extremely meaningful ones as the range of problems they cover is rather limited.

**Local minimax regret.** To achieve our goal we introduce a “local” notion of minimax regret inspired by the bound of Simchowitz and Foster (2020) in a reinforcement learning setting. For any policy and given *any* reference point  $\theta_*$  we search for the *hardest* nearby alternative which shares the same problem-dependent constant as  $\theta_*$ . Formally for  $\epsilon > 0$  we define the local minima (expected) regret:

$$\text{MinimaxRegret}_{\theta_*}^T(\epsilon) := \min_{\pi} \max_{\|\theta - \theta_*\| \leq \epsilon} \mathbb{E}[\text{Regret}_{\theta}^{\pi}(T)] .$$

The following theorem proves that the regret incurred by GLM-UCB+ and OFU-GLB in the long-term regime (large  $T$ ) is *minimax-optimal* w.r.t  $T$ ,  $d$  and  $\kappa_*$ .

**Theorem 3.4.1** (Problem-dependent Logistic Bandit lower-bound). *Let  $\mathcal{A} = \mathcal{S}_d(0, 1)$ . For any problem instance  $\theta_*$  and for  $T \geq d^2 \kappa_*(\theta_*)$ , there exist  $\epsilon_T$  small enough such that:*

- (1)  $\frac{5}{6} \kappa_*(\theta_*) \leq \kappa_*(\theta) \leq \frac{6}{5} \kappa_*(\theta_*)$  for all  $\theta \in \{ \|\theta - \theta_*\| \leq \epsilon_T \}$ ,
- (2)  $\text{MinimaxRegret}_{\theta_*}^T(\epsilon_T) = \Omega \left( d \sqrt{T / \kappa_*(\theta_*)} \right)$ .

This local lower-bound naturally implies a global one, stated below.

**Corollary 3.4.1** (Global Logistic Bandit Lower-Bound). *Let  $\mathcal{A} = \mathcal{S}_d(0, 1)$ . For any policy  $\pi$  and for any tuple  $(T, d, \kappa)$  such that  $T \geq d^2 \kappa$ , there exists a problem  $\theta$  such that  $\kappa_*(\theta) = \kappa$  and:*

$$\text{Regret}_{\theta}^{\pi}(T) = \Omega \left( d \sqrt{T / \kappa} \right).$$

### 3.4.2 Proof of the lower-bound

**High level idea.** We discuss here the construction of our local lower-bound. Let  $\theta_*$  denote a fixed nominal instance and  $\pi$  a policy which has low-regret when playing against  $\theta_*$ . Our strategy is to find an alternative problem  $\theta'$  which satisfies the two following *conflicting* criteria: **(1)**  $\pi$  has the same behavior against both  $\theta_*$  and  $\theta'$  and **(2)**  $\theta'$  is *far* from  $\theta_*$  so that the optimal arms  $a_*(\theta_*)$  and  $a_*(\theta')$  significantly *differ*. When playing against  $\theta_*$  we can expect  $\pi$  to produce a trajectory where most of the time  $a_t \approx a_*(\theta_*)$ . Indeed since:

$$\text{Regret}_{\theta_*}^{\pi}(T) \propto \sum_{t=1}^T \|a_t - a_*(\theta_*)\|^2,$$

a small regret against  $\theta_*$  implies an accurate tracking of  $a_*(\theta_*)$ . Notice that when  $\mathcal{A} = \mathcal{S}_d(0, 1)$  the optimal arm  $a_*(\theta_*)$  is co-linear with  $\theta_*$ . As a consequence there are  $d - 1$  directions (the orthogonal complement of  $\theta_*$ ) where  $\theta_*$  is poorly estimated. This suggest that parameters laying in  $\mathcal{H}_{\perp}^*$  (the hyperplane supported by  $\theta_*$ , cf. Fig. 3.2) can easily be confused with  $\theta_*$  for the policy  $\pi$ . This notion of *distinguishability* between parameters can be formalized through a discrepancy measures  $d_T(\theta_*, \theta')$  which quantifies how easy it is for  $\pi$  to determine if the rewards it receives are generated by either  $\theta_*$  or  $\theta'$ . For any  $\theta' \in \mathcal{H}_{\perp}^*$  it scales as follow:

$$d_T(\theta_*, \theta') \approx \sqrt{\frac{T}{\kappa_*(\theta_*)}} \|\theta_* - \theta'\|^2.$$

This scaling is intuitive; the larger  $T$ , the more occasions for  $\pi$  to separate  $\theta_*$  from  $\theta'$ . Further, the larger  $\kappa_*$ , the smaller the conditional variance of the rewards and the longer it takes to correctly estimate an arm's mean reward and determine whether it was generated by  $\theta_*$  or  $\theta'$ . To satisfy **(1)** we must choose  $\theta'$  so that  $d_T(\theta_*, \theta')$  is small; the trade-off with **(2)** suggests picking  $\theta'$  such that:

$$\|\theta' - \theta_*\|^2 \approx \sqrt{\frac{\kappa_*(\theta_*)}{T}}. \quad (3.16)$$

Under such conditions,  $\pi$  cannot separate  $\theta_*$  from  $\theta'$  and must therefore *act* similarly against both parameters (i.e most of the time we will have  $x_t \approx x_*(\theta_*)$  against  $\theta'$ ). Easy computations

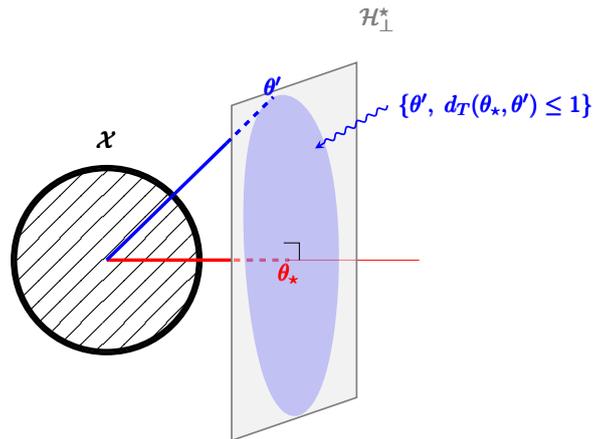


Figure 3.2: Illustration of the construction behind the local lower-bound.

show that the regret of  $\pi$  against  $\theta'$  then writes:

$$\begin{aligned} \text{Regret}_{\theta'}^{\pi}(T) &\approx \frac{1}{\kappa_{\star}(\theta_{\star})} \sum_{t=1}^T \|a_t - a_{\star}(\theta')\|^2, \\ &\approx \frac{1}{\kappa_{\star}(\theta_{\star})} \sum_{t=1}^T \|a_{\star}(\theta_{\star}) - a_{\star}(\theta')\|^2, \\ &\approx \frac{1}{\kappa_{\star}(\theta_{\star})} T \|\theta_{\star} - \theta'\|^2. \end{aligned}$$

which gives the announced behavior after replacing  $\|\theta_{\star} - \theta'\|$  by the scaling suggested by the trade-off between **(1)** and **(2)** presented in Eq. (3.16).

**Formal proof.** We follow Lattimore and Szepesvári (2020) and will note  $(\Omega_t, \mathcal{F}_t, \mathbb{P}_{\pi\theta})$  the *canonical* bandit probability space at round  $t$  under the parameter  $\theta$ . A thorough definition of this probability space can be found in (Lattimore and Szepesvári, 2020, Section 4.7). To simplify notations, we will denote  $\mathbb{P}_{\theta} = \mathbb{P}_{\pi\theta}$  the probability measure of the random sequence  $\{a_1, r_2, \dots, a_T, r_{T+1}\}$ , obtained by having  $\pi$  interact with the environment parameter  $\theta$ . We work in a logistic bandit setting meaning that at any round  $t$  and conditionally on  $a_t$  being played:

$$r_{t+1} \sim \text{Bernoulli}(\mu(a_t^{\top} \theta))$$

where  $\mu(z) = (1 + \exp(-z))^{-1}$  is the logistic function. We fix the policy  $\pi$  for now and start the proof with the following result which ties the regret incurred against  $\theta$  to the *tracking* of  $a_{\star}(\theta)$ .

**Proposition 3.4.1.** *For all  $\theta \in \mathbb{R}^d$  the following holds:*

$$\text{Regret}_{\theta}^{\pi}(T) \geq \frac{\|\theta\|}{\kappa_{\star}(\theta)} \sum_{i=1}^d \mathbb{E}_{\theta} \left[ \sum_{t=1}^T [a_{\star}(\theta) - a_t]_i^2 \right]. \quad (3.17)$$

Let us fix  $\theta_{\star}$  as an arbitrary parameter that will serve as our reference point and let  $\{e_i\}_{i=1}^d$  the canonical basis of  $\mathbb{R}^d$ . Without loss of generality we assume that  $\theta_{\star} = \|\theta_{\star}\| e_1$ . With such notations, we now introduce the set of *unidentifiable* parameters:

$$\Xi_{\varepsilon} := \left\{ \theta_{\star} + \varepsilon \sum_{i=2}^d v_i e_i, v \in \{-1, 1\}^{d-1} \right\} \subset \mathcal{H}_{\perp}^{\star},$$

where  $\varepsilon$  is a small positive scalar to be tuned later.  $\Xi_\varepsilon$  is a set of slightly perturbed versions of  $\theta_\star$ . The goal is to set  $\varepsilon$  small enough so that **(1)** a policy interacting with any  $\theta \in \Xi_\varepsilon$  is unable to tell with high confidence which parameter generates the rewards but however large enough so **(2)** the policy cannot perform simultaneously well on every  $\theta \in \Xi_\varepsilon$ . To find such  $\varepsilon$  we are going to reason by contradiction and assume that the policy  $\pi$  performs well for every  $\theta \in \Xi_\varepsilon$ . Note that all the elements  $\theta$ s of  $\Xi_\varepsilon$  have the same norm and because  $\mathcal{A} = \mathcal{S}_d(0, 1)$  they also share the same  $\kappa_\star(\theta)$ , which we will denote  $\kappa_\varepsilon$  for short. Note also that  $\kappa_\varepsilon \geq \kappa_\star(\theta_\star)$ . We make the following hypothesis which will lead us to a contradiction for the right value of  $\varepsilon$ .

**Hypothesis 3.4.1.** *There exists a universal constant  $C$  such that:*

$$\forall \theta \in \Xi_\varepsilon \text{ we have } \text{Regret}_\theta^\pi(T) < Cd\sqrt{T/\kappa_\varepsilon}.$$

Without loss of generality we will take  $C = 1$ . We now use [Proposition 3.4.1](#) and introduce the optimal action for our reference point  $\theta_\star$ . By doing so we obtain the following result which proof is deferred to [Section 3.1.2](#).

**Lemma 3.4.1.** *For each  $\theta \in \Xi_\varepsilon$  and any direction  $i \in [d, 2]$  let us introduce the event:*

$$A_i(\theta) := \left\{ [a_\star(\theta) - a_\star(\theta_\star)]_i \cdot \left[ \frac{1}{T} \sum_{t=1}^T a_t - a_\star(\theta_\star) \right]_i \geq 0 \right\}.$$

Then for any  $\theta \in \Xi_\varepsilon$  we have:

$$\text{Regret}_\theta^\pi(T) \geq \frac{T\varepsilon^2}{2\kappa_\varepsilon \|\theta_\star\|} \sum_{i=2}^d \mathbb{P}_\theta(A_i(\theta)).$$

The goal is now to find one  $\theta \in \Xi_\varepsilon$  such that the above lower-bound is large. This can be done thanks to a *averaging hammer* as in ([Lattimore and Szepesvári, 2020](#), Section 24.1). We will need a *flipping* operator  $\text{Flip}_i(\cdot)$  which for any  $\theta \in \Xi_\varepsilon$  changes the sign of the  $i^{\text{th}}$  coordinate of  $\theta$ . Formally, let:

$$[\text{Flip}_i(\theta)]_i = -[\theta]_i \quad \text{and} \quad [\text{Flip}_i(\theta)]_j = [\theta]_j \quad \text{for all } j \neq i. \quad (3.18)$$

In the following lemma we show that the average value of  $\sum_{i=2}^d \mathbb{P}_\theta(A_i(\theta))$  over  $\Xi_\varepsilon$  is linked to the average relative entropy (denoted  $D_{\text{KL}}$ ) between the probability measures induced by *flipped* versions of  $\theta$ . The proof is deferred to [Section 3.1.3](#).

**Lemma 3.4.2** (Averaging Hammer). *The following holds:*

$$\frac{1}{|\Xi_\varepsilon|} \sum_{\theta \in \Xi_\varepsilon} \sum_{i=2}^d \mathbb{P}_\theta(A_i(\theta)) \geq \frac{d}{4} - \frac{\sqrt{d}}{2} \sqrt{\frac{1}{|\Xi_\varepsilon|} \sum_{\theta \in \Xi_\varepsilon} \sum_{i=2}^d D_{\text{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)})}.$$

We now have to characterize this average relative entropy. This is done in the following Lemma thanks to [Hypothesis 3.4.1](#); the proof is presented in [Section 3.1.4](#).

**Lemma 3.4.3** (Average Relative Entropy). *Under [Hypothesis 3.4.1](#) we have:*

$$\frac{1}{|\Xi_\varepsilon|} \sum_{\theta \in \Xi_\varepsilon} \sum_{i=2}^d D_{\text{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)}) \leq \frac{2}{\kappa_\varepsilon} dT\varepsilon^4 \exp(4\varepsilon) + 4d\varepsilon^2 \exp(4\varepsilon) \left(6 + \frac{d}{2}\varepsilon^2\right) \sqrt{\frac{T}{\kappa_\varepsilon}}.$$

Combining [Lemmas 3.4.2](#) and [3.4.3](#) we therefore obtain that:

$$\frac{1}{|\Xi_\varepsilon|} \sum_{\theta \in \Xi_\varepsilon} \sum_{i=2}^d \mathbb{P}_\theta(A_i(\theta)) \geq \frac{d}{4} \left[ 1 - 2 \left( 2\epsilon^4 \frac{T}{\kappa_\epsilon} + 24\epsilon^2 \sqrt{\frac{T}{\kappa_\epsilon}} + 2d\epsilon^4 \sqrt{\frac{T}{\kappa_\epsilon}} \right)^{1/2} \exp(2\epsilon) \right]$$

Because this results holds for an average over  $\Xi_\varepsilon$ , it must still be true for at least one  $\tilde{\theta} \in \Xi_\varepsilon$ . In other words, there exists  $\tilde{\theta} \in \Xi_\varepsilon$  such that:

$$\sum_{i=2}^d \mathbb{P}_{\tilde{\theta}}(A_i(\tilde{\theta})) \geq \frac{d}{4} \left[ 1 - 2 \left( 2\epsilon^4 \frac{T}{\kappa_\epsilon} + 24\epsilon^2 \sqrt{\frac{T}{\kappa_\epsilon}} + 2d\epsilon^4 \sqrt{\frac{T}{\kappa_\epsilon}} \right)^{1/2} \exp(2\epsilon) \right].$$

Thanks to [Lemma 3.4.1](#) and by using  $\kappa_\varepsilon \geq \kappa_\star(\theta_\star)$  we have that:

$$\text{Regret}_{\tilde{\theta}}^\pi(T) \geq dT \frac{\epsilon^2}{8 \|\theta_\star\| \kappa_\epsilon} \left[ 1 - 2 \left( 2\epsilon^4 \frac{T}{\kappa_\star(\theta_\star)} + 24\epsilon^2 \sqrt{\frac{T}{\kappa_\star(\theta_\star)}} + 2d\epsilon^4 \sqrt{\frac{T}{\kappa_\star(\theta_\star)}} \right)^{1/2} \exp(2\epsilon) \right].$$

Now is the time to tune  $\epsilon$ . Taking  $\epsilon^2 = \frac{1}{32} \sqrt{\frac{\kappa_\star(\theta_\star)}{T}}$  yields after some computations that:

$$\text{Regret}_{\tilde{\theta}}^\pi(T) \geq \frac{d\sqrt{T}}{256 \|\theta_\star\|} \frac{\sqrt{\kappa_\star(\theta_\star)}}{\kappa_\epsilon} \left( 1 - 2 \left( \frac{24576}{32^4} + \frac{2}{32^4} d \sqrt{\frac{\kappa_\star(\theta_\star)}{T}} \right)^{1/2} \exp \left( \frac{2}{\sqrt{32}} \sqrt{\frac{\kappa_\star(\theta_\star)}{T}} \right) \right).$$

When  $T \geq d^2 \kappa_\star(\theta_\star)$  we obtain after some computations that:

$$\text{Regret}_{\tilde{\theta}}^\pi(T) \geq \frac{d\sqrt{T}}{512 \|\theta_\star\|} \frac{\sqrt{\kappa_\star(\theta_\star)}}{\kappa_\epsilon}.$$

The only missing step requires tying  $\kappa_\star(\theta_\star)$  and  $\kappa_\varepsilon$ . Because  $\varepsilon$  is small  $\theta_\star$  and any nearby alternative  $\theta \in \Xi_\varepsilon$  share the same problem-dependent constants  $\kappa_\star$ . Indeed by a direct application of [Lemma 1.B.3](#) (a self-concordance control result) for any  $\theta \in \Xi_\varepsilon$ :

$$\kappa_\varepsilon \exp(-\sqrt{d}\epsilon) \leq \kappa_\star(\theta_\star) \leq \kappa_\varepsilon \exp(\sqrt{d}\epsilon)$$

Therefore since  $d\epsilon^2 = (1/32)d\sqrt{\kappa_\star(\theta_\star)/T} \leq 1/32$  when  $T \geq d^2 \kappa_\star(\theta_\star)$  we obtain that:

$$5\kappa_\varepsilon/6 \leq \kappa_\star(\theta_\star) \leq 6\kappa_\varepsilon/5.$$

which proves the claim 2. of the theorem. To sum-up, we have shown that when [Hypothesis 3.4.1](#) holds it exists  $\tilde{\theta} \in \Xi_\varepsilon$  with  $\epsilon^2 = \frac{1}{32} \sqrt{\kappa_\star(\theta_\star)/T}$  such that if  $T \geq d^2 \kappa_\star(\theta_\star)$ :

$$\text{Regret}_{\tilde{\theta}}^\pi(T) = \Omega \left( d\sqrt{T/\kappa_\star(\theta_\star)} \right).$$

Of course if [Hypothesis 3.4.1](#) does not hold the above result is guaranteed by definition. We have therefore proven that for any policy if  $T \geq d^2 \kappa_\varepsilon$  and  $\epsilon^2 = \frac{1}{32} \sqrt{\kappa_\varepsilon/T}$ :

$$\max_{\|\theta - \theta_\star\|^2 \leq d\epsilon^2} \text{Regret}_{\tilde{\theta}}^\pi(T) = \Omega \left( d\sqrt{T/\kappa_\star(\theta_\star)} \right).$$

This holds for any policy  $\pi$  which proves the point 1. of the claimed result.

### 3.5 Numerical simulations

We finish this chapter with this short experimental section reporting numerical simulations illustrating our theoretical results. All experiments are run on Logistic Bandit instances. We start by comparing the performance of **GLM-UCB** with **OFU-GLB-r**. The reason for omitting **GLM-UCB+** and **OFU-GLB** is practical; indeed those algorithms require running non-convex optimization routines that cannot be bypassed. On the contrary and as we emphasized earlier **OFU-GLB-r** is fully tractable and requires only solving convex programs. Note that in all generality **GLM-UCB** also requires solving a non-convex optimization program at every round; it can however be bypassed whenever  $\hat{\theta}_t \in \Theta$ . We enforce this by adopting a strong enough regularization and check that it yields in practice the desired behavior. The results presented in [Fig. 3.3](#) and [Section 3.5](#), corroborate our theoretical analysis: **(1)** **OFU-GLB-r** displays a clear advantage over previous **GLM-UCB** ([Figs. 3.3a](#) and [3.3b](#)) and **(2)** in the Logistic Bandit case a higher level of non-linearity (i.e higher values of  $\kappa_*$ ) is actually beneficial (see [Fig. 3.4b](#)) for **OFU-GLB-r**. This cannot be the case for **GLM-UCB** as its performances can only degrade when the level of non-linearity increases as confirmed in [Fig. 3.4a](#).

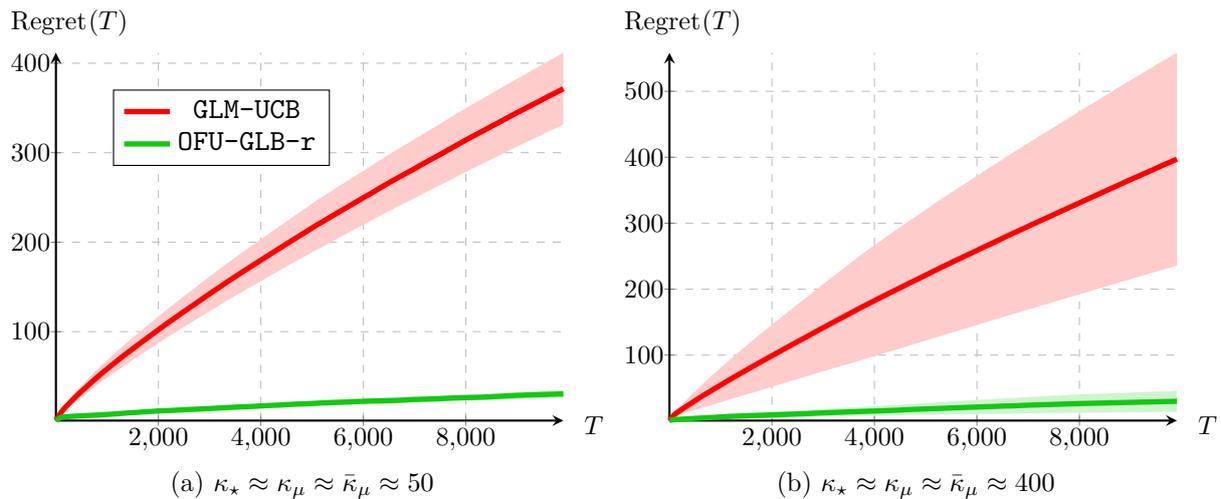


Figure 3.3: Empirical comparison of GLM-UCB and OFU-GLB-r on two LogB toy experiments. The regret curves are averaged over 50 independent runs. Standard-deviation is reported in shaded colors around the averaged cumulative regret. The arm-set  $\mathcal{A}$  is composed of 40 arms drawn uniformly at random in the 2-dimensional ball at the beginning of each run. We provide both algorithms with the perfect knowledge of  $\|\theta_\star\|$  - that is  $S = \|\theta_\star\|$ . This allows to approximate the unit-ball case as in this scenario  $\kappa_\star \approx \kappa_\mu \approx \bar{\kappa}_\mu$ . GLM-UCB requires a knowledge of  $\bar{\kappa}_\mu$ ; we provide it with an upper-bound on this quantity, computed on the unit-ball - that is for  $\mathcal{A} = \mathcal{B}_2(0, 1)$ . For both the considered cases where we vary  $\|\theta_\star\|$  (leading to different reward sensitivity problem-dependent constants) OFU-GLB-r largely outperforms GLM-UCB as predicted by the different regret upper-bounds.

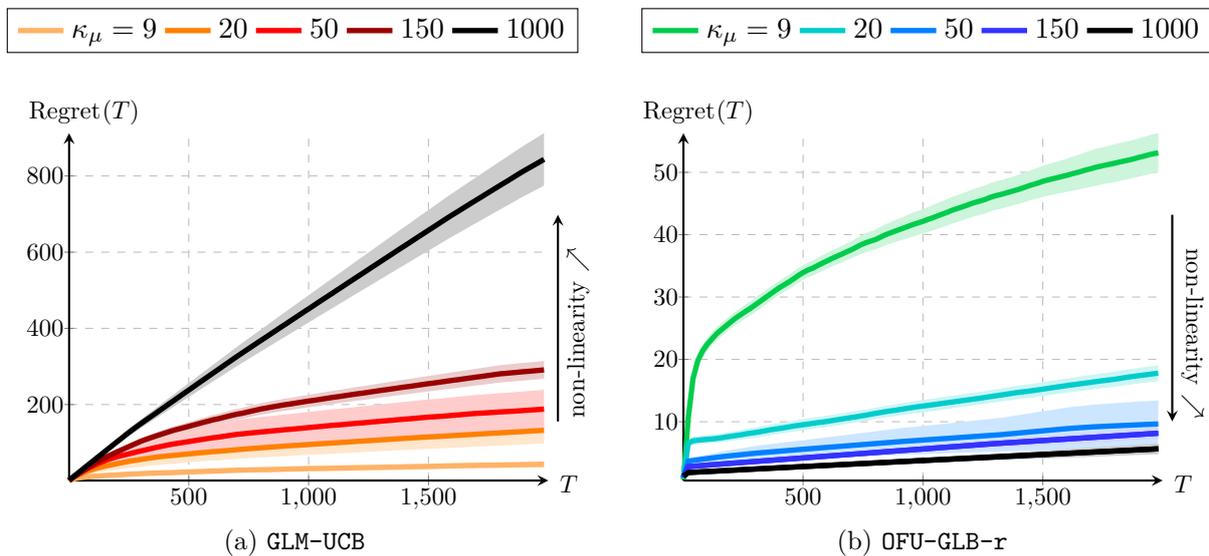


Figure 3.4: Comparing the effect of non-linearity on GLM-UCB and OFU-GLB-r by varying the level of non-linearity in a Logistic Bandit setting. We reproduce the experimental set-up of Fig. 3.3. As predicted by the different regret bounds the performance of GLM-UCB degrades when the level of non-linearity increases (e.g when  $\|\theta_\star\|$  or equivalently  $\kappa_\star \approx \kappa_\mu \approx \bar{\kappa}_\mu$  increases). In contrast the performance of OFU-GLB *improves* when the level non-linearity increases.

# Appendix

## Appendix 3.A Proof of Proposition 3.A.1

**Proposition 3.A.1** (Regret and exploration bonus). *Recall the prediction error  $\Delta_t(a) = |\mu(a^\top \tilde{\theta}_t) - \mu(a^\top \theta_\star)|$ . If for all  $a \in \mathcal{A}$  and  $t \in [T]$  we have  $\varepsilon_t(a) \geq \Delta_t(a)$  then:*

$$\text{Regret}_{\theta_\star}(T) \leq 2 \sum_{t=1}^T \varepsilon_t(a_t) .$$

*Proof.* By removing and adding  $\sum_{t=1}^T \mu(a_\star^\top \tilde{\theta}_t)$  and  $\sum_{t=1}^T \mu(a_t^\top \tilde{\theta}_t)$  one has that:

$$\begin{aligned} \text{Regret}_{\theta_\star}(T) &= \sum_{t=1}^T \mu(a_\star^\top \theta_\star) - \mu(a_t^\top \theta_\star) \\ &= \left[ \sum_{t=1}^T \mu(a_\star^\top \theta_\star) - \mu(a_\star^\top \tilde{\theta}_t) \right] + \left[ \sum_{t=1}^T \mu(a_t^\top \tilde{\theta}_t) - \mu(a_t^\top \theta_\star) \right] + \left[ \sum_{t=1}^T \mu(a_\star^\top \tilde{\theta}_t) - \mu(a_t^\top \tilde{\theta}_t) \right] \\ &\leq \sum_{t=1}^T \Delta_t(a_\star) + \sum_{t=1}^T \Delta_t(a_t) + \left[ \sum_{t=1}^T \mu(a_\star^\top \tilde{\theta}_t) - \mu(a_t^\top \tilde{\theta}_t) \right] \\ &\leq \sum_{t=1}^T \Delta_t(a_\star) + \sum_{t=1}^T \Delta_t(a_t) + \sum_{t=1}^T \varepsilon_t(a_t) - \sum_{t=1}^T \varepsilon_t(a_\star) . \end{aligned}$$

where we last used the definition of the action selection process  $a_t = \arg \max_{a \in \mathcal{A}} \mu(a^\top \tilde{\theta}_t) + \varepsilon_t(a)$  which yields:

$$\mu(a_t^\top \tilde{\theta}_t) + \varepsilon_t(a_t) \geq \mu(a_\star^\top \tilde{\theta}_t) + \varepsilon_t(a_\star) .$$

If we have  $\varepsilon_t(a) \geq \Delta_t(a)$  for all  $t \in [T]$  and  $a \in \mathcal{A}$  we obtained the claimed result. ■

## Appendix 3.B Proof of Lemma 3.1.1

**Lemma 3.1.1** (Confident prediction-error upper-bound). *Under the event  $\{\theta_\star \in \mathcal{C}_t(\delta), \forall t \geq 1\}$  for any  $a \in \mathcal{A}$  and  $t \geq 1$ :*

$$\Delta_t(a) \leq 2(1 + 2S)\dot{\mu}(a^\top \tilde{\theta}_t) \|a\|_{\mathbf{H}_t^{-1}(\tilde{\theta}_t)} \gamma_t(\delta) + 2(1 + 2S)^2 \bar{\kappa}_\mu \gamma_t^2(\delta) \|a\|_{\mathbf{V}_t^{-1}}^2 = \varepsilon_t(a) .$$

*Proof.* By a second-order Taylor expansion:

$$\Delta_t(a) \leq \dot{\mu}(a^\top \tilde{\theta}_t) |a^\top (\tilde{\theta}_t - \theta_\star)| + (a^\top (\tilde{\theta}_t - \theta_\star))^2 \int_{v=0}^1 (1-v) |\ddot{\mu}(a^\top \theta_\star + va^\top (\tilde{\theta}_t - \theta_\star))| dv .$$

By [Assumption 1.4.1](#) we have  $|\ddot{\mu}| \leq \bar{\mu}$ ; therefore  $\max_{a \in \mathcal{A}, \theta \in \Theta} |\ddot{\mu}| \leq \bar{L}_\mu$  and:

$$\begin{aligned} \Delta_t(a) &\leq \dot{\mu}(a^\top \tilde{\theta}_t) |a^\top (\tilde{\theta}_t - \theta_\star)| + (\bar{L}_\mu/2) (a^\top (\tilde{\theta}_t - \theta_\star))^2 \\ &\leq \dot{\mu}(a^\top \tilde{\theta}_t) \|a\|_{\mathbf{H}_t^{-1}(\tilde{\theta}_t)} \left\| \tilde{\theta}_t - \theta_\star \right\|_{\mathbf{H}_t(\tilde{\theta}_t)} + (\bar{L}_\mu/2) \|a\|_{\mathbf{H}_t^{-1}(\tilde{\theta}_t)}^2 \left\| \tilde{\theta}_t - \theta_\star \right\|_{\mathbf{H}_t(\tilde{\theta}_t)}^2 \end{aligned}$$

by the Cauchy-Schwarz inequality. By [Corollary 2.3.1](#) if  $\theta_\star \in \mathcal{C}_t(\delta)$  then  $\theta_\star \in \mathcal{C}'_t(\delta)$  and therefore:

$$\left\| \theta_\star - \tilde{\theta}_t \right\|_{\mathbf{H}_t(\theta)} \leq 2(1 + 2S)\gamma_t(\delta) .$$

Plugging this result in the prediction error bound we obtain:

$$\begin{aligned} \Delta_t(a) &\leq 2(1 + 2S)\gamma_t(\delta) \left[ \dot{\mu}(a^\top \tilde{\theta}_t) \|a\|_{\mathbf{H}_t^{-1}(\tilde{\theta}_t)} + (1 + 2S)\gamma_t(\delta) \bar{L}_\mu \|a\|_{\mathbf{H}_t^{-1}(\tilde{\theta}_t)}^2 \right] , \\ &\leq 2(1 + 2S)\gamma_t(\delta) \left[ \dot{\mu}(a^\top \tilde{\theta}_t) \|a\|_{\mathbf{H}_t^{-1}(\tilde{\theta}_t)} + (1 + 2S)\gamma_t(\delta) \bar{\kappa}_\mu \|a\|_{\mathbf{V}_t^{-1}}^2 \right] . \end{aligned}$$

where we last used  $\mathbf{H}_t(\theta) \geq \bar{\ell}_\mu \mathbf{V}_t^{-1}$  for  $\theta \in \Theta$  and  $\bar{\kappa}_\mu = \bar{L}_\mu / \bar{\ell}_\mu$ . ■

### Appendix 3.C Proof of Eq. (3.7)

By a simple Taylor-expansion:

$$\dot{\mu}(a_t^\top \tilde{\theta}_t) \leq \dot{\mu}(a_t^\top \bar{\theta}_t) + \bar{L}_\mu |a_t^\top (\tilde{\theta}_t - \bar{\theta}_t)| \quad (3.19)$$

$$\leq \dot{\mu}(a_t^\top \bar{\theta}_t) + \bar{L}_\mu \|a\|_{\mathbf{G}_t^{-1}(\tilde{\theta}_t, \bar{\theta}_t)} \|\tilde{\theta}_t - \bar{\theta}_t\|_{\mathbf{G}_t(\tilde{\theta}_t, \bar{\theta}_t)} \quad (\text{Cauchy-Schwarz})$$

$$\leq \dot{\mu}(a_t^\top \bar{\theta}_t) + \bar{L}_\mu \|a\|_{\mathbf{G}_t^{-1}(\tilde{\theta}_t, \bar{\theta}_t)} \|g_t(\tilde{\theta}_t) - g_t(\bar{\theta}_t)\|_{\mathbf{G}_t^{-1}(\tilde{\theta}_t, \bar{\theta}_t)} \quad (\text{Eq. (1.21)})$$

$$\leq \dot{\mu}(a_t^\top \bar{\theta}_t) + 2\sqrt{1 + 2S\bar{L}_\mu} \gamma_t(\delta) \|a\|_{\mathbf{G}_t^{-1}(\tilde{\theta}_t, \bar{\theta}_t)}$$

$$\leq \dot{\mu}(a_t^\top \bar{\theta}_t) + 2\sqrt{1 + 2S(\bar{L}_\mu/\sqrt{\bar{\ell}_\mu})} \gamma_t(\delta) \|a\|_{\mathbf{V}_t^{-1}} \quad (\text{Eq. (1.23)}) \quad (3.20)$$

In the second to last inequality we used that:

$$\begin{aligned} \|g_t(\tilde{\theta}_t) - g_t(\bar{\theta}_t)\|_{\mathbf{G}_t^{-1}(\tilde{\theta}_t, \bar{\theta}_t)} &\leq \|g_t(\bar{\theta}_t) - g_t(\hat{\theta}_t)\|_{\mathbf{G}_t^{-1}(\tilde{\theta}_t, \bar{\theta}_t)} + \|g_t(\bar{\theta}_t) - g_t(\hat{\theta}_t)\|_{\mathbf{G}_t^{-1}(\tilde{\theta}_t, \bar{\theta}_t)} \\ &\leq \sqrt{1 + 2S} \left( \|g_t(\bar{\theta}_t) - g_t(\hat{\theta}_t)\|_{\mathbf{H}_t^{-1}(\bar{\theta}_t)} + \|g_t(\tilde{\theta}_t) - g_t(\hat{\theta}_t)\|_{\mathbf{H}_t^{-1}(\bar{\theta}_t)} \right) \\ &\leq 2\sqrt{1 + 2S} \gamma_t(\delta) \end{aligned}$$

where we first used Eq. (1.28) and then the fact that  $\tilde{\theta}_t, \bar{\theta}_t \in \mathcal{C}_t(\delta)$ . Using Eq. (3.20) we therefore have that:

$$\begin{aligned} R_1(T) &\leq \sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t) \|a_t\|_{\mathbf{H}_t^{-1}(\bar{\theta}_t)} + 2\sqrt{1 + 2S(\bar{L}_\mu/\sqrt{\bar{\ell}_\mu})} \bar{\gamma}_T(\delta) \sum_{t=1}^T \|a_t\|_{\mathbf{H}_t^{-1}(\bar{\theta}_t)} \|a_t\|_{\mathbf{V}_t^{-1}} \\ &\leq \sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t) \|a_t\|_{\mathbf{H}_t^{-1}(\bar{\theta}_t)} + 2\sqrt{1 + 2S\bar{\kappa}_\mu} \bar{\gamma}_T(\delta) \sum_{t=1}^T \|a_t\|_{\mathbf{V}_t^{-1}}^2 \\ &= \sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t) \|a_t\|_{\mathbf{H}_t^{-1}(\bar{\theta}_t)} + 2\sqrt{1 + 2S\bar{\kappa}_\mu} \bar{\gamma}_T(\delta) R_2(T) \end{aligned}$$

which proves the claimed result.

### Appendix 3.D Proof of Lemma 3.1.2

Recall we are trying to bound  $\sum_{t=1}^T \sqrt{\dot{\mu}(a_t^\top \bar{\theta}_t)} \|\bar{a}_t\|_{\mathbf{L}_t^{-1}}$  where  $\bar{a}_t = \sqrt{\dot{\mu}(a_t^\top \bar{\theta}_t)} a_t$  and  $\mathbf{L}_t = \sum_{s=1}^{t-1} \bar{a}_s \bar{a}_s^\top + \lambda_t \mathbf{I}_d$ . Using the Cauchy-Schwarz inequality followed by a naive application of the Elliptical Potential lemma would yield the following bound:

$$\begin{aligned} \sum_{t=1}^T \sqrt{\dot{\mu}(a_t^\top \bar{\theta}_t)} \|\bar{a}_t\|_{\mathbf{L}_t^{-1}} &\leq \sqrt{\sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t)} \sqrt{\sum_{t=1}^T \|\bar{a}_t\|_{\mathbf{L}_t^{-1}}^2} \\ &\leq \sqrt{2d\bar{L}_\mu \log(\lambda_T + \bar{L}_\mu T/d)} \sqrt{\sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t)}, \end{aligned}$$

where the last inequality is obtained thanks to the Elliptical Potential lemma and the fact that all generality  $\|\bar{a}_t\| \leq \sqrt{\bar{L}_\mu}$ . When  $\bar{L}_\mu$  is large (for instance in the Poisson Bandit case) this bound is way off; the dependency in  $\bar{L}_\mu$  is an artifact of a loose analysis. This dependency is erased whenever one can guarantee that  $\|\bar{a}_t\|_{\mathbf{L}_t^{-1}} \leq 1$  for all  $t \geq 1$ . While this cannot happen at every

round (except if the regularization is set to  $\bar{L}_\mu$  which ultimately degrades the final regret bound) it holds as soon as the action  $a_t$  (or an “almost” co-linear action) is played once. This hints that the condition  $\|\bar{a}_t\|_{\mathbf{L}_t^{-1}} \leq 1$  holds for most rounds; we can just discard the rounds for which this does not hold in the analysis and bound the regret by its maximal value on such rounds. This analysis is also slightly off but allows to defer this annoying technicality to a second-order term in the regret, as formalized in [Lemma 3.1.2](#).

**Lemma 3.1.2.** *The following holds:*

$$\sum_{t=1}^T \sqrt{\dot{\mu}(a_t^\top \bar{\theta}_t)} \|\bar{a}_t\|_{\mathbf{L}_t^{-1}} \leq \sqrt{2d \log(\lambda_T + T/d)} \sqrt{\sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t)} + 2d\bar{L}_\mu^2 \log(\lambda_T + \bar{L}_\mu T/d).$$

*Proof.* Define the following set of rounds:

$$\mathcal{T} = \left\{ t \in [T], \|a_t\|_{\mathbf{L}_t^{-1}} \geq 1 \right\}.$$

By a straight-forward bound we have:

$$\begin{aligned} |\mathcal{T}| &\leq \sum_{t \in \mathcal{T}} \|\bar{a}_t\|_{\mathbf{L}_t^{-1}} && (\|\bar{a}_t\|_{\mathbf{L}_t^{-1}} \geq 1 \text{ for } t \in \mathcal{T}) \\ &\leq \sum_{t \in \mathcal{T}} \|\bar{a}_t\|_{\mathbf{L}_t^{-1}}^2 && (\|\bar{a}_t\|_{\mathbf{L}_t^{-1}} \geq 1 \text{ for } t \in \mathcal{T}) \\ &\leq \sum_{t=1}^T \|\bar{a}_t\|_{\mathbf{L}_t^{-1}}^2 && (\mathcal{T} \in [T]) \\ &\leq 2d\bar{L}_\mu \log(\lambda_T + \bar{L}_\mu T/d), \end{aligned}$$

where we last used the Elliptical Potential Lemma (see [Lemma B.3](#)). Furthermore:

$$\begin{aligned} \sum_{t=1}^T \sqrt{\dot{\mu}(a_t^\top \bar{\theta}_t)} \|\bar{a}_t\|_{\mathbf{L}_t^{-1}} \mathbf{1}(a_t \in \mathcal{T}) &\leq |\mathcal{T}| \bar{L}_\mu \\ &\leq 2d\bar{L}_\mu^2 \log(\lambda_T + \bar{L}_\mu T/d), \end{aligned} \tag{3.21}$$

by some simple upper-bounding leveraging the fact that  $\mathbf{L}_t \succeq \lambda_1 \mathbf{I}_d \succeq \mathbf{I}_d$  and  $\dot{\mu}(a_t^\top \bar{\theta}_t) \leq \bar{L}_\mu$ . On the other hand:

$$\begin{aligned} \sum_{t=1}^T \sqrt{\dot{\mu}(a_t^\top \bar{\theta}_t)} \|\bar{a}_t\|_{\mathbf{L}_t^{-1}} \mathbf{1}(a_t \notin \mathcal{T}) &= \sum_{t=1}^T \sqrt{\dot{\mu}(a_t^\top \bar{\theta}_t)} \|\bar{a}_t\|_{\mathbf{L}_t^{-1}} \mathbf{1}(\|a_t\|_{\mathbf{L}_t^{-1}} < 1) \\ &\leq \sqrt{\sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t) \mathbf{1}(\|a_t\|_{\mathbf{L}_t^{-1}} < 1)} \sqrt{\sum_{t=1}^T \|\bar{a}_t\|_{\mathbf{L}_t^{-1}}^2 \mathbf{1}(\|a_t\|_{\mathbf{L}_t^{-1}} < 1)} \\ &\leq \sqrt{\sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t)} \sqrt{\sum_{t=1}^T \|\bar{a}_t\|_{\mathbf{L}_t^{-1}}^2 \mathbf{1}(\|a_t\|_{\mathbf{L}_t^{-1}} < 1)} \\ &\leq \sqrt{2d \log(\lambda_T + T/d)} \sqrt{\sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t)} \end{aligned} \tag{3.22}$$

where the second inequality is obtained by Cauchy-Schwarz, the third by using  $\dot{\mu} > 0$  and the last by the use of the Elliptical Lemma with the condition  $\|a_t\|_{\mathbf{L}_t^{-1}} \leq 1$  for every round  $T$  (up

to straight-forward re-indexing). The announced result is obtained by assembling Eqs. (3.21) and (3.22);

$$\begin{aligned} \sum_{t=1}^T \sqrt{\dot{\mu}(a_t^\top \bar{\theta}_t)} \|\bar{a}_t\|_{\mathbf{L}_t^{-1}} &= \sum_{t=1}^T \sqrt{\dot{\mu}(a_t^\top \bar{\theta}_t)} \|\bar{a}_t\|_{\mathbf{L}_t^{-1}} \mathbf{1}(a_t \notin \mathcal{T}) + \sum_{t=1}^T \sqrt{\dot{\mu}(a_t^\top \bar{\theta}_t)} \|\bar{a}_t\|_{\mathbf{L}_t^{-1}} \mathbf{1}(a_t \in \mathcal{T}) \\ &\leq \sqrt{2d \log(\lambda_T + T/d)} \sqrt{\sum_{t=1}^T \dot{\mu}(a_t^\top \bar{\theta}_t)} + 2d \bar{L}_\mu^2 \log(\lambda_T + \bar{L}_\mu T/d). \end{aligned}$$

■

### Appendix 3.E Proof of Eq. (3.10)

By using successively a first-order Taylor expansion, Assumption 1.4.1, Eq. (1.19) and the definition of the regret the following set of inequalities hold:

$$\begin{aligned} \sum_{t=1}^T \dot{\mu}(a_t^\top \theta_\star) &\leq T \dot{\mu}(a_\star^\top \theta_\star) + \sum_{t=1}^T \left| \int_{v=0}^1 \dot{\mu}(a_\star^\top \theta_\star + v(a_t - a_\star)^\top \theta_\star) dv \right| |(a_\star - a_t)^\top \theta_\star| \\ &\leq T \dot{\mu}(a_\star^\top \theta_\star) + \sum_{t=1}^T \left[ \int_{v=0}^1 \dot{\mu}(a_\star^\top \theta_\star + v(a_t - a_\star)^\top \theta_\star) dv \right] (a_\star - a_t)^\top \theta_\star \quad (|\dot{\mu}| \leq \dot{\mu}) \\ &= T \dot{\mu}(a_\star^\top \theta_\star) + \sum_{t=1}^T \alpha(\theta_\star, a_\star, a_t) (a_\star - a_t)^\top \theta_\star \\ &= T \dot{\mu}(a_\star^\top \theta_\star) + \sum_{t=1}^T \mu(a_\star^\top \theta_\star) - \mu(a_t^\top \theta_\star) \\ &= T \dot{\mu}(a_\star^\top \theta_\star) + \text{Regret}_{\theta_\star}(T). \end{aligned}$$

### Appendix 3.F Proof of Eq. (3.12)

Recall that both  $\theta_t, \theta_\star \in \Theta$ . Under the event  $\{\theta_\star \in \mathcal{C}_t(\delta)\}$ :

$$\begin{aligned} \|\theta_t - \theta_\star\|_{\mathbf{H}_t(\theta_\star)} &\leq \sqrt{1 + 2S} \|\theta_t - \theta_\star\|_{\mathbf{G}_t(\theta_t, \theta_\star)} && \text{(Eq. (1.28))} \\ &= \sqrt{1 + 2S} \|g_t(\theta_t) - g_t(\theta_\star)\|_{\mathbf{G}_t^{-1}(\theta_t, \theta_\star)} && \text{(Eq. (1.21))} \\ &\leq \sqrt{1 + 2S} \left( \|g_t(\theta_t) - g_t(\hat{\theta}_t)\|_{\mathbf{G}_t^{-1}(\theta_t, \theta_\star)} + \|g_t(\theta_\star) - g_t(\hat{\theta}_t)\|_{\mathbf{G}_t^{-1}(\theta_t, \theta_\star)} \right) \\ &\leq (1 + 2S) \left( \|g_t(\theta_t) - g_t(\hat{\theta}_t)\|_{\mathbf{H}_t^{-1}(\theta_t)} + \|g_t(\theta_\star) - g_t(\hat{\theta}_t)\|_{\mathbf{H}_t^{-1}(\theta_\star)} \right) && \text{(Eq. (1.28))} \\ &\leq 2(1 + 2S) \gamma_t(\delta). \end{aligned}$$

### Appendix 3.G Proof of Theorem 3.3.1

**Theorem 3.3.1.** *Let  $\delta \in (0, 1]$ . On any LogB problem the regret of OFU-GLB satisfies with probability at least  $1 - \delta$ :*

$$\text{Regret}_{\theta_\star}(T) = \mathcal{O} \left( d \log(T) \sqrt{\dot{\mu}(a_\star^\top \theta_\star)} T + d^2 \log(T)^2 + S \mu(a_\star^\top \theta_\star) \sum_{t=1}^T \mathbf{1}(a_t \in \mathcal{A}_-) \right).$$

*Proof.* As usual we work under the event  $\{\theta_\star \in \mathcal{C}_t(\delta) \text{ for all } t \geq 1\}$  which holds with probability at least  $1 - \delta$ . We start by performing a second order Taylor expansion of the regret.

$$\begin{aligned} \text{Regret}_{\theta_\star}(T) &= \sum_{t=1}^T \mu(a_\star^\top \theta_\star) - \mu(a_t^\top \theta_\star) \\ &= \underbrace{\sum_{t=1}^T \dot{\mu}(a_t^\top \theta_\star) (a_\star - a_t)^\top \theta_\star}_{R'_1(T)} + \underbrace{\sum_{t=1}^T \tilde{\vartheta}_t \left( (a_\star - a_t)^\top \theta_\star \right)^2}_{R'_2(T)}. \end{aligned}$$

where we defined  $\tilde{\vartheta}_t = \int_{v=0}^1 (1-v) \ddot{\mu} \left( a_t^\top \theta_\star + v(a_\star - a_t)^\top \theta_\star \right) dv$ . We start by examining  $R'_1(T)$ ; because by optimism  $a_t^\top \theta_t \geq a_\star^\top \theta_\star$  we have:

$$R'_1(T) \leq \sum_{t=1}^T \dot{\mu}(a_t^\top \theta_\star) a_t^\top (\theta_t - \theta_\star)$$

which we already bounded in the proof of [Theorem 3.2.2](#). Re-using the same argument yields:

$$R'_1(T) \leq 2\sqrt{2}f(T) \sqrt{T \dot{\mu}(a_\star^\top \theta_\star) + \text{Regret}_{\theta_\star}(T)}$$

where  $f(T) = (1 + 2S)\bar{\gamma}_T(\delta) \sqrt{d \log(\lambda_T + T/(4d))}$  (we used  $\bar{L}_\mu \leq 1/4$  in the LogB). As before this yields an implicit second-order polynomial inequation on the regret. Solving it yields;

$$\text{Regret}_{\theta_\star}(T) \leq 4\sqrt{2}f(T) \sqrt{\dot{\mu}(a_\star^\top \theta_\star)T} + 16f(T)^2 + 2R'_2(T). \quad (3.23)$$

We can now turn our attention to bounding  $R'_2(T)$  to finish the proof. The following holds:

$$R_2(T) = \sum_{t=1}^T \tilde{\vartheta}_t \left\{ (a_\star - a_t)^\top \theta_\star \right\}^2 \mathbb{1}(a_t \in \mathcal{A}_-) + \sum_{t=1}^T \tilde{\vartheta}_t \left\{ (a_\star - a_t)^\top \theta_\star \right\}^2 \mathbb{1}(a_t \in \mathcal{A}_+), \quad (3.24)$$

with  $\mathcal{A}_+ = \mathcal{A} \setminus \mathcal{A}_-$ . We start by bounding the most-left term in the above inequality. Note that by self-concordance ( $|\ddot{\mu}| \leq \dot{\mu}$ ) of the logistic function we have  $\tilde{\vartheta}_t \leq \alpha(\theta_\star, a_\star, a_t)$  and therefore:

$$\begin{aligned} \sum_{t=1}^T \tilde{\vartheta}_t \left\{ (a_\star - a_t)^\top \theta_\star \right\}^2 \mathbb{1}(a_t \in \mathcal{A}_-) &\leq \sum_{t=1}^T \alpha(\theta_\star, a_\star, a_t) \left\{ (a_\star - a_t)^\top \theta_\star \right\}^2 \mathbb{1}(a_t \in \mathcal{A}_-) \\ &\leq 2S \sum_{t=1}^T \alpha(\theta_\star, a_\star, a_t) \left\{ (a_\star - a_t)^\top \theta_\star \right\} \mathbb{1}(a_t \in \mathcal{A}_-) \\ &= 2S \sum_{t=1}^T \left[ \mu(a_\star^\top \theta_\star) - \mu(a_t^\top \theta_\star) \right] \mathbb{1}(a_t \in \mathcal{A}_-) \\ &\leq 2S \mu(a_\star^\top \theta_\star) \sum_{t=1}^T \mathbb{1}(a_t \in \mathcal{A}_-) \end{aligned} \quad (3.25)$$

where we used  $\|\theta_\star\| \leq S$  and  $\|x\| \leq 1$  (for any  $x \in \mathcal{A}$ ) in the second-inequality and the mean-value theorem for the equality which follows. We turn to bounding the most r.h.s term in [Eq. \(3.24\)](#). We start with the case  $a_\star^\top \theta_\star \geq 0$ . We therefore look at the following definition for the detrimental arms:

$$\mathcal{A}_- = \left\{ a \in \mathcal{A} \mid a^\top \theta_\star \leq -1 \right\}.$$

Fix  $t$  and assume that  $a_t \in \mathcal{A}_+$ . Note that when  $a_t^\top \theta_\star \geq 0$  we inherit  $\tilde{\vartheta}_t \leq 0$  from the fact that  $\dot{\mu}(z) \leq 0$  for all  $z \geq 0$ . Using this fact ( $\dot{\mu} \leq 0$  on  $\mathbb{R}^+$ ) we can show that if  $a_t^\top \theta_\star \leq 0$ :

$$\begin{aligned} \tilde{\vartheta}_t &\leq \int_{v=0}^1 (1-v) \dot{\mu} \left( (1-v) a_t^\top \theta_\star \right) dv \\ &\leq \int_{v=0}^1 \dot{\mu} \left( (1-v) a_t^\top \theta_\star \right) dv && (\dot{\mu} \leq |\dot{\mu}| \leq \dot{\mu}) \\ &\leq \dot{\mu}(a_t^\top \theta_\star) \int_{v=0}^1 \exp(v |a_t^\top \theta_\star|) && (\text{Lemma 1.B.3}) \\ &\leq e^1 \dot{\mu}(a_t^\top \theta_\star) && (-1 \leq a_t^\top \theta_\star \leq 0) \end{aligned}$$

where in the last inequality we used  $a_t^\top \theta_\star \geq [-1, 0]$  since  $a_t \in \mathcal{A}_+$  and  $a_t^\top \theta_\star \leq 0$  by assumption. Packing this results together we showed that:

$$\begin{aligned} \tilde{\vartheta}_t \mathbb{1}(a_t \in \mathcal{A}_+) &\leq e^1 \dot{\mu}(a_t^\top \theta_\star) \mathbb{1}(a_t \in \mathcal{A}_+, a_t^\top \theta_\star \leq 0) + 0 \cdot \mathbb{1}(a_t \in \mathcal{A}_+, a_t^\top \theta_\star \geq 0) \\ &\leq e^1 \dot{\mu}(a_t^\top \theta_\star) \mathbb{1}(a_t \in \mathcal{A}_+) \\ &\leq e^1 \dot{\mu}(a_t^\top \theta_\star) \end{aligned}$$

Therefore we obtain:

$$\begin{aligned} \sum_{t=1}^T \tilde{\vartheta}_t \left\{ (a_\star - a_t)^\top \theta_\star \right\}^2 \mathbb{1}(a_t \in \mathcal{A}_+) &\leq e^1 \sum_{t=1}^T \dot{\mu}(a_t^\top \theta_\star) \left\{ (a_\star - a_t)^\top \theta_\star \right\}^2 \\ &\leq e^1 \sum_{t=1}^T \dot{\mu}(a_t^\top \theta_\star) \left\{ a_t^\top (\theta_t - \theta_\star) \right\}^2 && (\text{optimism}) \\ &\leq 4e^1 (1 + 2S)^2 \bar{\gamma}_T^2(\delta) \sum_{t=1}^T \dot{\mu}(a_t^\top \theta_\star) \|a_t\|_{\mathbf{H}_t^{-1}(\theta_\star)}^2 && (\text{Eq. (3.12)}) \\ &\leq 8e^1 f(T)^2 && (3.26) \end{aligned}$$

where we last used the Elliptical Potential lemma. We now consider the case  $a_\star^\top \theta_\star \leq 0$ . The definition of  $\mathcal{A}_-$  becomes:

$$\mathcal{A}_- = \left\{ a \mid \dot{\mu}(a^\top \theta_\star) \leq \dot{\mu}(a_\star^\top \theta_\star) / 2 \right\}.$$

Fix  $t$  and assume that  $a_t \in \mathcal{A}_+$ . Thanks to  $|\dot{\mu}| \leq \dot{\mu}$ :

$$\begin{aligned} \tilde{\vartheta}_t &\leq \alpha(\theta_\star, a_\star, a_t) \\ &\leq \dot{\mu}(a_\star^\top \theta_\star) && (a_t^\top \theta_\star \leq a_\star^\top \theta_\star \leq 0 \text{ and } \dot{\mu} \text{ increasing on } \mathbb{R}^-) \\ &\leq 2\dot{\mu}(a_t^\top \theta_\star) && (x \in \mathcal{A}_+) \end{aligned}$$

Therefore we obtain:

$$\begin{aligned} \sum_{t=1}^T \tilde{\vartheta}_t \left\{ (a_\star - a_t)^\top \theta_\star \right\}^2 \mathbb{1}(a_t \in \mathcal{A}_+) &\leq 2 \sum_{t=1}^T \dot{\mu}(a_t^\top \theta_\star) \left\{ (a_\star - a_t)^\top \theta_\star \right\}^2 \\ &\leq 2 \sum_{t=1}^T \dot{\mu}(a_t^\top \theta_\star) \left\{ a_t^\top (\theta_\star - \theta_t) \right\}^2 && (\text{optimism}) \\ &\leq 8(1 + 2S)^2 \bar{\gamma}_T^2(\delta) \sum_{t=1}^T \dot{\mu}(a_t^\top \theta_\star) \|a_t\|_{\mathbf{H}_t^{-1}(\theta_\star)}^2 && (\text{Eq. (3.12)}) \\ &\leq 16f(T)^2 && (3.27) \end{aligned}$$

Assembling Eq. (3.24)-(3.25)-(3.26)-(3.27) we obtain that:

$$R_2(T) \leq 22f(T)^2 + 2S\mu(a_\star^\top \theta_\star) \sum_{t=1}^T \mathbf{1}(a_t \in \mathcal{A}_-).$$

Merging this result with Eq. (3.23) yields the bound;

$$\text{Regret}_{\theta_\star}(T) \leq 4\sqrt{2}f(T)\sqrt{\hat{\mu}(a_\star^\top \theta_\star)T} + 60f(T)^2 + 2S\mu(a_\star^\top \theta_\star) \sum_{t=1}^T \mathbf{1}(a_t \in \mathcal{A}_-).$$

Taking the minimum between this bound and Theorem 3.2.2 complemented by the fact that  $f(T) = \mathcal{O}(d \log(T))$  yields the announced result. ■

## Appendix 3.H Proof of Proposition 3.3.1

**Proposition 3.3.1** (Length of transitory regime). *Let  $\{a_1, \dots, a_T\}$  be generated by OFU-GLB on a LogB problem with arm-set  $\mathcal{A}$ . The following holds with high probability:*

$$\mu(a_\star^\top \theta_\star) \sum_{t=1}^T \mathbf{1}(a_t \in \mathcal{A}_-) = \tilde{\mathcal{O}}(d^2 + dK) \quad \text{if } |\mathcal{A}_-| \leq K, \quad (3.13)$$

$$\mu(a_\star^\top \theta_\star) \sum_{t=1}^T \mathbf{1}(a_t \in \mathcal{A}_-) = \tilde{\mathcal{O}}(d^3) \quad \text{if } \mathcal{A} = \mathcal{B}_d(0, 1). \quad (3.14)$$

### Proof of Eq. (3.13)

*Proof.* As usual we work under the event  $\{\forall t \geq 1, \theta_\star \in \mathcal{C}_t(\delta)\}$  which holds with probability  $1 - \delta$ . We assume that there is a finite number  $K$  of detrimental arms. We will separate three cases to ease the analysis:

- (1)  $a_\star^\top \theta_\star \geq 0$ ,
  - (2)  $a_\star^\top \theta_\star \leq -1$ ,
  - (3)  $a_\star^\top \theta_\star \in [-1, 0]$ .
- (3.28)

(1)  $a_\star^\top \theta_\star \geq 0$ . In this configuration we have  $\mathcal{A}_- = \{a \in \mathcal{A}, a^\top \theta_\star \leq -1\}$ . This implies that detrimental arms have a large (constant) gap. Indeed for any  $a \in \mathcal{A}_-$ :

$$\mu(a_\star^\top \theta_\star) - \mu(a^\top \theta_\star) \geq \mu(a_\star^\top \theta_\star) - \mu(-1) \geq 1/2 - \mu(-1) \geq 1/5. \quad (3.29)$$

We can use this result to show that OFU-GLB plays detrimental arms only logarithmically often. Indeed, for any  $a \in \mathcal{A}_-$  let  $\tau_a$  be the last time-step when  $a$  is played, and  $N_a$  the number of time  $a$  was played over the whole horizon. Formally:

$$\tau_a = \max_t \{t \in [T] \mid a_t = a\} \quad \text{and} \quad N_a = \sum_{t=1}^T \mathbf{1}(a_t = a) = \sum_{t=1}^{\tau_a} \mathbf{1}(a_t = a).$$

Fix  $a \in \mathcal{A}_-$  and let  $\tau = \tau_a$  (*i.e.*  $a_\tau = a$ ). Thanks to Eq. (3.29) and the mean-value theorem:

$$\begin{aligned}
1/5 &\leq \mu(a_*^\top \theta_*) - \mu(a_\tau^\top \theta_*) \\
&\leq \mu(a_\tau^\top \theta_\tau) - \mu(a_\tau^\top \theta_*) && \text{(optimism)} \\
&\leq \alpha(a_\tau, \theta_\tau, \theta_*) a_\tau^\top (\theta_\tau - \theta_*) && \text{(mean-value theorem)} \\
&= \alpha(a_\tau, \theta_\tau, \theta_*) a_\tau^\top \mathbf{G}_\tau^{-1}(\theta_\tau, \theta_*) (g_\tau(\theta_\tau) - g_\tau(\theta_*)) && \text{(Eq. (1.21))} \\
&\leq \alpha(a_\tau, \theta_\tau, \theta_*) \|a_\tau\|_{\mathbf{G}_\tau^{-1}(\theta_\tau, \theta_*)} \|g_\tau(\theta_\tau) - g_\tau(\theta_*)\|_{\mathbf{G}_\tau^{-1}(\theta_\tau, \theta_*)} && \text{(Cauchy-Schwarz)} \\
&\leq 2\sqrt{1 + 2S}\gamma_\tau(\delta) \alpha(a_\tau, \theta_\tau, \theta_*) \|a_\tau\|_{\mathbf{G}_\tau^{-1}(\theta_\tau, \theta_*)} && (3.30)
\end{aligned}$$

where we last used  $\|g_t(\theta_t) - g_t(\theta_*)\|_{\mathbf{G}_t^{-1}(\theta_t, \theta_*)} \leq 2\sqrt{1 + 2S}\gamma_t(\delta)$  (see proof of Eq. (3.12) in Appendix 3.F). Note also that  $\mathbf{G}_\tau(\theta_\tau, \theta_*) \succeq N_a \alpha(a, \theta_\tau, \theta_*) a a^\top + \lambda_\tau \mathbf{I}_d$ . It is therefore easy to show (for instance using the Sherman-Morison formula) that  $\|a_\tau\|_{\mathbf{G}_\tau^{-1}(\theta_\tau, \theta_*)}^2 \leq (\alpha(a_\tau, \theta_\tau, \theta_*) N_a)^{-1}$ . We therefore finally obtain by injecting this into Eq. (3.30):

$$\begin{aligned}
N_a &\leq 100(1 + 2S)\gamma_\tau(\delta)^2 \alpha(a_\tau, \theta_\tau, \theta_*) \\
&\leq 25(1 + 2S)\gamma_\tau(\delta)^2 && (\alpha \leq \sup \dot{\mu} \leq 1/4)
\end{aligned}$$

Remember that this results holds for *any*  $a \in \mathcal{A}_-$ . Therefore using  $\mu \leq 1$ :

$$\mu(a_*^\top \theta_*) \sum_{t=1}^T \mathbb{1}(a_t \in \mathcal{A}_-) = \sum_{a \in \mathcal{A}_-} N_a \leq 25(1 + 2S) |\mathcal{A}_-| \bar{\gamma}_T(\delta)^2.$$

Using the fact that  $\bar{\gamma}_t(\delta) = \mathcal{O}(\sqrt{d \log(T)})$  we obtain the announced result:

$$\mu(a_*^\top \theta_*) \sum_{t=1}^T \mathbb{1}(a_t \in \mathcal{A}_-) = \tilde{\mathcal{O}}(d^2 + dK).$$

(2)  $a_*^\top \theta_* < -1$ . In this configuration one necessarily has  $a^\top \theta_* \leq -1$  and  $\mu(a^\top \theta_*) \leq \mu(a_*^\top \theta_*) \leq \mu(-1) \leq 1/2$  for any  $a \in \mathcal{A}$ . We start by characterizing the gap of detrimental arms which are now defined by  $\mathcal{A}_- = \{a \in \mathcal{A}, \dot{\mu}(a^\top \theta_*) \leq \dot{\mu}(a_*^\top \theta_*)/2\}$ . From  $\dot{\mu} = \mu(1 - \mu)$  we get that for any  $a \in \mathcal{A}_-$ :

$$\begin{aligned}
\mu(a^\top \theta_*) &\leq \frac{\mu(a_*^\top \theta_*)}{2} \frac{1 - \mu(a_*^\top \theta_*)}{1 - \mu(a^\top \theta_*)} \\
&\leq \mu(a_*^\top \theta_*)/2 && (\mu(a^\top \theta_*) \leq \mu(a_*^\top \theta_*))
\end{aligned}$$

and therefore for any  $a \in \mathcal{A}_-$ :

$$\mu(a_*^\top \theta_*) - \mu(a^\top \theta_*) \geq \mu(a_*^\top \theta_*)/2 \tag{3.31}$$

Note the difference with case (1) since here the gap is no longer lower-bounded by a universal constant. Fix  $a \in \mathcal{A}_-$  and let  $\tau = \tau_a$  (*i.e.*  $a_\tau = a$ ). Using the mean-value theorem we obtain:

$$\begin{aligned}
\mu(a_*^\top \theta_*)/2 &\leq \mu(a_*^\top \theta_*) - \mu(a_\tau^\top \theta_*) \\
&\leq \alpha(\theta_*, a_*, a_\tau) \theta_*^\top (a_* - a_\tau) \\
&\leq \alpha(\theta_*, a_*, a_\tau) a_\tau^\top (\theta_\tau - \theta_*) && \text{(optimism)} \\
&\leq \alpha(\theta_*, a_*, a_\tau) a_\tau^\top \mathbf{G}_\tau^{-1}(\theta_\tau, \theta_*) (g_\tau(\theta_\tau) - g_\tau(\theta_*)) && \text{(Eq. (1.21))} \\
&\leq \alpha(\theta_*, a_*, a_\tau) \|a_\tau\|_{\mathbf{G}_\tau^{-1}(\theta_\tau, \theta_*)} \|g_\tau(\theta_\tau) - g_\tau(\theta_*)\|_{\mathbf{G}_\tau^{-1}(\theta_\tau, \theta_*)} && \text{(Cauchy-Schwarz)} \\
&\leq 2\sqrt{1 + 2S}\gamma_\tau(\delta) \alpha(\theta_*, a_*, a_\tau) \|a_\tau\|_{\mathbf{G}_\tau^{-1}(\theta_\tau, \theta_*)} \\
&\leq 2\sqrt{1 + 2S}\gamma_\tau(\delta) \dot{\mu}(a_*^\top \theta_*) \|a_\tau\|_{\mathbf{G}_\tau^{-1}(\theta_\tau, \theta_*)} && (3.32)
\end{aligned}$$

where we used the fact that  $\dot{\mu}$  is increasing on  $[a_\tau^\top \theta_\star, a_\star^\top \theta_\star]$  which yields  $\alpha(\theta_\star, a_\star, a_\tau) \leq \dot{\mu}(a_\star^\top \theta_\star)$ . We now need to separate two cases:

(2.1)  $a^\top \theta_\tau \leq 0$ . Thanks to optimism (*i.e.*  $a^\top \theta_\tau \geq a_\star^\top \theta_\star$ ) and the monotonicity (increasing) of  $\dot{\mu}$  in  $\mathbb{R}^-$  we obtain that  $\dot{\mu}(a_\star^\top \theta_\star) \leq \dot{\mu}(a^\top \theta_\tau)$ . Further:

$$\begin{aligned} \|a_\tau\|_{\mathbf{G}_\tau^{-1}(\theta_\tau, \theta_\star)} &\leq \sqrt{1 + 2S} \|a_\tau\|_{\mathbf{H}_\tau^{-1}(\theta_\tau)} && \text{(Eq. (1.28))} \\ &\leq \sqrt{1 + 2S} (N_a \dot{\mu}(a^\top \theta_\tau))^{-1/2} && \text{(Sherman-Morison)} \\ &\leq \sqrt{1 + 2S} (N_a \dot{\mu}(a_\star^\top \theta_\star))^{-1/2} && \text{(3.33)} \end{aligned}$$

(2.1)  $a^\top \theta_\tau \geq 0$

$$\begin{aligned} \|a_\tau\|_{\mathbf{G}_\tau^{-1}(\theta_\tau, \theta_\star)} &\leq (N_a \alpha(a, \theta_\tau, \theta_\star))^{-1/2} && \text{(Sherman-Morison)} \\ &\leq N_a^{-1/2} \left( \frac{a^\top \theta_\tau - a^\top \theta_\star}{\mu(a^\top \theta_\tau) - \mu(a^\top \theta_\star)} \right)^{1/2} && \text{(mean-value theorem)} \\ &\leq N_a^{-1/2} \sqrt{2S} \left( \mu(a^\top \theta_\tau) - \mu(a^\top \theta_\star) \right)^{-1/2} && (\|a\| \leq 1, \theta_\tau, \theta_\star \in \Theta) \\ &\leq N_a^{-1/2} \sqrt{2S} \left( 1/2 - \mu(a^\top \theta_\star) \right)^{-1/2} && (a^\top \theta_\tau \geq 0 \Rightarrow \mu(a^\top \theta_\tau) \geq 1/2) \\ &\leq N_a^{-1/2} \sqrt{2S} \left( 1/2 - \mu(-1) \right)^{-1/2} && (a^\top \theta_\star \leq 0 \Rightarrow \mu(a^\top \theta_\star) \leq \mu(-1)) \\ &\leq 5N_a^{-1/2} \sqrt{2S} \\ &\leq 5(N_a \dot{\mu}(a_\star^\top \theta_\star))^{-1/2} \sqrt{2S} && (0 \leq \dot{\mu} \leq 1) \end{aligned} \tag{3.34}$$

Therefore combining Eqs. (3.33) and (3.34) we obtain that whichever we are in case (2.1) or (2.2) for any  $a \in \mathcal{A}_-$ :

$$\|a_\tau\|_{\mathbf{G}_\tau^{-1}(\theta_\tau, \theta_\star)} \leq 5\sqrt{1 + 2S} \left( N_a \dot{\mu}(a_\star^\top \theta_\star) \right)^{-1/2} \gamma_\tau(\delta)$$

Plugging this result in Eq. (3.32) we obtain that:

$$\mu(a_\star^\top \theta_\star)/2 \leq 10(1 + 2S)N_a^{-1/2} \left( \dot{\mu}(a_\star^\top \theta_\star) \right)^{1/2} \gamma_\tau(\delta)$$

Therefore for any  $a \in \mathcal{A}_-$  and since  $\dot{\mu} = \mu(1 - \mu) \leq \mu$ :

$$N_a \leq 400(1 + 2S)^2 \frac{\dot{\mu}(a_\star^\top \theta_\star)}{\mu(a_\star^\top \theta_\star)^2} \gamma_\tau(\delta)^2 \leq \frac{400(1 + 2S)^2}{\mu(a_\star^\top \theta_\star)} \gamma_\tau(\delta)^2 (\dot{\mu} \leq \mu) \tag{3.35}$$

Following the same reasoning as far case (1) we arrived to the same result;

$$\mu(a_\star^\top \theta_\star) \sum_{t=1}^T \mathbf{1}(a_t \in \mathcal{A}_-) = \sum_{a \in \mathcal{A}_-} N_a \leq 400(1 + 2S)^2 |\mathcal{A}_-| \bar{\gamma}_T(\delta)^2.$$

(3)  $a_\star^\top \theta_\star \in [-1, 0]$  In this configuration  $\mathcal{A}_- = \{a \in \mathcal{A}, \dot{\mu}(a^\top \theta_\star) \leq \dot{\mu}(a_\star^\top \theta_\star)/2\}$ . We can directly re-use the characterization of the sub-optimality gap for detrimental arms of Eq. (3.31). This yields that for any  $a \in \mathcal{A}_-$ :

$$\begin{aligned} \mu(a_\star^\top \theta_\star) - \mu(a^\top \theta_\star) &\geq \dot{\mu}(a_\star^\top \theta_\star)/2 \\ &\geq \dot{\mu}(-1)/2 \geq 9/200 \end{aligned}$$

We are therefore in the same configuration as in case (1) (the sub-optimality gap of detrimental arms is lower-bounded by a universal constant). Following the same reasoning yields to the announced claim. This finishes the proof. ■

**Proof of Eq. (3.14)**

*Proof.* As usual we work under the event  $\{\forall t \geq 1, \theta_* \in \mathcal{C}_t(\delta)\}$  which holds with probability  $1 - \delta$ . In the unit ball configuration  $a_*(\theta) = \theta/\|\theta\|$  for any  $\theta \in \Theta$ . Further, this guarantees that  $a_*(\theta)^\top \theta \geq 0$  for all  $\theta \in \Theta$ . In particular,  $a_*(\theta_*)^\top \theta_* \geq 0$  and we have the following definition for the detrimental arms:

$$\mathcal{A}_- = \left\{ a \in \mathcal{A}, a^\top \theta_* \leq -1 \right\}.$$

As before the objective of the proof is to bound the number of time detrimental arms are played by OFU-GLB within  $T$  rounds. We collect such rounds in the following set:

$$\mathcal{T} := \{t \leq T \text{ s.t. } a_t \in \mathcal{A}_-\}, \quad (3.36)$$

We are going to decompose the set  $\mathcal{T}$  in distinct subsets each one being of small cardinality. Formally, we construct  $\{\mathcal{T}_i\}_{i \geq 1}$  through the following backward induction.

1. **Initialization.**  $\mathcal{T}_0 = \emptyset, i = 0$ .

2. **Backward induction.** While  $\bigcup_{j \geq 1} \mathcal{T}_j \neq \mathcal{T}$ , we increment  $i$  by 1, and define

$$\begin{aligned} \tau_i &= \max \left\{ t \in \mathcal{T}, t \notin \bigcup_{j < i} \mathcal{T}_j \right\}, \\ \mathcal{T}_i &= \left\{ t \leq \tau_i, t \notin \bigcup_{j < i} \mathcal{T}_j, a_t^\top \theta_{\tau_i} \geq 0, a_t \in \mathcal{A}_- \right\}. \end{aligned} \quad (3.37)$$

Such construction immediately implies that  $\{\mathcal{T}_i\}_{i \geq 1}$  is a partition of  $\mathcal{T}$ .

**Proposition 3.H.1.** *Let  $\mathcal{T}$  and  $\{\mathcal{T}_i\}_{i \geq 1}$  be defined as in Eq. (3.36) and Eq. (3.37), and let  $N$  be the number of subsets  $\{\mathcal{T}_i\}$ . Then:*

$$\bigcup_{i=1}^N \mathcal{T}_i = \mathcal{T}; \quad \mathcal{T}_i \cap \mathcal{T}_j = \emptyset, \forall i \neq j; \quad N \leq (d+1).$$

*Proof of Proposition 3.H.1.* The fact that  $\bigcup_{i=1}^N \mathcal{T}_i$  is a partition of  $\mathcal{T}$  directly follows from its construction. Thus, we only have to prove that  $N \leq (d+1)$ . By construction, of the time steps  $\tau_i$  for  $i = 1, \dots, N$ , we have that

$$\forall j > i, \quad a_{\tau_i}^\top \theta_{\tau_j} < 0,$$

and since  $\theta_{\tau_i}$  is co-linear with  $a_{\tau_i}$ , we obtain

$$\forall j, i \in [N], \quad a_{\tau_i}^\top a_{\tau_j} < 0.$$

We conclude by using Lemma. 19 in Dong et al. (2019), which states that it can only exist at least  $d+1$  such arms, and hence such time steps. As a result  $N \leq (d+1)$ . ■

From the definition of  $\bigcup_{i=1}^N \mathcal{T}_i$  and Proposition 3.H.1, we have that

$$|\mathcal{T}| = \sum_{i=1}^N |\mathcal{T}_i| \leq (d+1) \max_{i=1, \dots, N} |\mathcal{T}_i|.$$

As a result, we only have to bound  $|\mathcal{T}_i|$  for any  $i \in [N]$  to conclude the proof. Notice that  $\tau_i$  is the last time step in  $\mathcal{T}_i$  and that for all  $t \in \mathcal{T}_i$ ,  $a_t^\top \theta_* \leq -1$  (from the definition of  $\mathcal{T}$ ) while  $a_t^\top \theta_{\tau_i} \geq 0$  (from the construction of the partition). Hence for all  $t \in \mathcal{T}_i$ :

$$\begin{aligned}
\mu(0) - \mu(-1) &\leq \mu(a_t^\top \theta_{\tau_i}) - \mu(a_t^\top \theta_\star) \\
&= \alpha(a_t, \theta_{\tau_i}, \theta_\star) a_t^\top (\theta_{\tau_i} - \theta_\star) && \text{(mean-value theorem)} \\
&\leq \alpha(a_t, \theta_{\tau_i}, \theta_\star) \|a_t\|_{\mathbf{G}_{\tau_i}^{-1}(\theta_{\tau_i}, \theta_\star)} \|\theta_{\tau_i} - \theta_\star\|_{\mathbf{G}_{\tau_i}(\theta_{\tau_i}, \theta_\star)} && \text{(Cauchy-Schwarz)} \\
&\leq 2\sqrt{1+2S}\gamma_{\tau_i}(\delta) \alpha(a_t, \theta_{\tau_i}, \theta_\star) \|a_t\|_{\mathbf{G}_{\tau_i}^{-1}(\theta_{\tau_i}, \theta_\star)} && (\theta_\star \in \mathcal{C}_t(\delta)) \\
&\leq \frac{\sqrt{1+2S}}{2} \gamma_{\tau_i}(\delta) \|a_t\|_{\mathbf{G}_{\tau_i}^{-1}(\theta_{\tau_i}, \theta_\star)}. && (\alpha \leq \sup \dot{\mu} \leq 1/4) \quad (3.38)
\end{aligned}$$

Further for all  $t \in \mathcal{T}_i$ ,  $a_t^\top \theta_\star \leq -1$  and  $a_t^\top \theta_{\tau_i} \geq 0$  leads to

$$\begin{aligned}
\alpha(a_t, \theta_{\tau_i}, \theta_\star) &= \frac{\mu(a_t^\top \theta_{\tau_i}) - \mu(a_t^\top \theta_\star)}{a_t^\top (\theta_{\tau_i} - \theta_\star)} \\
&\geq \frac{\mu(0) - \mu(-1)}{2S}. \quad (\|a\| \leq 1, \theta_{\tau_i}, \theta_\star \in \Theta)
\end{aligned}$$

As a result, let  $\bar{\mathbf{V}}_{\tau_i} := \sum_{s \in \mathcal{T}_i} a_s a_s^\top + \lambda_{\tau_i} \mathbf{I}_d$ , one obtains,

$$\mathbf{G}_{\tau_i}(\theta_{\tau_i}, \theta_\star) \succeq \sum_{s \in \mathcal{T}_i} \alpha(a_s, \theta_{\tau_i}, \theta_\star) a_s a_s^\top + \lambda_{\tau_i} \mathbf{I}_d \succeq \frac{\mu(0) - \mu(-1)}{2S} \bar{\mathbf{V}}_{\tau_i},$$

which combined with Eq. (3.38) leads to:

$$(\mu(0) - \mu(-1))^{3/2} \leq \sqrt{S/2} \sqrt{1+2S} \gamma_{\tau_i}(\delta) \|a_t\|_{\bar{\mathbf{V}}_{\tau_i}^{-1}}. \quad (3.39)$$

Taking the square and summing over  $t \in \mathcal{T}_i$  yields:

$$\begin{aligned}
(\mu(0) - \mu(-1))^3 |\mathcal{T}_i| &\leq (S/2)(1+2S)\gamma_{\tau_i}(\delta)^2 \sum_{t \in \mathcal{T}_i} \|a_t\|_{\bar{\mathbf{V}}_{\tau_i}^{-1}}^2 \\
&\leq (S/2)(1+2S)\gamma_{\tau_i}(\delta)^2 \text{Tr} \left( \bar{\mathbf{V}}_{\tau_i}^{-1} \sum_{t \in \mathcal{T}_i} a_t a_t^\top \right) \\
&\leq (S/2)(1+2S)\gamma_{\tau_i}(\delta)^2 d
\end{aligned}$$

and therefore  $|\mathcal{T}_i| \leq 50(1+2S)^2 d \gamma_{\tau_i}^2(\delta)$ . Therefore;

$$|\mathcal{T}| = \sum_{i=1}^N |\mathcal{T}_i| \leq 50(1+2S)^2 (d+1) d \bar{\gamma}_T(\delta)^2,$$

which finishes the proof since  $\bar{\gamma}_T(\delta) = \mathcal{O}(\sqrt{d \log(T)})$ . ■

### Appendix 3.I Proof of intermediary results for Theorem 3.4.1

Throughout the proof we actually assume for ease of exposition that  $\|\theta_\star\| \geq 1$  which implies that  $\kappa_\star(\theta_\star) \geq 5$ . It can be avoided and we make it here to simplify computations and avoid clutter. Note that  $\kappa(\theta_\star) \geq 5$  is precisely the region of interest for this lower-bound, *i.e.* large values of  $\kappa$ . In the same spirit also make the following assumption on  $\varepsilon$ :

$$\varepsilon \leq \|\theta_\star\| / \sqrt{d-1}. \quad (3.40)$$

which can trivially imposed when setting the final value of  $\varepsilon$ .

### 3.I.1 Proof of Proposition 3.4.1

**Proposition 3.4.1.** *For all  $\theta \in \mathbb{R}^d$  the following holds:*

$$\text{Regret}_\theta^\pi(T) \geq \frac{\|\theta\|}{\kappa_\star(\theta)} \sum_{i=1}^d \mathbb{E}_\theta \left[ \sum_{t=1}^T [a_\star(\theta) - a_t]_i^2 \right]. \quad (3.17)$$

Further if  $\|\theta\| \geq 1$ :

$$\text{Regret}_\theta^\pi(T) \geq \frac{1}{6} \mathbb{E}_\theta \left[ \sum_{t=1}^T \dot{\mu}(a_t^\top \theta) \|a_\star(\theta) - a_t\|^2 \right]. \quad (3.18)$$

*Proof.* We start by proving the second result. By definition of the regret:

$$\begin{aligned} \text{Regret}_\theta^\pi(T) &= \mathbb{E}_\theta \left[ \sum_{t=1}^T \mu(a_\star(\theta)^\top \theta) - \mu(a_t^\top \theta) \right] \\ &= \mathbb{E}_\theta \left[ \sum_{t=1}^T \alpha(\theta, a_\star(\theta), a_t) \left( a_\star(\theta)^\top \theta - a_t^\top \theta \right) \right] && \text{(mean-value theorem)} \\ &\geq \mathbb{E}_\theta \left[ \sum_{t=1}^T \frac{\dot{\mu}(a_t^\top \theta)}{1 + |\theta^\top (a_\star(\theta) - a_t)|} \left( a_\star(\theta)^\top \theta - a_t^\top \theta \right) \right] && \text{(Lemma 1.B.1)} \\ &\geq \frac{1}{1 + 2\|\theta\|} \mathbb{E}_\theta \left[ \sum_{t=1}^T \dot{\mu}(a_t^\top \theta) \left( a_\star(\theta)^\top \theta - a_t^\top \theta \right) \right] && (\|a\| \leq 1 \forall a \in \mathcal{A}) \\ &\geq \frac{\|\theta\|}{1 + 2\|\theta\|} \mathbb{E}_\theta \left[ \sum_{t=1}^T \dot{\mu}(a_t^\top \theta) \left( 1 - a_t^\top \frac{\theta}{\|\theta\|} \right) \right] \\ &\geq \frac{\|\theta\|}{2 + 4\|\theta\|} \mathbb{E}_\theta \left[ \sum_{t=1}^T \dot{\mu}(a_t^\top \theta) \|a_\star(\theta) - a_t\|^2 \right] \end{aligned}$$

where in the last line we used that for all  $x, y \in \mathcal{S}_d(0, 1)$  we have  $1 - x^\top y = \frac{1}{2} \|x - y\|^2$ . Using the fact that  $\|\theta\| \geq 1$  yields the second result. A similar bound can be written by using  $\alpha(\theta, a_\star(\theta), a_t) \geq \dot{\mu}(a_\star(\theta)^\top \theta)$ . Namely, we obtain:

$$\begin{aligned} \text{Regret}_\theta^\pi(T) &\geq \mathbb{E}_\theta \left[ \sum_{t=1}^T \dot{\mu}(a_\star(\theta)^\top \theta) \left( a_\star(\theta)^\top \theta - a_t^\top \theta \right) \right] \\ &\geq \frac{\|\theta\|}{\kappa_\star(\theta)} \mathbb{E}_\theta \left[ \sum_{t=1}^T \|a_\star(\theta) - a_t\|^2 \right] \\ &\geq \frac{\|\theta\|}{\kappa_\star(\theta)} \mathbb{E}_\theta \left[ \sum_{t=1}^T \|a_\star(\theta) - a_t\|^2 \right] && (\|\theta_\star\| \geq 1) \\ &\geq \frac{\|\theta\|}{\kappa_\star(\theta)} \mathbb{E}_\theta \left[ \sum_{t=1}^T \sum_{i=1}^d [a_\star(\theta) - a_t]_i^2 \right] \end{aligned}$$

Using the linearity of the expectation delivers the first claim. ■

### 3.I.2 Proof of Lemma 3.4.1

**Lemma 3.4.1.** *For each  $\theta \in \Xi_\varepsilon$  and any direction  $i \in [d, 2]$  let us introduce the event:*

$$A_i(\theta) := \left\{ [a_\star(\theta) - a_\star(\theta_\star)]_i \cdot \left[ \frac{1}{T} \sum_{t=1}^T a_t - a_\star(\theta_\star) \right]_i \geq 0 \right\}.$$

Then for any  $\theta \in \Xi_\varepsilon$  we have:

$$\text{Regret}_\theta^\pi(T) \geq \frac{T\epsilon^2}{2\kappa_\epsilon \|\theta_\star\|} \sum_{i=2}^d \mathbb{P}_\theta(A_i(\theta)).$$

*Proof.* From Proposition 3.4.1 we have that:

$$\begin{aligned} \text{Regret}_\theta^\pi(T) &\geq \frac{\|\theta\|}{\kappa(\theta)} \sum_{i=1}^d \mathbb{E}_\theta \left[ \sum_{t=1}^T [a_\star(\theta) - a_t]_i^2 \right] \\ &\geq \frac{\|\theta\|}{\kappa(\theta)} \sum_{i=1}^d \mathbb{E}_\theta \left[ \sum_{t=1}^T [a_\star(\theta) - a_t]_i^2 \mathbf{1} \{A_i(\theta)\} \right] \\ &= \frac{\|\theta\|}{\kappa(\theta)} \sum_{i=1}^d \mathbb{E}_\theta \left[ \sum_{t=1}^T [a_\star(\theta) - a_\star(\theta_\star) + a_\star(\theta_\star) - a_t]_i^2 \mathbf{1} \{A_i(\theta)\} \right] \\ &= \frac{\|\theta\|}{\kappa(\theta)} \sum_{i=1}^d [a_\star(\theta) - a_\star(\theta_\star)]_i^2 \mathbb{E}_\theta [\mathbf{1} \{A_i(\theta)\}] \\ &\quad + \frac{\|\theta\|}{\kappa(\theta)} \sum_{i=1}^d \mathbb{E}_\theta \left[ \sum_{t=1}^T [a_\star(\theta_\star) - a_t]_i^2 \mathbf{1} \{A_i(\theta)\} \right] \\ &\quad + \frac{2T\|\theta\|}{\kappa(\theta)} \sum_{i=1}^d \mathbb{E}_\theta \left[ \mathbf{1} \{A_i(\theta)\} \left[ a_\star(\theta_\star) - \frac{1}{T} \sum_{t=1}^T a_t \right]_i [a_\star(\theta) - a_\star(\theta_\star)]_i \right] \\ &\geq \frac{\|\theta\|}{\kappa(\theta)} T \sum_{i=1}^d [a_\star(\theta) - a_\star(\theta_\star)]_i^2 \mathbb{E}_\theta [\mathbf{1} \{A_i(\theta)\}] \end{aligned}$$

where in the last line we lower-bounded the last two terms by 0 (this was done for the second term thanks to the definition of  $A_i(\theta)$ ). Some easy computations yield the result:

$$\begin{aligned} \text{Regret}_\theta^\pi(T) &\geq T \frac{\|\theta\|}{\kappa_\epsilon} \frac{\epsilon^2}{\|\theta_\star\|^2 + (d-1)\epsilon^2} \sum_{i=2}^d \mathbb{E}_\theta [\mathbf{1} \{A_i(\theta)\}] \\ &\geq T \frac{\|\theta\|}{\kappa_\epsilon} \frac{\epsilon^2}{2\|\theta_\star\|^2} \sum_{i=2}^d \mathbb{E}_\theta [\mathbf{1} \{A_i(\theta)\}] \quad (\text{Equation (3.40)}) \\ &= \frac{T\epsilon^2}{2\kappa_\epsilon \|\theta_\star\|} \sum_{i=2}^d \mathbb{P}_\theta(A_i(\theta)) \end{aligned}$$

■

### 3.I.3 Proof of Lemma 3.4.2

**Lemma 3.4.2** (Averaging Hammer). *The following holds:*

$$\frac{1}{|\Xi_\varepsilon|} \sum_{\theta \in \Xi_\varepsilon} \sum_{i=2}^d \mathbb{P}_\theta(A_i(\theta)) \geq \frac{d}{4} - \frac{\sqrt{d}}{2} \sqrt{\frac{1}{|\Xi_\varepsilon|} \sum_{\theta \in \Xi_\varepsilon} \sum_{i=2}^d D_{KL}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)})}.$$

*Proof.* Let us fix  $\theta \in \Theta$  and  $i \in [2, d]$ . Note that:

$$\begin{aligned} \mathbb{P}_{\text{Flip}_i(\theta)}(A_i(\text{Flip}_i(\theta))) &\geq \mathbb{P}_\theta(A_i(\text{Flip}_i(\theta))) - D_{\text{TV}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)}) \\ &\geq \mathbb{P}_\theta(A_i(\text{Flip}_i(\theta))) - \sqrt{\frac{1}{2} D_{\text{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)})} \quad (\text{ Pinsker inequality}) \\ &\geq \mathbb{P}_\theta(A_i^C(\theta)) - \sqrt{\frac{1}{2} D_{\text{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)})} \end{aligned} \quad (3.41)$$

where  $D_{\text{KL}}$  denotes the relative entropy, and where we used the fact that:

$$\begin{aligned}
 A_i(\text{Flip}_i(\theta)) &= \left\{ [a_\star(\text{Flip}_i(\theta)) - a_\star(\theta_\star)]_i \cdot \left[ \frac{1}{T} \sum_{t=1}^T a_t - a_\star(\theta_\star) \right]_i \geq 0 \right\} && \text{(definition)} \\
 &= \left\{ [a_\star(\text{Flip}_i(\theta))]_i \cdot \left[ \frac{1}{T} \sum_{t=1}^T a_t \right]_i \geq 0 \right\} && (a_\star(\theta_\star)_i = 0) \\
 &= \left\{ -[a_\star(\theta)]_i \cdot \left[ \frac{1}{T} \sum_{t=1}^T a_t \right]_i \geq 0 \right\} && ([\text{Flip}_i(\theta)]_i = -[\theta]_i) \\
 &= A_i(\theta)^C
 \end{aligned}$$

In the following, we denote  $\Xi_i^+ := \{\theta \in \Xi_\varepsilon \text{ such that } \text{sign}([\theta]_i) > 0\}$  and  $\Xi_i^- := \{\theta \in \Xi_\varepsilon \text{ such that } \text{sign}([\theta]_i) < 0\}$ . Then by averaging over  $\Xi_\varepsilon$ :

$$\begin{aligned}
 \frac{1}{|\Xi_\varepsilon|} \sum_{\theta \in \Xi_\varepsilon} \sum_{i=2}^d \mathbb{P}_\theta(A_i(\theta)) &= \frac{1}{|\Xi_\varepsilon|} \sum_{i=2}^d \sum_{\theta \in \Xi_\varepsilon} \mathbb{P}_\theta(A_i(\theta)) \\
 &= \frac{1}{|\Xi_\varepsilon|} \sum_{i=2}^d \sum_{\theta \in \Xi_i^+} \left( \mathbb{P}_\theta(A_i(\theta)) + \mathbb{P}_{\text{Flip}_i(\theta)}(A_i(\text{Flip}_i(\theta))) \right) \\
 &\geq \frac{1}{|\Xi_\varepsilon|} \sum_{i=2}^d \sum_{\theta \in \Xi_i^+} \mathbb{P}_\theta(A_i(\theta)) + \mathbb{P}_\theta(A_i^C(\theta)) - \sqrt{\frac{1}{2} D_{\text{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)})} \quad \text{(Equation (3.41))} \\
 &\geq \frac{1}{|\Xi_\varepsilon|} \sum_{i=2}^d \sum_{\theta \in \Xi_i^+} 1 - \sqrt{\frac{1}{2} D_{\text{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)})}
 \end{aligned}$$

Repeating the same operation but referencing to  $\Xi_i^-$  we easily get that:

$$\begin{aligned}
 \frac{2}{|\Xi_\varepsilon|} \sum_{\theta \in \Xi_\varepsilon} \sum_{i=2}^d \mathbb{P}_\theta(A_i(\theta)) &\geq \frac{1}{|\Xi_\varepsilon|} \sum_{i=2}^d \sum_{\theta \in \Xi_i^+ \cup \Xi_i^-} 1 - \sqrt{\frac{1}{2} D_{\text{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)})} \\
 &= \frac{1}{|\Xi_\varepsilon|} \sum_{i=2}^d \sum_{\theta \in \Xi_\varepsilon} 1 - \sqrt{\frac{1}{2} D_{\text{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)})} \\
 &= (d-1) - \sum_{i=2}^d \frac{1}{|\Xi_\varepsilon|} \sum_{\theta \in \Xi_\varepsilon} \sqrt{\frac{1}{2} D_{\text{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)})} \\
 &\geq \frac{d}{2} - \sum_{i=2}^d \frac{1}{|\Xi_\varepsilon|} \sum_{\theta \in \Xi_\varepsilon} \sqrt{\frac{1}{2} D_{\text{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)})} \quad (d \geq 1) \\
 &\geq \frac{d}{2} - \frac{1}{\sqrt{2}} \sum_{i=2}^d \sqrt{\frac{1}{|\Xi_\varepsilon|} \sum_{\theta \in \Xi_\varepsilon} D_{\text{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)})} \quad \text{(Jensen inequality)} \\
 &\geq \frac{d}{2} - \sqrt{\frac{d-1}{2}} \sqrt{\sum_{i=2}^d \frac{1}{|\Xi_\varepsilon|} \sum_{\theta \in \Xi_\varepsilon} D_{\text{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)})} \quad \text{(Cauchy-Schwartz)} \\
 &\geq \frac{d}{2} - \sqrt{d} \sqrt{\sum_{i=2}^d \frac{1}{|\Xi_\varepsilon|} \sum_{\theta \in \Xi_\varepsilon} D_{\text{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)})}
 \end{aligned}$$

which proves the announced result. ■

### 3.I.4 Proof of Lemma 3.4.3

**Lemma 3.4.3** (Average Relative Entropy). *Under Hypothesis 3.4.1 we have:*

$$\frac{1}{|\Xi_\epsilon|} \sum_{\theta \in \Xi_\epsilon} \sum_{i=2}^d D_{KL} \left( \mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)} \right) \leq \frac{2}{\kappa_\epsilon} dT\epsilon^4 \exp(4\epsilon) + 4d\epsilon^2 \exp(4\epsilon) \left(6 + \frac{d}{2}\epsilon^2\right) \sqrt{\frac{T}{\kappa_\epsilon}}.$$

We will use the following result to control the relative entropy between two different parameters. It is a consequence of the relative entropy decomposition presented in [Lattimore and Szepesvári \(2020\)](#) along with the fact that the relative entropy is dominated by the chi-square divergence.

**Lemma 3.I.1** (Relative Entropy Decomposition). *For any  $\theta, \theta'$  we have that:*

$$D_{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \leq \mathbb{E}_\theta \left[ \sum_{t=1}^T \frac{\left( \mu(a_t^\top \theta) - \mu(a_t^\top \theta') \right)^2}{\dot{\mu}(a_t^\top \theta')} \right]$$

*Proof.* Denote  $P_a^\theta = \mathbb{P}_\theta(r|a)$ . Thanks to ([Lattimore and Szepesvári, 2020](#), Section 24.1) we have:

$$\begin{aligned} D_{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) &= \mathbb{E}_\theta \left[ \sum_{t=1}^T D_{KL} \left( P_{a_t}^\theta, P_{a_t}^{\theta'} \right) \right] \\ &= \mathbb{E}_\theta \left[ \sum_{t=1}^T D_{KL} \left( \text{Bernoulli}(a_t^\top \theta), \text{Bernoulli}(a_t^\top \theta') \right) \right] \\ &\leq \mathbb{E}_\theta \left[ \sum_{t=1}^T D_{\chi^2} \left( \text{Bernoulli}(a_t^\top \theta), \text{Bernoulli}(a_t^\top \theta') \right) \right] \end{aligned}$$

where we used  $D_{KL} \leq D_{\chi^2}$  ([Tsybakov, 2008](#), Chapter 2). Using the expression of the  $\chi^2$ -divergence for Bernoulli random variables finishes the proof.  $\blacksquare$

Applying this result between  $\mathbb{P}_\theta$  and  $\mathbb{P}_{\text{Flip}_i(\theta)}$  yields:

$$\begin{aligned} D_{KL}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)}) &\leq \mathbb{E}_\theta \left[ \sum_{t=1}^T \frac{\left( \mu(a_t^\top \theta) - \mu(a_t^\top \text{Flip}_i(\theta)) \right)^2}{\dot{\mu}(a_t^\top \text{Flip}_i(\theta))} \right] \\ &\leq \mathbb{E}_\theta \left[ \sum_{t=1}^T \frac{\alpha^2(a_t, \theta, \text{Flip}_i(\theta))}{\dot{\mu}(a_t^\top \text{Flip}_i(\theta))} \left\{ a_t^\top (\theta - \text{Flip}_i(\theta)) \right\}^2 \right] \quad (\text{mean-value theorem}) \end{aligned}$$

We are now going to link  $\alpha^2(a_t, \theta, \text{Flip}_i(\theta))$  to  $\dot{\mu}(a_t^\top \text{Flip}_i(\theta))$  and  $\dot{\mu}(a_t^\top \theta)$  thanks to the self-concordance. Indeed, it is easy to show (see the proof of Lemma 1.B.1) that for all  $z_1, z_2$  we have  $\dot{\mu}(z_1) \leq \dot{\mu}(z_2) \exp(|z_1 - z_2|)$ . We therefore have the following inequalities:

$$\begin{aligned} \alpha(a_t, \theta, \text{Flip}_i(\theta)) &\leq \dot{\mu}(a_t^\top \theta) \exp \left( \left| a_t^\top (\theta - \text{Flip}_i(\theta)) \right| \right) \quad \text{and} \\ \alpha(a_t, \theta, \text{Flip}_i(\theta)) &\leq \dot{\mu}(a_t^\top \text{Flip}_i(\theta)) \exp \left( \left| a_t^\top (\theta - \text{Flip}_i(\theta)) \right| \right) \end{aligned}$$

Plugging this in the relative entropy decomposition we obtain:

$$\begin{aligned}
 D_{\text{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)}) &\leq \mathbb{E}_\theta \left[ \sum_{t=1}^T \frac{\alpha^2(a_t, \theta, \text{Flip}_i(\theta))}{\dot{\mu}(a_t^\top \text{Flip}_i(\theta))} \left\{ a_t^\top (\theta - \text{Flip}_i(\theta)) \right\}^2 \right] \\
 &\leq \mathbb{E}_\theta \left[ \sum_{t=1}^T \dot{\mu}(a_t^\top \theta) \left\{ a_t^\top (\theta - \text{Flip}_i(\theta)) \right\}^2 \right] \exp \left( 2 \left| a_t^\top (\theta - \text{Flip}_i(\theta)) \right| \right) \\
 &\leq \exp(4\epsilon) \mathbb{E}_\theta \left[ \sum_{t=1}^T \dot{\mu}(a_t^\top \theta) \left\{ a_t^\top (\theta - \text{Flip}_i(\theta)) \right\}^2 \right] \\
 &\leq 2\epsilon^2 \exp(4\epsilon) \mathbb{E}_\theta \left[ \sum_{t=1}^T \dot{\mu}(a_t^\top \theta) [a_t]_i^2 \right] \\
 &= 2\epsilon^2 \exp(4\epsilon) \mathbb{E}_\theta \left[ \sum_{t=1}^T \dot{\mu}(a_t^\top \theta) [a_t - a_\star(\theta) + a_\star(\theta)]_i^2 \right] \\
 &\leq 4\epsilon^2 \exp(4\epsilon) \mathbb{E}_\theta \left[ \sum_{t=1}^T \dot{\mu}(a_t^\top \theta) [a_t - a_\star(\theta)]_i^2 + \sum_{t=1}^T \dot{\mu}(a_t^\top \theta) [a_\star(\theta)]_i^2 \right]
 \end{aligned}$$

where we last used the fact that  $(a + b)^2 \leq 2(a^2 + b^2)$ . Therefore by summing over  $d$ :

$$\begin{aligned}
 \sum_{d=2}^d D_{\text{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)}) &\leq 4\epsilon^2 \exp(4\epsilon) \mathbb{E}_\theta \left[ \sum_{t=1}^T \sum_{i=2}^d \dot{\mu}(a_t^\top \theta) [a_t - a_\star(\theta)]_i^2 + \sum_{t=1}^T \sum_{i=2}^d \dot{\mu}(a_t^\top \theta) [a_\star(\theta)]_i^2 \right] \\
 &\leq 4\epsilon^2 \exp(4\epsilon) \mathbb{E}_\theta \left[ \sum_{t=1}^T \sum_{i=1}^d \dot{\mu}(a_t^\top \theta) [a_t - a_\star(\theta)]_i^2 + \sum_{t=1}^T \sum_{i=1}^d \dot{\mu}(a_t^\top \theta) [a_\star(\theta)]_i^2 \right] \\
 &\leq 4\epsilon^2 \exp(4\epsilon) \mathbb{E}_\theta \left[ \sum_{t=1}^T \dot{\mu}(a_t^\top \theta) \|a_t - a_\star(\theta)\|^2 + d \frac{\epsilon^2}{\|\theta_\star\|^2 + (d-1)\epsilon^2} \sum_{t=1}^T \dot{\mu}(a_t^\top \theta) \right] \\
 &\leq 4\epsilon^2 \exp(4\epsilon) \mathbb{E}_\theta \left[ \sum_{t=1}^T \dot{\mu}(a_t^\top \theta) \|a_t - a_\star(\theta)\|^2 + \frac{d}{2} \epsilon^2 \sum_{t=1}^T \dot{\mu}(a_t^\top \theta) \right]
 \end{aligned}$$

where we used Equation (3.40) and the fact that  $\|\theta_\star\| \geq 1$ . Using Proposition 3.4.1 we obtain:

$$\sum_{d=2}^d D_{\text{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)}) \leq 4\epsilon^2 \exp(4\epsilon) \left( 6\text{Regret}_\theta^\pi(T) + \frac{d}{2} \epsilon^2 \mathbb{E}_\theta \left[ \sum_{t=1}^T \dot{\mu}(a_t^\top \theta) \right] \right) \quad (3.42)$$

We finish the proof by resorting to a Taylor expansion of  $\dot{\mu}(a_t^\top \theta)$ . Formally:

$$\sum_{t=1}^T \dot{\mu}(a_t^\top \theta) \leq \sum_{t=1}^T \left[ \dot{\mu}(a_\star(\theta)^\top \theta) + \left| \int_{v=0}^1 \ddot{\mu}(a_\star(\theta)^\top \theta + v\theta^\top (a_t - a_\star(\theta))) dv \right| \left| \theta^\top (a_\star(\theta) - a_t) \right| \right]$$

Using the fact that  $|\ddot{\mu}| \leq \mu$  and  $a_\star(\theta)^\top \theta \geq a_t^\top \theta$  we obtain that:

$$\begin{aligned}
 \mathbb{E}_\theta \left[ \sum_{t=1}^T \dot{\mu}(a_t^\top \theta) \right] &\leq \mathbb{E}_\theta \left[ \sum_{t=1}^T \left[ \dot{\mu}(a_\star(\theta)^\top \theta) + \alpha(\theta, a_\star(\theta), a_t) \theta^\top (a_\star(\theta) - a_t) \right] \right] \\
 &= \frac{T}{\kappa_\epsilon} + \mathbb{E}_\theta \left[ \sum_{t=1}^T \alpha(\theta, a_\star(\theta), a_t) \theta^\top (a_\star(\theta) - a_t) \right] \\
 &= \frac{T}{\kappa} + \text{Regret}_\theta^\pi(T)
 \end{aligned}$$

where we used the mean value theorem in the last line (see for instance the beginning of the proof of Proposition 3.4.1). Plugging this result in Equation (3.42) we obtain:

$$\sum_{d=2}^d D_{\text{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)}) \leq 4\epsilon^2 \exp(4\epsilon) \left( 6\text{Regret}_\theta^\pi(T) + \frac{d}{2}\epsilon^2 \left( \frac{T}{\kappa_\epsilon} + \text{Regret}_\theta^\pi(T) \right) \right)$$

Averaging over  $\Xi_\epsilon$  and since by Hypothesis (3.4.1) we know that  $\text{Regret}_\theta^\pi(T) \leq d\sqrt{T/\kappa_\epsilon}$  we obtain the announced result.

# Extensions to Non-Stationary Environments

---

The main goal of this chapter is to evaluate the portability of the tools we introduced so far for the more challenging non-stationary bandit problem. In this setting, ubiquitous in real-life scenarios, the ground truth  $\theta_*$  is allowed to change over time. This adds a *tracking* challenge on top of the traditional *learning* difficulties of the bandit problem. While MAB and LB algorithms have been successfully adapted to address this additional challenge, the non-stationary GLB case has been relatively under-explored. We first extend our approach to a family of non-stationary environments known as piece-wise stationary for which we show that the conclusion obtained in the stationary case still hold. We then turn our attention to a greater class of non-stationary environments characterized by a general notion of non-stationarity known as the variation-budget. In this setting we show that even extending the linearization approach of [Filippi et al. \(2010\)](#) is surprisingly challenging; we propose a first algorithm answering this challenge, but leave open the question of the optimal handling of non-linearity in this more challenging setting.

## Outline

---

4.1	The learning problem . . . . .	104
4.1.1	Setting and non-stationary metrics . . . . .	104
4.1.2	Forgetting strategies . . . . .	104
4.2	Piece-wise stationary environments . . . . .	105
4.2.1	Confidence sets . . . . .	105
4.2.2	Algorithms and regret upper-bounds . . . . .	107
4.2.3	Sketch of proof for the sliding-window strategy . . . . .	108
4.3	Drifting environments . . . . .	110
4.3.1	Motivation and Challenges . . . . .	110
4.3.2	A linearization algorithm . . . . .	112
4.3.3	Sketch of proof . . . . .	114

---

## 4.1 The learning problem

### 4.1.1 Setting and non-stationary metrics

**Setting.** In this chapter we consider the non-stationary Generalized Linear Bandit setting, which extends the stationary case laid out in [Section 1.3.1](#). The only difference comes from the nature of the ground truth generating the rewards. Indeed, we now assume that prior to the experiment the environment selects a *sequence* of parameters  $\{\theta_\star^t\}_{t=1}^T$ , such that the reward  $r_{t+1}$  generated at round  $t$  after the agent after plays  $a_t$  checks:

$$\mathbb{E}[r_{t+1}|\mathcal{F}_t] = \mu(a_t^\top \theta_\star^t), \quad (4.1)$$

$$\text{and } \text{Var}[r_{t+1}|\mathcal{F}_t] = \dot{\mu}(a_t^\top \theta_\star^t). \quad (4.2)$$

where  $\mu$  is strictly increasing and satisfies the self-concordant property (see [Assumption 1.4.1](#)). We shall work under the same boundedness assumptions as in the stationary setting and consider that [Assumptions 1.3.1](#) and [1.3.2](#) hold at each round  $t$ .

**Dynamic regret.** The focus in this chapter is on the *dynamic* regret:

$$\text{Regret}_{\theta_\star^1:T}^\pi(T) := \sum_{t=1}^T a_{\star,t}^\top \theta_\star^t - \sum_{t=1}^T a_t^\top \theta_\star^t,$$

where  $a_{\star,t} := \arg \max_{\mathcal{A}} \mu(a^\top \theta_\star^t)$  is the optimal action in-hindsight at each round  $t$ .

**Non-stationarity metrics.** The ground truth parameter  $\theta_\star^t$  is allowed to change in an *arbitrary* fashion; however, we assume that the agent has access to a non-stationarity metric which indicates the overall evolution of the environment over the horizon  $T$ . In this chapter we will cover two such metrics, each adapted to a particular nature of non-stationarity. The first is well fitted for piece-wise stationary environments which undergo brutal changes at  $o(T)$  rounds; this is measured by the quantity  $\Gamma_T$  which counts this number of “jumps” in the reward signal:

$$\Gamma_T = \sum_{t=2}^T \mathbb{1}(\theta_\star^t \neq \theta_\star^{t-1}).$$

The second fits *drifting* environments where the ground truth parameter can change at every round but by a small amount; this is measured by the *variation-budget*  $B_T$  defined as follows:

$$B_T := \sum_{t=2}^T \|\theta_\star^t - \theta_\star^{t-1}\|.$$

In each of the considered case, we will assume that the agent has the knowledge of the relevant non-stationary metric (or more realistically an upper-bound of it).

### 4.1.2 Forgetting strategies

A popular strategy to deal with non-stationarity is to resort to so-called *forgetting* strategies. The main concept is to forget old information as its relevance for taking decision in the present becomes questionable. In other words, such strategies *discard* the knowledge obtained by interacting with the  $\{\theta_\star^s\}$  for  $s \ll t$  as the associated data might not carry anymore a useful signal about the current version of the environment  $\theta_\star^t$ . This can be done in either smooth or abrupt fashions; in the non-stationary bandit literature, the two most popular forgetting mechanisms are *sliding-windows* (abrupt) and *discounts* (smooth).

**Piece-wise stationarity** Forgetting mechanisms problems were first studied by [Garivier and Moulines \(2011\)](#) who analyzed both the discount and sliding-window version of the UCB algorithm in piece-wise stationary MAB settings (the discounted UCB algorithm was introduced by [Kocsis and Szepesvári \(2006\)](#) but missed regret guarantees). Under a minimal gap assumption they proved that both algorithm enjoy a  $\tilde{\mathcal{O}}(\sqrt{\Gamma_T T})$  dynamic regret upper-bounds, matching the dynamic regret lower-bound presented in the same paper. Their approach was extended by [Russac et al. \(2020\)](#) to the general GLB setting by combining it with the linearization approach of [Filippi et al. \(2010\)](#). They show that both the sliding-window and discounted version of GLM-UCB enjoy dynamic regret upper-bounds of the form  $\tilde{\mathcal{O}}(\bar{\kappa}_\mu T^{2/3} \Gamma_T^{1/3})$ . The degradation of the bound compared to ([Garivier and Moulines, 2011](#)) comes from the fact that no gap assumption are made by [Russac et al. \(2020\)](#); actually 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 E of [Russac et al. \(2021\)](#)).

**Drifting environments.** The results obtained in the MAB piece-wise stationary setting were paralleled by similar achievements in drifting environments by [Besbes et al. \(2014\)](#) who achieved a  $\tilde{\mathcal{O}}(T^{2/3} B_T^{1/3})$  dynamic regret upper-bound and prove a matching dynamic regret lower-bound. There exists many attempts to extend this result to Linear Bandit environments; [Cheung et al. \(2019b\)](#) developed dynamic policies by resorting to a sliding-window, [Russac et al. \(2019\)](#) introduced a similar approach based on an exponential moving average, and [Zhao et al. \(2020\)](#) advocated for a simpler restart-based solution. All three aforementioned approaches claim regret bounds of the form  $\tilde{\mathcal{O}}(B_T^{1/3} T^{2/3})$  matching the results of [Besbes et al. \(2014\)](#) obtained in the MAB setting. Unfortunately, an error in their analysis was recently pointed out by [Touati and Vincent \(2021\)](#). A correct analysis yields degraded regret bounds, scaling as  $\tilde{\mathcal{O}}(B_T^{1/4} T^{3/4})$ . [Cheung et al. \(2019b\)](#); [Zhao et al. \(2020\)](#) also introduced linearization extensions of their LB-based algorithms to handle general GLB settings; unfortunately, their respective analyses again suffer from important caveats and overlooks the fact that as we shall see, the linearization approach of [Filippi et al. \(2010\)](#) requires important modifications in drifting environments.

## 4.2 Piece-wise stationary environments

We dedicate this section to extending our tools from the stationary case to the piece-wise stationary setting. Our goal is to mirror the improvements made over the linearization approach of [Filippi et al. \(2010\)](#) in the non-stationary case, this time over the algorithms of [Russac et al. \(2019\)](#). The first step in this direction is the derivation of the appropriate confidence sets, by leveraging our weighted concentration inequality from [Theorem 2.4.1](#).

### 4.2.1 Confidence sets

**Estimators.** We will consider here two forgetting mechanisms; the sliding-window and discounted strategies. Both rely on the same principle, which is to discard old information when estimating the current version  $\theta_\star^t$  of the environment. This can be formalized by introducing a general *weighted* quasi-maximum likelihood estimator:

$$\hat{\theta}_t^w := \arg \min_{\theta \in \Theta} \left\{ \mathcal{L}_t^w(\theta) := \sum_{s=1}^{t-1} w_{t-1-s} \left[ b(a_s^\top \theta) - r_{s+1} a_s^\top \theta \right] + \lambda \|\theta\|^2 / 2 \right\},$$

where  $\{w_t\}_{t=1}^T$  is a sequence of deterministic weights in the interval  $[0, 1]$ . As in the stationary case we will use the Hessian of the associated log-loss;

$$\mathbf{H}_t^w(\theta) := \sum_{s=1}^{t-1} \dot{\mu}(a_s^\top \theta) w_{t-1-s} a_s a_s^\top + \lambda_t \mathbf{I}_d .$$

as well as the notation  $g_t(\theta) = \sum_{s=1}^{t-1} w_{t-1-s} \mu(a_s^\top \theta) a_s + \lambda_t \theta$ . The sliding-window estimator simply discards data that was obtained more than  $D$  time steps ago, where  $D \in \mathbb{N}$  is an hyper-parameter dictating the length of the sliding-window. Formally, this estimator is obtained by setting  $w_{t-1-s} = \mathbf{1}(s \geq t-D)$ . The discount strategy follows a smoother approach and applies weights which gets smaller as the data gets older. Formally, for  $\gamma \in (0, 1)$  it is obtained by setting  $w_{t-1-s} = \gamma^{t-1-s}$ . In the rest of this chapter we replace the index “ $w$ ” in all notations by “SW” (resp. “ED”) whenever we refer to their sliding-window (resp. exponentially discounted) versions.

**Remark 4.2.1.** *The two strategies are obviously quite similar and undergo the same type of analysis. In particular for the discount strategy it is important to define an effective window-size  $D$  for the discounted strategy, which correspond to the rounds  $s \in [t-1-D, t-1]$  for which the weights  $\gamma^{t-1-s}$  are non-negligible. In the discounted approach this is a purely analytical quantity, defined by introducing a condition of the type:  $\gamma^D = 1/T^2$ , yielding  $D = 2 \log(T)/\log(1/\gamma)$ . In the following we willingly confuse it with the memory length of the sliding window estimator as both play the same conceptual role.*

**Confidence sets.** An important concept when studying forgetting mechanisms in piece-wise stationary settings is that there necessarily exists some continuous blocks of rounds on which the environment is stationary, and that are far enough from breakpoints so that the dominant terms in the estimators are all generated by the same ground truth parameter. On such blocks, we can repeat the analysis that was conducted in the stationary case. Formally, we are interested in the following set of rounds:

$$\mathcal{T}_D := \left\{ t \in [T], \theta_\star^t = \theta_\star^s \text{ for all } s \in [\max(1, t-D), t-1] \right\} .$$

Leveraging [Theorem 2.4.1](#) along with some simple upper-bounding allows to construct confidence sets for  $\theta_\star^t$  at each round  $t \in \mathcal{T}_D$ , as detailed in the following theorem. The proof is deferred to [Appendix 4.A](#). We will use the notation:

$$\nu_t(\delta) := \sqrt{\lambda_t(S + (2\sigma)^{-1})} + \frac{\sigma d}{\sqrt{\lambda_t}} \log\left(4T(1 + \sigma^2 D/(d\lambda_t))/\delta\right) .$$

Note that for  $\lambda_t = d \log(2+t)$  we still have  $\bar{\nu}_T(\delta) = \max_{t \in [T]} \nu_t(\delta) = \mathcal{O}(\sqrt{d \log(T/\delta)})$ .

**Theorem 4.2.1** (Variance-sensitive confidence sets based on forgetting estimators). *Let  $\delta \in (0, 1]$  and define the following sets:*

$$\begin{aligned} \mathcal{C}_t^{\text{SW}}(\delta) &:= \left\{ \theta \in \Theta, \left\| g_t^{\text{SW}}(\theta) - g_t^{\text{SW}}(\hat{\theta}_t^{\text{SW}}) \right\|_{\mathbf{H}_t^{\text{SW}}(\theta)^{-1}} \leq \nu_t(\delta) \right\} , \\ \mathcal{C}_t^{\text{ED}}(\delta) &:= \left\{ \theta \in \Theta, \left\| g_t^{\text{ED}}(\theta) - g_t^{\text{ED}}(\hat{\theta}_t^{\text{ED}}) \right\|_{\mathbf{H}_t^{\text{ED}}(\theta)^{-1}} \leq \nu_t(\delta) + 2(S\bar{L}_\mu + \sigma)\gamma^D/(1-\gamma) \right\} . \end{aligned}$$

The events  $\{\forall t \in \mathcal{T}_D, \theta_\star^t \in \mathcal{C}_t^{\text{SW}}(\delta)\}$  and  $\{\forall t \in \mathcal{T}_D, \theta_\star^t \in \mathcal{C}_t^{\text{ED}}(\delta)\}$  hold with probability at least  $1 - \delta$ .

**Algorithm 7** OFU-GLB-SW

---

**input:** Arm set  $\mathcal{A}$ , regularization coefficients  $\{\lambda_t\}_t$ , failure level  $\delta$ , admissible parameter set  $\Theta$ , sliding-window size  $D$ .  
Set  $\mathbf{H}_1 \leftarrow \lambda_1 \mathbf{I}_d$ ,  $\hat{\theta}_1^{\text{SW}} \leftarrow 0_d$ .  
**for**  $t \in [1, T]$  **do**  
    Solve  $a_t \in \arg \max_{\mathcal{A}} \max_{\theta \in \mathcal{C}_t^{\text{SW}}(\delta)} a^\top \theta$ .  $\triangleright$  *planning*  
    Play the arm  $a_t$  and observe reward  $r_{t+1}$ .  
    Update the estimator  $\hat{\theta}_{t+1}^{\text{SW}}$  and the confidence interval  $\mathcal{C}_t^{\text{SW}}(\delta)$ .  $\triangleright$  *learning*  
**end for**

---

As can be expected, both confidence sets have a similar structure. However because the discounted strategy keeps track of all past interactions (the weights are never truly zero, unlike for the sliding window strategy) the radius of the confidence set is augmented by an additive constant which depends on  $\bar{L}_\mu$ . This may feel like a miss; the goal behind such confidence sets is indeed to reduce the dependencies w.r.t the reward sensitivity constants in order to achieve improved tightness. Fortunately this additive term is truly a *negligible* second-order term. It indeed scales with  $\gamma^D$  which we saw in [Remark 4.2.1](#) is typically  $o(1)$  for the values of  $D$  that are chosen at analytical time.

### 4.2.2 Algorithms and regret upper-bounds

Equipped with the confidence sets from the previous section the design of the non-stationary GLB algorithm is straight-forward and re-employs the parameter-search optimistic approach of the previous chapter. Formally OFU-GLB-SW follows the strategy:

$$\text{play } a_t \in \arg \max_{a \in \mathcal{A}} \max_{\theta \in \mathcal{C}_t^{\text{SW}}(\delta)} a^\top \theta ,$$

while OFU-GLB-D uses the discount-based confidence set:

$$\text{play } a_t \in \arg \max_{a \in \mathcal{A}} \max_{\theta \in \mathcal{C}_t^{\text{ED}}(\delta)} a^\top \theta .$$

For the sake of completeness, we give the pseudo-code of OFU-GLB-SW in [Algorithm 7](#). The following theorem states the regret upper-bounds for those two algorithms and improves over the results presented in [Russac et al. \(2021\)](#). The proof of the first claim is deferred to the following section; the second is proven by directly replicating the bounding strategy. We use the following notation for the *averaged* reward sensitivity at the optimal action:

$$\ell_\mu^* := \sum_{t=1}^T \dot{\mu}(a_{\star,t}^\top \theta_\star^t) / T . \quad (4.3)$$

**Theorem 4.2.2** (Regret bounds for forgetting strategies). *Let  $\delta \in (0, 1]$ . When setting  $D = (T/\Gamma_T)^{2/3}$  (resp.  $\gamma = 1 - (T/\Gamma_T)^{2/3}$ ) the dynamic regret of OFU-GLB-SW (resp. OFU-GLB-D) satisfies with probability at least  $1 - \delta$ :*

$$\text{Regret}_{\theta_\star^1:T}(T) = \tilde{\mathcal{O}} \left( T^{2/3} \Gamma_T^{1/3} \left( d \log(T/\delta) \sqrt{\ell_\mu^* + \bar{L}_\mu} \right) + T^{1/3} \Gamma_T^{2/3} d^2 \log(T/\delta)^2 \left( 1 + \bar{L}_\mu / \ell_\mu + \bar{L}_\mu^2 \right) \right) .$$

A few comments are in order. The first-order term of the regret bounds presented in [Theorem 4.2.2](#) present a similar structure as in the stationary case; in the long-run, what matters is the reward sensitivity around the optimal action (here, its averaged version  $\ell_\mu^*$ ). In contrast

with the stationary case the maximal reward sensitivity  $\bar{L}_\mu$  also plays a role in this first-order term; as it will become clear in the analysis, this term is linked to an irreducible regret that the algorithm suffers after a switch of the reward signal. As in the stationary case, the effects of non-stationary embodied by the ratio  $\bar{L}_\mu/\ell_\mu$  are deferred to a second-order term of the regret. Note that the first-order term matches the rates of the regret bound of similar algorithms in the MAB setting, when no minimal gap assumption are made. Finally, those regret bounds are obtained under knowledge of  $\Gamma_T$ ; in practical case where  $\Gamma_T$  is unknown this can be alleviated by a Bandit-over-Bandit approach (see [Cheung et al. \(2019b\)](#)) with little cost on the regret.

**Remark** (About the bonus-based version). *The non-stationary setting further highlights the difference between the exploration-bonus and parameter-search approaches. Indeed we were not able to derive similar bounds for the bonus-based approach laid out in [Russac et al. \(2021\)](#); in the non-stationary setting one cannot guarantee that the set of information-preserving parameters (which is required in the stationary bonus version) is not empty. While this is most probably an analysis issue, it highlights the superiority from an analysis viewpoint of the parameter-search approach in non-linear parametric bandit problems.*

### 4.2.3 Sketch of proof for the sliding-window strategy

We provide here the sketch of proof for the regret upper-bound of OFU-GLB-SW. The proof starts by decomposing the dynamic regret over rounds in or out of  $\mathcal{T}_D$ .

$$\begin{aligned} \text{Regret}_{\theta_*^1:T}(T) &= \sum_{t=1}^T \mu(a_{*,t}^\top \theta_*^t) - \mu(a_t^\top \theta_*^t) \\ &= \underbrace{\sum_{t=1}^T [\mu(a_{*,t}^\top \theta_*^t) - \mu(a_t^\top \theta_*^t)] \mathbf{1}(t \in \mathcal{T}_D)}_{R_T} + \sum_{t=1}^T [\mu(a_{*,t}^\top \theta_*^t) - \mu(a_t^\top \theta_*^t)] \mathbf{1}(t \notin \mathcal{T}_D). \end{aligned}$$

Note that there are at most  $D\Gamma_T$  rounds outside of  $\mathcal{T}_D$ ; on such rounds the instantaneous regret can be maximal - at worst  $2S\bar{L}_\mu$ . In other words;

$$\sum_{t=1}^T [\mu(a_{*,t}^\top \theta_*^t) - \mu(a_t^\top \theta_*^t)] \mathbf{1}(t \notin \mathcal{T}_D) \leq 2S\bar{L}_\mu D\Gamma_T.$$

We now turn our attention to rounds in  $\mathcal{T}_D$ . We will work under the event  $\{\forall t \in \mathcal{T}_D, \theta_*^t \in \mathcal{C}_t^{\text{sw}(\delta)}\}$  which according to [Theorem 4.2.1](#) holds with probability at least  $1 - \delta$ . Denote  $\tilde{\theta}_t \in \arg \max_{\theta \in \mathcal{C}_t^{\text{sw}(\delta)}} a_t^\top \theta$ ; by the optimism property we have that  $a_t^\top \tilde{\theta}_t \geq a_{*,t}^\top \theta_*^t$  for all  $t \in \mathcal{T}_D$ . Applying this property followed by a second-order Taylor expansion we obtain:

$$\begin{aligned} R_T &= \sum_{t=1}^T [\mu(a_{*,t}^\top \theta_*^t) - \mu(a_t^\top \theta_*^t)] \mathbf{1}(t \in \mathcal{T}_D), \\ &\leq \sum_{t=1}^T [\mu(a_t^\top \tilde{\theta}_t) - \mu(a_t^\top \theta_*^t)] \mathbf{1}(t \in \mathcal{T}_D), \\ &\leq \sum_{t=1}^T \left[ \dot{\mu}(a_t^\top \theta_*^t) a_t^\top (\tilde{\theta}_t - \theta_*^t) + \bar{L}_\mu \left( a_t^\top (\tilde{\theta}_t - \theta_*^t) \right)^2 / 2 \right] \mathbf{1}(t \in \mathcal{T}_D). \end{aligned}$$

We prove in [Appendix 4.B](#) that under the event  $\{\forall t \in \mathcal{T}_D, \theta_*^t \in \mathcal{C}_t^{\text{sw}(\delta)}\}$  for all  $t \in \mathcal{T}_D$ :

$$\left\| \tilde{\theta}_t - \theta_*^t \right\|_{\mathbf{H}_t^{\text{sw}(\delta)}} \leq 2(1 + 2S)\nu_t(\delta). \quad (4.4)$$

After applying the Cauchy-Schwarz inequality to our bound on  $R_T$  we obtain:

$$\begin{aligned} R_T &\leq 2(1 + 2S)\nu_t(\delta) \sum_{t=1}^T \dot{\mu}(a_t^\top \theta_\star^t) \|a_t\|_{\mathbf{H}_t^{\text{SW}}(\theta_\star^t)^{-1}} \mathbf{1}(t \in \mathcal{T}_D) \\ &\quad + 2(1 + 2S)^2 \nu_t(\delta)^2 \bar{L}_\mu / \ell_\mu \sum_{t=1}^T \|a_t\|_{(\mathbf{V}_t^{\text{SW}})^{-1}}^2 \mathbf{1}(t \in \mathcal{T}_D), \end{aligned} \quad (4.5)$$

where  $\mathbf{V}_t^{\text{SW}} = \sum_{s=\max(t-D,1)}^{t-1} a_s a_s^\top + (\lambda_t / \ell_\mu) \mathbf{I}_d$ . The second term can be bounded thanks to a repeated application of the Elliptical Potential lemma on the right decomposition of  $\mathcal{T}_D$ . The proof of the following bound is purely technical so we defer it to [Appendix 4.C](#):

$$\sum_{t=1}^T \|a_t\|_{(\mathbf{V}_t^{\text{SW}})^{-1}}^2 \mathbf{1}(t \in \mathcal{T}_D) \leq 2\lceil T/D \rceil d \log(\lambda_T + T\bar{\ell}_\mu/d) \quad (4.6)$$

Similarly the first term in [Eq. \(4.5\)](#) is bounded by deriving the piece-wise stationary version of [Lemma 3.1.2](#); formally, it writes:

$$\begin{aligned} \sum_{t=1}^T \dot{\mu}(a_t^\top \theta_\star^t) \|a_t\|_{\mathbf{H}_t^{\text{SW}}(\theta_\star^t)^{-1}} \mathbf{1}(t \in \mathcal{T}_D) &\leq \sqrt{2\lceil T/D \rceil d \log(\lambda_T + T/d)} \sqrt{\sum_{t=1}^T \dot{\mu}(a_t^\top \theta_\star^t)} \\ &\quad + 2\lceil T/D \rceil d \bar{L}_\mu^2 \log(\lambda_T + \bar{L}_\mu T/d). \end{aligned} \quad (4.7)$$

The proof of this bound is given in [Appendix 4.D](#). Following the proof for the stationary case and using the self-concordance property it is easy to show that:

$$\begin{aligned} \sqrt{\sum_{t=1}^T \dot{\mu}(a_t^\top \theta_\star^t)} &\leq \sqrt{\sum_{t=1}^T \dot{\mu}(a_{\star,t}^\top \theta_\star^t)} + \sqrt{\text{Regret}_{\theta_\star^{1:t}}(T)} \\ &= \sqrt{T\ell_\mu^\star} + \sqrt{\text{Regret}_{\theta_\star^{1:t}}(T)}. \end{aligned} \quad (4.8)$$

with  $\ell_\mu^\star$  defined in [Eq. \(4.3\)](#). By denoting  $f(T) := (1+2S) \max(1, \nu_t(\delta)) \sqrt{d \log(\lambda_T + \max(1, \bar{L}_\mu)T/D)}$  and assembling [Eqs. \(4.5\)](#) to [\(4.8\)](#) we obtain after some simple-upper bounding:

$$R_T \leq 4f(T)^2 (\bar{L}_\mu / \ell_\mu + \bar{L}_\mu^2) \lceil T/D \rceil + 4\sqrt{\lceil T/D \rceil} f(T) (\sqrt{\ell_\mu^\star T} + \sqrt{\text{Regret}_{\theta_\star^{1:t}}(T)}).$$

Injecting into our initial bound on  $\text{Regret}_{\theta_\star^{1:t}}(T)$  we obtain:

$$\text{Regret}_{\theta_\star^{1:t}}(T) \leq 2S\bar{L}_\mu D \Gamma_T + 4f(T)^2 (\bar{L}_\mu / \ell_\mu + \bar{L}_\mu^2) \left\lceil \frac{T}{D} \right\rceil + 4\sqrt{\left\lceil \frac{T}{D} \right\rceil} f(T) (\sqrt{\ell_\mu^\star T} + \sqrt{\text{Regret}_{\theta_\star^{1:t}}(T)}).$$

Solving it leads to:

$$\text{Regret}_{\theta_\star^{1:t}}(T) \leq 8 \left\lceil \frac{T}{\sqrt{D}} \right\rceil f(T) \sqrt{\ell_\mu^\star} + 4S\bar{L}_\mu D \Gamma_T + 8 \left\lceil \frac{T}{D} \right\rceil f(T)^2 (1 + \bar{L}_\mu / \ell_\mu + \bar{L}_\mu^2).$$

We have left to tune the length  $D$  of the sliding-window. Choosing  $D = T^{2/3} \Gamma_T^{-2/3}$  yields:

$$\text{Regret}_{\theta_\star^{1:t}}(T) \leq T^{2/3} \Gamma_T^{1/3} (8f(T) \sqrt{\ell_\mu^\star} + 4S\bar{L}_\mu) + 8T^{1/3} \Gamma_T^{2/3} f(T)^2 (1 + \bar{L}_\mu / \ell_\mu + \bar{L}_\mu^2).$$

Realizing that  $f(T) = \mathcal{O}(d \log(T/\delta))$  yields the desired result.

## 4.3 Drifting environments

The treatment of drifting environments reveals to be more complex; as anticipated earlier and further detailed later in this section, even in the linear case current analyses show a gap with the known minimax-rates obtained in the MAB setting. In the GLB setting, things are even more blurred; as we shall see, even the linearization step introduced by [Filippi et al. \(2010\)](#) shows its limits and requires generalization.

For this reason we slightly switch gears in this section; we keep the discussion to a high-level to illustrate the difficulties and remaining challenges in this particular non-stationary setting. Furthermore, we shall put aside our on-going study of the effects of non-linearity; the prime focus is simply to obtain the first valid regret bound for drifting GLBs by generalizing the projection step of [Filippi et al. \(2010\)](#). The technical details, the different algorithms and the proof of their regret guarantees can be found by the interested reader in [Faury et al. \(2021\)](#).

### 4.3.1 Motivation and Challenges

**On the limits of piece-wise stationarity.** The piece-wise stationarity measure  $\Gamma_T$  is poorly suited to drifting environments as in such settings it 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 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 a larger spectrum of scenarios

**Existing work, linear case.** In the linear case, all approaches discussed hereinbefore follow this general path and announce regret rates of the form:

$$\text{Regret}_{\theta^*_{1:t}}(T) = \tilde{O}\left(B_T^{1/3}T^{2/3}\right).$$

It turns out that this rate can only hold with a relatively strong assumption on the geometry of the arm sets  $\mathcal{A}$  (at least with the existing analysis). In the general case, a correct analysis yields a degraded rate which does not match the lower-bound of [Cheung et al. \(2019b\)](#). 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  $\{\theta_s^*\}_{s=1}^{t-1}$  that generated observed rewards. The learning guarantees of forgetting policies can be extended from existing stationary analyses (e.g, [Abbasi-Yadkori et al., 2011](#)). This fundamentally requires an on-policy approach: deviations can only be measured in the directions that were played. Practically speaking, this means that the right metric to derive confidence intervals is  $\|\cdot\|_{\mathbf{V}_t}$  where  $\mathbf{V}_t = \sum_{s=1}^{t-1} w_{s,t} a_s a_s^\top + \lambda \mathbf{I}_d$ . On the other hand, the tracking aspect is inherited from the drift of  $\theta_{\star}^{t-1}$  to  $\theta_{\star}^t$  which induces an incompressible estimation error. It is therefore fundamentally tied to the variation-budget  $B_T$ , which is an off-policy metric (*i.e* independent of the trajectory that was played) characterized by the  $\ell_2$  norm. Both aspects are conflicting sources of regret; reaching optimality requires finding the correct balance between the two of them. Naturally, 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. If specializing to the sliding-window mechanism, the error can be traced back to the following statement:

$$\forall t \leq T, \quad \left\| \mathbf{V}_t^{-1} \sum_{s=t-D}^{t-1} a_s a_s^\top (\theta_\star^s - \theta_\star^t) \right\|_2 \leq \sum_{s=t-D}^{t-1} \|\theta_\star^s - \theta_\star^{s+1}\|, \quad (4.9)$$

which links the deviation between  $\theta_\star^t$  and  $\theta_\star^s$  in each direction  $a_s$  (on-policy) to the variation-budget over the length of the sliding window (off-policy). Such a statement appears several times in the literature, for instance in (Cheung et al., 2019b, Appendix B), (Russac et al., 2019, Appendix B.3) and (Zhao et al., 2020, Appendix A). Unfortunately, this is in general false. The approach followed by previous works ties the left-hand side of Equation (4.9) to  $\lambda_{\max}$ , the highest eigenvalue of the matrix  $\mathbf{V}_t^{-1} \sum_{s=t-D}^{t-1} a_s a_s^\top$ . They then proceed to show that the latter is smaller than a universal constant (one which does not depend on the dimension  $d$  or the sliding-window's length  $D$ ). The first step of this reasoning is false; indeed,  $\mathbf{V}_t^{-1} \sum_{s=t-D}^{t-1} a_s a_s^\top$  being not a symmetric matrix, its operator norm cannot be bounded by its larger eigenvalue; actually, one can easily design counter-examples where the two are arbitrarily different. This indicates that the impact of this mistake on the validity of the regret bound is significant; the matrix  $\mathbf{V}_t^{-1} \sum_{s=t-D}^{t-1} a_s a_s^\top$  being dependent of the algorithm's behavior, we cannot, in all generality, discard *a-priori* the events that such counter-examples arise. We can however look at *sufficient* conditions for the current analysis to hold. In particular, it is sufficient that  $\mathbf{V}_t^{-1} \sum_{s=t-D}^{t-1} a_s a_s^\top$  is a *symmetric* matrix. Equivalently, we can require for the two positive semidefinite matrices  $\mathbf{V}_t^{-1}$  and  $\sum_{s=t-D}^{t-1} a_s a_s^\top$  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  $\mathcal{A}$ ; for instance, when  $\mathcal{A}$  lies along an orthogonal basis.

**Proposition 4.3.1.** *Let  $\{e_i\}_{i=1}^d$  be an orthonormal basis of  $\mathbb{R}^d$  and  $\mathcal{A}$  be such that for all  $x \in \mathcal{A}$ , there exists  $\alpha \in \mathbb{R}$ ,  $i \in [1, d]$  such that  $x = \alpha e_i$ . Then on the non-stationary LB problem, forgetting strategies achieve a regret upper-bound of the form  $\text{Regret}_{\theta_\star^{1:t}}(T) = \tilde{\mathcal{O}}(B_T^{1/3} T^{2/3})$ .*

A correct treatment of the tracking error in the general case was recently proposed by (Touati and Vincent, 2021, Section 5). They showed that a correct bounding of the left-hand side of Equation (4.9) leads to the following control of the tracking error:

$$\forall t \leq T, \quad \left\| \mathbf{V}_t^{-1} \sum_{s=t-D}^{t-1} a_s a_s^\top (\theta_\star^s - \theta_\star^t) \right\|_2 \leq \sqrt{dD} \sum_{s=t-D}^{t-1} \|\theta_\star^s - \theta_\star^{s+1}\|.$$

The apparition of the sliding-window's length  $D$  in this bound eventually shifts its optimal value (in terms of learning v.s tracking regret balance) and yields degraded rates.

**Proposition 4.3.2** (Touati and Vincent (2021)). *Under general arm-set geometry, forgetting strategies achieve a regret upper-bound of the form  $\text{Regret}_{\theta_\star^{1:t}}(T) = \tilde{\mathcal{O}}(B_T^{1/4} T^{3/4})$  on the non-stationary LB problem.*

**Extension to GLBs.** (Cheung et al., 2019b; Zhao et al., 2020) extended their LB analysis to GLBs. With the exception of an inflation of the exploration bonus to account for non-linearity, their algorithm remains the same. They claim the following regret upper-bound:

$$\text{Regret}_{\theta_\star^{1:t}}(T) = \tilde{\mathcal{O}}\left(\bar{\kappa}_\mu B_T^{1/3} T^{2/3}\right),$$

which stands as a natural extension of the existing stationary bounds from Filippi et al. (2010). Not only does the previous remark regarding the validity of the rates (w.r.t  $T$  and  $B_T$ ) pass on to this setting, their approach also disregards the fundamental non-linear aspect of GLBs. 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  $\theta_\star^t$ ). From the assumption that  $\|\theta_\star^t\| \leq S$  for all  $t \in [T]$  this translates in the structural 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. 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  $\Theta$ . The situation is worse under non-stationarity since  $\hat{\theta}_t$  can be  $B_t$  far from  $\Theta$ . 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  $\bar{\kappa}_\mu$  which captures the degree of non-linearity of the inverse link function. For the highly non-linear logistic function we have seen that  $\bar{\kappa}_\mu \geq e^S$ . If we were to inflate the radius of the admissible set  $\Theta$  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  $R_\mu$  would be multiplied by a factor  $e^{\delta_S}$ ! Because the regret bound scales linearly with this quantity, this exponential growth would lead to prohibitively deficient performance guarantees.

**Challenges.** Filippi et al. (2010) countered the aforementioned difficulty by introducing a *projection* step, mapping  $\hat{\theta}_t$  back to an admissible parameter  $\tilde{\theta}_t \in \Theta$ . The latter is then used to predict the performance of the available actions. Their projection step essentially incorporates the prior knowledge  $\theta_\star \in \Theta$  without degrading the learning guarantees of the maximum likelihood estimator. The situation is different here as under parameter drift one needs to preserve both the learning and tracking guarantees of  $\hat{\theta}_t$ . However, and as previously discussed, both mechanisms have different dynamics and are characterized by different metrics. This leads to a tension in the design of the projection as this requires to incorporate the knowledge  $\{\theta_\star^t\} \in \Theta$ , without degrading neither the learning nor the tracking guarantees. This situation therefore calls for a generalization of the projection step of Filippi et al. (2010), in order to adapt to both sources of deviation.

### 4.3.2 A linearization algorithm

We provide in this section such a generalization. The main idea is to compute the optimal translation under the tracking metric of the confidence set characterized by the learning metric. We illustrate this idea in Figure 4.1.

**The algorithm.** The algorithmic idea to generalize the projection step of Filippi et al. (2010) is the same regardless of the forgetting mechanism that is chosen. We focused in the previous chapter on the sliding-window approach; we illustrate here the exponential-weight discount mechanism. Recall that our goal here is not to focus on the effects of non-linearity (embodied by the reward sensitivity constants) but rather on obtaining non-trivial regret bounds w.r.t  $d$  and  $T$  in the non-linear drifting setting. Therefore to simplify matter we will consider in the following a *bonus-based* approach. It operates in two steps; (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 at round  $t$ ) and (2) the construction of a suitable exploration bonus to compensate for prediction errors. The first step builds on the following set, linked to the

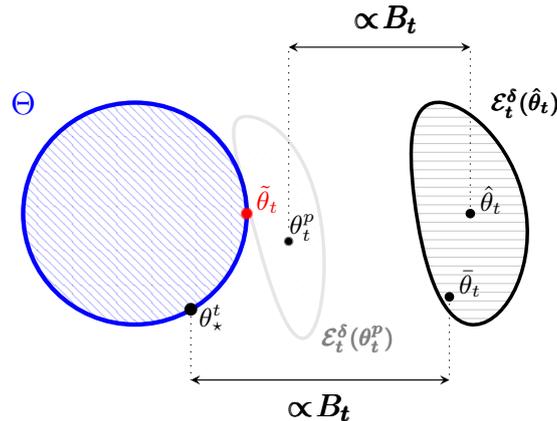


Figure 4.1: Illustration of the generalized projection step. The confidence set  $\mathcal{E}_t^\delta(\hat{\theta}_t)$  is obtained by leveraging the learning guarantees of the estimator  $\hat{\theta}_t$ . It concentrates around  $\hat{\theta}_t$ , which tracks the sequence  $\{\theta_\star^s\}_{s=1}^{t-1}$  and contrary to the stationary case can lie outside  $\Theta$ . Our solution involves finding a translation  $\theta_t^p$  of  $\hat{\theta}_t$  such that the confidence set centered at this new reference point effectively intersects  $\Theta$ , say at  $\tilde{\theta}_t$ . This translation occurs in a metric that preserves the tracking guarantees of  $\theta_t$ . The deviations  $(\theta_t^p \leftrightarrow \hat{\theta}_t)$  and  $(\theta_t \leftrightarrow \theta_\star^t)$  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.

deviation incurred through the learning process:

$$\mathcal{E}_t^\delta(\theta) := \left\{ \theta' \in \mathbb{R}^d \text{ s.t. } \left\| g_t^{\text{ED}}(\theta') - g_t^{\text{ED}}(\theta) \right\|_{(\mathbf{V}_t^{\text{ED}})^{-1}} \leq \beta_t(\delta) \right\}, \quad (4.10)$$

where  $\mathbf{V}_t^{\text{ED}} = \sum_{s=1}^{t-1} \gamma^{t-1-s} a_s a_s^\top + (\lambda/\bar{\ell}_\mu) \mathbf{I}_d$ <sup>1</sup> and  $\beta_t(\delta) = \sqrt{\lambda} S + \sigma \sqrt{2 \log(1/\delta) + d \log \left( 1 + \frac{\bar{\ell}_\mu (1-\gamma^{2t})}{\lambda d (1-\gamma^2)} \right)}$ .

We start by identifying an intermediary parameter  $\theta_t^p$ , solution of the following constrained optimization program (ties can be broken arbitrarily):

$$\theta_t^p \in \arg \min_{\theta \in \mathbb{R}^d} \left\{ \left\| g_t^{\text{ED}}(\theta) - g_t^{\text{ED}}(\hat{\theta}_t^{\text{ED}}) \right\|_{\mathbf{V}_t^{-2}} \text{ s.t. } \Theta \cap \mathcal{E}_t^\delta(\theta) \neq \emptyset \right\}. \quad (4.11)$$

This optimization program is well-posed as it consists in minimizing a smooth function over a non-empty compact set ( $\{\theta \text{ s.t. } \Theta \cap \mathcal{E}_t^\delta(\theta) \neq \emptyset\}$  always contains  $0_d$ ). Once  $\theta_t^p$  is computed, the algorithm simply chooses any parameter  $\tilde{\theta}_t \in \Theta \cap \mathcal{E}_t^\delta(\theta_t^p)$ .

**Remark 4.3.1.** 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  $\hat{\theta}_t^{\text{ED}}$ ) does not intersect the admissible set  $\Theta$ . Our strategy for finding  $\tilde{\theta}_t$  is then to compute an appropriate “vibration”  $\mathcal{E}_t^\delta(\theta_t^p)$  of  $\mathcal{E}_t^\delta(\hat{\theta}_t^{\text{ED}})$  which does intersect  $\Theta$ , while minimizing the deviation between  $\theta_t^p$  and  $\hat{\theta}_t^{\text{ED}}$  according to a metric related to the tracking error (through the map  $g_t^{\text{ED}}$  and the squared inverse of the design matrix). Note that the procedure to find  $\tilde{\theta}_t$  can be simplified; it is equivalent to directly solving:

$$\begin{pmatrix} \tilde{\theta}_t \\ \eta_t^p \end{pmatrix} \in \arg \min_{\theta' \in \mathbb{R}^d, \eta \in \mathbb{R}^d} \left\{ \left\| g_t^{\text{ED}}(\theta') + \beta_t(\delta) (\mathbf{V}_t^{\text{ED}})^{1/2} \eta - g_t^{\text{ED}}(\hat{\theta}_t^{\text{ED}}) \right\|_{\mathbf{V}_t^{-2}} \text{ s.t. } \|\theta'\| \leq S, \|\eta\| \leq 1 \right\}. \quad (4.12)$$

<sup>1</sup>Throughout this section we set  $\lambda_t = \lambda$  for all  $t$

The exploration bonus at round  $t$  for a given arm  $a \in \mathcal{A}$  is defined as  $\varepsilon_t(a) = 2\bar{\kappa}_\mu\beta_t(\delta)\|a\|_{(\mathbf{V}_t^{\text{ED}})^{-1}}$ . The algorithm follows an optimistic strategy by boosting the predicted reward associated with  $\tilde{\theta}_t$  by  $\varepsilon_t(a)$  and playing :

$$a_t \in \arg \max_{a \in \mathcal{A}_t} \mu(a^\top \tilde{\theta}_t) + \varepsilon_t(a).$$

**Regret guarantees.** As in the linear case, we can recover the minimax rates with sufficient assumptions on the arm-set geometry.

**Proposition 4.3.3.** *Let  $\{e_i\}_{i=1}^d$  be an orthonormal basis of  $\mathbb{R}^d$  and  $\mathcal{A}$  be such that for all  $a \in \mathcal{A}$ , there exists  $\alpha \in \mathbb{R}$ ,  $i \in [1, d]$  such that  $a = \alpha e_i$ . Then forgetting strategies achieve a regret upper-bound of the form  $\text{Regret}_{\theta_*^1:t}(T) = \tilde{\mathcal{O}}(\bar{\kappa}_\mu B_T^{1/3} T^{2/3})$  on the drifting GLB problem.*

For general arm-set geometry, sub-linear rates can still be recovered however with a sensibly more serious degradation than in the linear case. The culprit for this deterioration remains conceptually the same: the transfer from the on-policy to the off-policy tracking error comes at an additional cost due to the non-linearity of the reward function.

**Proposition 4.3.4.** *Under general arm-set geometry, forgetting strategies achieve a regret upper-bound of the form  $\text{Regret}_{\theta_*^1:t}(T) = \tilde{\mathcal{O}}(\bar{\kappa}_\mu B_T^{1/5} T^{4/5})$  on the non-stationary GLB problem.*

One will notice the presence in the bound of the ratio  $\bar{\kappa}_\mu$ , typical of the linearization approach. This regret bound is therefore quite natural and extends the work of [Filippi et al. \(2010\)](#) to drifting 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.

### 4.3.3 Sketch of proof

In this section, we detail the key steps to prove the aforementioned regret bounds. 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 will only consider here orthogonal arm-sets; 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  $\hat{\theta}_t^{\text{ED}}$  from  $\theta_*^t$ . 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 phenomena:

$$\bar{\theta}_t := \arg \min_{\theta \in \mathbb{R}^d} \left( \sum_{s=1}^{t-1} \gamma^{t-1-s} \left[ b(a_s^\top \theta) - \mu(a_s^\top \theta_*^s) a_s^\top \theta \right] + (\lambda/2) \left\| \theta - \theta_*^t \right\|^2 \right). \quad (4.13)$$

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<sup>2</sup>. As a result, the deviation between  $\hat{\theta}_t^{\text{ED}}$  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  $\theta_*^t$  is a consequence of the unpredictable changes of the sequence  $\{\theta_*^s\}_s$  (*tracking*). The introduction of the reference point  $\bar{\theta}_t$  allows us to characterize both deviations separately in [Lemma 4.3.1](#) and [Lemma 4.3.2](#).

<sup>2</sup>Note the difference between  $\hat{\theta}_t$  and  $\bar{\theta}_t$ , where the rewards  $r_{t+1}$  are replaced by their conditional expected values  $\mu(a_s^\top \theta_*^s)$

**Lemma 4.3.1.** *[Learning] Let  $\delta \in (0, 1]$ . With probability at least  $1 - \delta$ :*

$$\text{for all } t \geq 1, \quad \bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t^{\text{ED}}) = \left\{ \theta \in \mathbb{R}^d \text{ s.t. } \left\| g_t^{\text{ED}}(\theta) - g_t^{\text{ED}}(\hat{\theta}_t^{\text{ED}}) \right\|_{(\mathbf{V}_t^{\text{ED}})^{-1}} \leq \beta_t(\delta) \right\}.$$

Lemma 4.3.1 ensures that with high probability the set  $\mathcal{E}_t^\delta(\hat{\theta}_t^{\text{ED}})$  is a *confidence set* for  $\bar{\theta}_t$ .

**Lemma 4.3.2.** *[Tracking with orthogonal action sets] Let  $D \in \mathbb{N}^*$ . The following holds:*

$$\|g_t^{\text{ED}}(\bar{\theta}_t) - g_t^{\text{ED}}(\theta_\star^t)\|_{(\mathbf{V}_t^{\text{ED}})^{-2}} \leq \frac{2\bar{L}_\mu S}{\lambda} \frac{\gamma^D}{1-\gamma} + \bar{L}_\mu \sum_{s=t-D}^{t-1} \left\| \theta_\star^s - \theta_\star^{s+1} \right\|.$$

Lemma 4.3.2 effectively links the deviation of  $\bar{\theta}_t$  from  $\theta_\star^t$  to the variation-budget  $B_T$  through the drift  $\sum_{s=t-D}^{t-1} \|\theta_\star^s - \theta_\star^{s+1}\|$ . The proof of this result borrows tools from [Russac et al. \(2019\)](#). The integer  $D$  appearing in Lemma 4.3.2 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.

**Remark 4.3.2.** *Behind the statement of Lemma 4.3.1 and Lemma 4.3.2 hides the main reason why the projection step of [Filippi et al. \(2010\)](#) needs to be generalized. Indeed, it appears that the deviations  $(\hat{\theta}_t^{\text{ED}} \leftrightarrow \bar{\theta}_t)$  and  $(\bar{\theta}_t \leftrightarrow \theta_\star^t)$  are controlled through different metrics ( $(\mathbf{V}_t^{\text{ED}})^{-1}$  and  $(\mathbf{V}_t^{\text{ED}})^{-2}$ , 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  $\Delta_t$  defined as follows for any arm  $a \in \mathcal{A}$ :

$$\Delta_t(a) := \left| \mu(a^\top \tilde{\theta}_t) - \mu(a^\top \theta_\star^t) \right|.$$

The next Lemma ties the cumulative pseudo-regret to the sum of prediction errors.

**Lemma 4.3.3.** *The following holds:*

$$\text{Regret}_{\theta_\star^t}(T) \leq 2\bar{\kappa}_\mu \sum_{t=1}^T \beta_t(\delta) \left[ \|a_t\|_{(\mathbf{V}_t^{\text{ED}})^{-1}} - \|a_{\star,t}\|_{(\mathbf{V}_t^{\text{ED}})^{-1}} \right] + \sum_{t=1}^T [\Delta_t(a_t) + \Delta_t(a_{\star,t})].$$

Thanks to Lemma 4.3.3 we are left to characterize the prediction error  $\Delta_t(a)$  for any  $a \in \mathcal{A}$ . To do so we rely on the mean-value theorem to ensure that it exists  $\hat{\theta}_t \in [\tilde{\theta}_t, \theta_\star^t]$  such that<sup>3</sup>:

$$\Delta_t(a) \leq \bar{L}_\mu a_t^\top \mathbf{H}_t^{\text{ED}}(\hat{\theta}_t) \left( g_t^{\text{ED}}(\tilde{\theta}_t) - g_t^{\text{ED}}(\theta_\star^t) \right), \quad (4.14)$$

Since  $\tilde{\theta}_t, \theta_\star^t \in \Theta$ , we obtain  $\hat{\theta}_t \in \Theta$  and we can use the lower bound  $\mathbf{H}_t^{\text{ED}}(\hat{\theta}_t) \succeq \bar{\ell}_\mu \mathbf{V}_t^{\text{ED}}$ .

**Remark 4.3.3.** *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., 2019a](#); [Zhao et al., 2020](#)). Indeed, if the prediction error is measured at  $\hat{\theta}_t$ , we are left with  $\hat{\theta}_t \in [\theta_\star^t, \hat{\theta}_t^{\text{ED}}]$ , and  $\hat{\theta}_t$  can lie outside of the admissible set  $\Theta$  (since  $\hat{\theta}_t$  can). The lower-bound linking  $\mathbf{H}_t^{\text{ED}}(\hat{\theta}_t)$  and  $\mathbf{V}_t^{\text{ED}}$  would therefore not hold. More precisely when  $\hat{\theta}_t \in [\theta_\star^t, \hat{\theta}_t^{\text{ED}}]$  not much can be said on the link between  $\mathbf{H}_t^{\text{ED}}(\hat{\theta}_t)$  and  $\mathbf{V}_t^{\text{ED}}$  without severely degrading the final regret guarantees.*

<sup>3</sup>Formally,  $\hat{\theta}_t \in [\tilde{\theta}_t, \theta_\star^t]$  means that there exists  $v \in [0, 1]$  such that  $\hat{\theta}_t = v\tilde{\theta}_t + (1-v)\theta_\star^t$ .

Adding and removing  $g_t^{\text{ED}}(\hat{\theta}_t^{\text{ED}}) + g_t^{\text{ED}}(\theta_t^p) + g_t^{\text{ED}}(\bar{\theta}_t)$  inside the inner-product in Equation (4.14), followed by easy manipulations yields:

$$\begin{aligned} \Delta_t(a) &\leq \underbrace{\bar{\kappa}_\mu \|a\|_{(\mathbf{V}_t^{\text{ED}})^{-1}} \left( \|g_t^{\text{ED}}(\tilde{\theta}_t) - g_t^{\text{ED}}(\theta_t^p)\|_{(\mathbf{V}_t^{\text{ED}})^{-1}} + \|g_t^{\text{ED}}(\bar{\theta}_t) - g_t^{\text{ED}}(\hat{\theta}_t)\|_{(\mathbf{V}_t^{\text{ED}})^{-1}} \right)}_{:=\Delta_t^{\text{learn}}(a)} \\ &\quad + \underbrace{\bar{\kappa}_\mu \|a\| \left( \|g_t^{\text{ED}}(\theta_t^p) - g_t^{\text{ED}}(\hat{\theta}_t)\|_{(\mathbf{V}_t^{\text{ED}})^{-2}} + \|g_t^{\text{ED}}(\bar{\theta}_t) - g_t^{\text{ED}}(\theta_\star^t)\|_{(\mathbf{V}_t^{\text{ED}})^{-2}} \right)}_{:=\Delta_t^{\text{track}}(a)}. \end{aligned}$$

**Leveraging the projection step** We can now bound the terms  $\Delta_t^{\text{learn}}(x)$  and  $\Delta_t^{\text{track}}(x)$  separately. Lemma 4.3.1 along with the design  $\tilde{\theta}_t \in \mathcal{E}_t^\delta(\theta_t^p)$  leads to:

$$\Delta_t^{\text{learn}}(a) \leq 2\bar{\kappa}_\mu \|a\|_{(\mathbf{V}_t^{\text{ED}})^{-1}} \beta_t(\delta) \quad \text{w.h.p} \quad (4.15)$$

The first term in  $\Delta_t^{\text{track}}(a)$  is kept under control by the specific design of the projection step.

**Lemma 4.3.4.** *Under the event  $\{\bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t^{\text{ED}})\}$  the following holds:*

$$\|g_t^{\text{ED}}(\theta_t^p) - g_t^{\text{ED}}(\hat{\theta}_t)\|_{(\mathbf{V}_t^{\text{ED}})^{-2}} \leq \|g_t^{\text{ED}}(\bar{\theta}_t) - g_t^{\text{ED}}(\theta_\star^t)\|_{(\mathbf{V}_t^{\text{ED}})^{-2}}.$$

As a result, bounding  $\Delta_t^{\text{track}}(a)$  reduces to bounding  $\|g_t^{\text{ED}}(\bar{\theta}_t) - g_t^{\text{ED}}(\theta_\star^t)\|_{(\mathbf{V}_t^{\text{ED}})^{-2}}$ . Combined with Lemma 4.3.2, this result states that the deviation between  $\theta_t^p$  and  $\hat{\theta}_t$  is characterized by  $B_t$ , the parameter-drift up to round  $t$ , as illustrated in Figure 4.1. This leads to:

$$\Delta_t^{\text{track}}(a) \leq 2\bar{\kappa}_\mu \|a\|_2 \left( \frac{2\bar{L}_\mu L^2 S}{\lambda} \frac{\gamma^D}{1-\gamma} + \bar{L}_\mu \sum_{s=t-D}^{t-1} \|\theta_\star^s - \theta_\star^{s+1}\| \right) \quad \text{w.h.p} \quad (4.16)$$

**Putting everything together.** Combining Equations (4.15) and (4.16) with Lemma 4.3.3 and the Elliptical Lemma yields:

$$\text{Regret}_{\theta_\star^t:t}(T) \leq C_1 \bar{\kappa}_\mu d T \log(1/\gamma) + C_2 \bar{\kappa}_\mu \gamma^D T / (1-\gamma) + C_3 \bar{\kappa}_\mu D B_T \quad \text{w.h.p}$$

where the constants  $C_1$ ,  $C_2$  and  $C_3$  hide  $\log(T)$  multiplicative dependencies. Setting the hyper-parameters  $D = \log(T)/(1-\gamma)$  and  $\gamma = 1 - (\frac{B_T}{dT})^{2/3}$  concludes the proof.

# Appendix

## Appendix 4.A Proof of Theorem 4.2.1

**Theorem 4.2.1** (Variance-sensitive confidence sets based on forgetting estimators). *Let  $\delta \in (0, 1]$  and define the following sets:*

$$\begin{aligned} \mathcal{C}_t^{\text{SW}}(\delta) &:= \left\{ \theta \in \Theta, \left\| g_t^{\text{SW}}(\theta) - g_t^{\text{SW}}(\hat{\theta}_t^{\text{SW}}) \right\|_{\mathbf{H}_t^{\text{SW}}(\theta)^{-1}} \leq \nu_t(\delta) \right\}, \\ \mathcal{C}_t^{\text{ED}}(\delta) &:= \left\{ \theta \in \Theta, \left\| g_t^{\text{ED}}(\theta) - g_t^{\text{ED}}(\hat{\theta}_t^{\text{ED}}) \right\|_{\mathbf{H}_t^{\text{ED}}(\theta)^{-1}} \leq \nu_t(\delta) + 2(S\bar{L}_\mu + \sigma)\gamma^D/(1-\gamma) \right\}. \end{aligned}$$

The events  $\{\forall t \in \mathcal{T}_D, \theta_\star^t \in \mathcal{C}_t^{\text{SW}}(\delta)\}$  and  $\{\forall t \in \mathcal{T}_D, \theta_\star^t \in \mathcal{C}_t^{\text{ED}}(\delta)\}$  hold with probability at least  $1 - \delta$ .

*Proof.* The sliding-window result is actually a direct corollary of [Theorem 2.2.1](#). Indeed for any  $t \in \mathcal{T}_D$  we know that  $\theta_\star^s = \theta_\star^t$  for any  $s \in [t - D, t]$ . Fix  $t \in \mathcal{T}_D$  and denote for simplicity  $\theta_\star^t = \theta_\star$ ; by characterization of the weighted maximum-likelihood estimator:

$$\begin{aligned} \left\| g_t^{\text{SW}}(\theta_\star^t) - g_t^{\text{SW}}(\hat{\theta}_t^{\text{SW}}) \right\|_{\mathbf{H}_t^{\text{SW}}(\theta_\star^t)^{-1}} &\leq \left\| \sum_{s=t-D}^{t-1} [\mu(a_s^\top \theta_\star^t) - r_{s+1}] a_s \right\|_{\mathbf{H}_t^{\text{SW}}(\theta_\star^t)^{-1}} + \lambda \left\| \theta_\star^t \right\|_{\mathbf{H}_t^{\text{SW}}(\theta_\star^t)^{-1}} \\ &\leq \left\| \sum_{s=t-D}^{t-1} [\mu(a_s^\top \theta_\star^t) - \mu(a_s^\top \theta_\star^s)] a_s + \sum_{s=t-D}^{t-1} \eta_{s+1} a_s \right\|_{\mathbf{H}_t^{\text{SW}}(\theta_\star^t)^{-1}} + \sqrt{\lambda_t} S \\ &= \left\| \sum_{s=t-D}^{t-1} \eta_{s+1} a_s \right\|_{\mathbf{H}_t^{\text{SW}}(\theta_\star^t)^{-1}} + \sqrt{\lambda_t} S, \end{aligned}$$

where we defined  $\eta_{s+1} = \mu(a_s^\top \theta_\star^t) - r_{s+1} = \mu(a_s^\top \theta_\star) - r_{s+1}$ . Thanks to [Eqs. \(4.1\) and \(4.2\)](#) we have that  $\mathbb{E}[\eta_{s+1} | \mathcal{F}_s] = 0$  and  $\text{Var}(\eta_{s+1} | \mathcal{F}_s) = \dot{\mu}(a_s^\top \theta_\star)$ . Furthermore we have that:

$$\begin{aligned} \mathbf{H}_t^{\text{SW}}(\theta_\star^t) &= \sum_{s=t-D}^{t-1} \dot{\mu}(a_s^\top \theta_\star^t) a_s a_s^\top + \lambda_t \mathbf{I}_d \\ &= \sum_{s=t-D}^{t-1} \dot{\mu}(a_s^\top \theta_\star) a_s a_s^\top + \lambda_t \mathbf{I}_d \end{aligned}$$

Therefore we can directly apply the same concentration inequality we used in the stationary case (see [Theorem 2.2.1](#)) to obtain after some simple upper-bounding that with probability at least  $1 - \delta$ :

$$\left\| g_t^{\text{SW}}(\theta_\star^t) - g_t^{\text{SW}}(\hat{\theta}_t^{\text{SW}}) \right\|_{\mathbf{H}_t^{\text{SW}}(\theta_\star^t)^{-1}} \leq \sqrt{\lambda_t} (S + (2\sigma)^{-1}) + \frac{\sigma d}{\sqrt{\lambda_t}} \log\left(4(1 + \sigma^2 D / (d\lambda_t)) / \delta\right)$$

Applying an union bound over  $\mathcal{T}_D$  and using  $|\mathcal{T}_D| \leq T$  yields the announced result for the sliding-window confidence set. We now turn our attention to the confidence set based on the discounted estimator, which requires at little more work. Again, we fix  $t \in \tau$  and denote  $\theta_\star^t = \theta_\star = \theta_\star^s$  for every  $s \in [t - D, t]$ . By first using the characterization of the weighted maximum-likelihood

estimator we have the following set of inequalities:

$$\begin{aligned}
\left\| g_t^{\text{ED}}(\theta_\star^t) - g_t^{\text{ED}}(\hat{\theta}_t^{\text{SW}}) \right\|_{\mathbf{H}_t^{\text{ED}}(\theta_\star^t)^{-1}} &= \left\| \sum_{s=1}^{t-1} \gamma^{t-1-s} \left[ \mu(a_s^\top \theta_\star^t) - r_{s+1} \right] a_s + \lambda_t \theta_\star^t \right\|_{\mathbf{H}_t^{\text{ED}}(\theta_\star^t)^{-1}} \\
&= \left\| \sum_{s=1}^{t-1} \gamma^{t-1-s} \left[ \mu(a_s^\top \theta_\star^t) - \mu(a_s^\top \theta_\star^s) - \eta_{s+1} \right] a_s + \lambda_t \theta_\star^t \right\|_{\mathbf{H}_t^{\text{ED}}(\theta_\star^t)^{-1}} \\
&\leq \left\| \sum_{s=1}^{t-1} \gamma^{t-1-s} \left[ \mu(a_s^\top \theta_\star^t) - \mu(a_s^\top \theta_\star^s) \right] a_s \right\|_{\mathbf{H}_t^{\text{ED}}(\theta_\star^t)^{-1}} + \\
&\quad \left\| \sum_{s=1}^{t-1} \gamma^{t-1-s} \eta_{s+1} a_s \right\|_{\mathbf{H}_t^{\text{ED}}(\theta_\star^t)^{-1}} + \sqrt{\lambda_t} S \quad (4.17)
\end{aligned}$$

Let us start by bounding the first term; the idea is that since  $\theta_\star^s = \theta_\star^t = \theta_\star$  for every  $s \in [t-D, t]$ , only the  $t-D-1$  first terms of the sum matters and they are all multiplied by a very small constant (at least  $\gamma^{t-D}$ ). This justifies a rather crude bound;

$$\begin{aligned}
\left\| \sum_{s=1}^{t-1} \gamma^{t-1-s} \left[ \mu(a_s^\top \theta_\star^t) - \mu(a_s^\top \theta_\star^s) \right] a_s \right\|_{\mathbf{H}_t^{\text{ED}}(\theta_\star^t)^{-1}} &= \left\| \sum_{s=1}^{t-1-D} \gamma^{t-1-s} \left[ \mu(a_s^\top \theta_\star^t) - \mu(a_s^\top \theta_\star^s) \right] a_s \right\|_{\mathbf{H}_t^{\text{ED}}(\theta_\star^t)^{-1}} \\
&\leq \sum_{s=1}^{t-1-D} \gamma^{t-1-s} \left| \mu(a_s^\top \theta_\star^t) - \mu(a_s^\top \theta_\star^s) \right| \|a_s\|_{\mathbf{H}_t^{\text{ED}}(\theta_\star^t)^{-1}} \\
&\leq 2S\bar{L}_\mu \lambda_t^{-1/2} \sum_{s=1}^{t-D-1} \gamma^{t-1-s} \\
&\leq 2S\bar{L}_\mu \lambda_t^{-1/2} \gamma^D / (1-\gamma) \quad (4.18)
\end{aligned}$$

where in the second to last inequality we use  $\mathbf{H}_t^{\text{ED}}(\theta_\star^t) \succeq \lambda_t \mathbf{I}_d$ ,  $\|a_s\| \leq 1$  and  $\mu(a_s^\top \theta_\star^t) - \mu(a_s^\top \theta_\star^s) \leq 2S\bar{L}_\mu$  since both  $\theta_\star^t$  and  $\theta_\star^s \in \Theta$ . We proceed in a similar fashion to bound the second in the r.h.s of Eq. (4.17). Indeed;

$$\begin{aligned}
\left\| \sum_{s=1}^{t-1} \gamma^{t-1-s} \eta_{s+1} a_s \right\|_{\mathbf{H}_t^{\text{ED}}(\theta_\star^t)^{-1}} &\leq \left\| \sum_{s=1}^{t-1-D} \gamma^{t-1-s} \eta_{s+1} a_s \right\|_{\mathbf{H}_t^{\text{ED}}(\theta_\star^t)^{-1}} + \left\| \sum_{s=t-D}^{t-1} \gamma^{t-1-s} \eta_{s+1} a_s \right\|_{\mathbf{H}_t^{\text{ED}}(\theta_\star^t)^{-1}} \\
&\leq 2\sigma \lambda_t^{-1/2} \sum_{s=1}^{t-1-D} \gamma_{t-1-s} + \gamma^{t-1} \left\| \sum_{s=t-D}^{t-1} \gamma^{-s} \eta_{s+1} a_s \right\|_{\mathbf{H}_t^{\text{ED}}(\theta_\star^t)^{-1}} \\
&\leq 2\sigma \lambda_t^{-1/2} \gamma^D / (1-\gamma) + \left\| \sum_{s=t-D}^{t-1} \gamma^{-s} \eta_{s+1} a_s \right\|_{\gamma^{-2(t-1)} \mathbf{H}_t^{\text{ED}}(\theta_\star^t)^{-1}} \quad (4.19)
\end{aligned}$$

The most r.h.s term in the above equation can be bounded by applying [Theorem 2.4.1](#). To see this

recall that  $\gamma \in (0, 1)$  and therefore we have the matrix inequality:

$$\begin{aligned}
 \gamma^{-2(t-1)} \mathbf{H}_t^{\text{ED}}(\theta_\star^t) &= \gamma^{-2(t-1)} \sum_{s=t-D}^{t-1} \gamma^{t-1-s} \dot{\mu}(a_s^\top \theta_\star) a_s a_s^\top + \lambda_t \gamma^{-2(t-1)} \mathbf{I}_d \\
 &\succeq \gamma^{-2(t-1)} \sum_{s=t-D}^{t-1} \gamma^{2(t-1-s)} \dot{\mu}(a_s^\top \theta_\star) a_s a_s^\top + \lambda_t \gamma^{-2(t-1)} \mathbf{I}_d \quad (\gamma \in (0, 1)) \\
 &= \sum_{s=t-D}^{t-1} \gamma^{-2s} \dot{\mu}(a_s^\top \theta_\star) a_s a_s^\top + \lambda_t \gamma^{-2(t-1)} \mathbf{I}_d \\
 &:= \widetilde{\mathbf{H}}_t^{\text{ED}}.
 \end{aligned}$$

where we used our temporary notation  $\theta_\star^s = \theta_\star$  for  $s \in [t-D, t]$ . Therefore:

$$\begin{aligned}
 \left\| \sum_{s=t-D}^{t-1} \gamma^{-s} \eta_{s+1} a_s \right\|_{\gamma^{2(t-1)} \mathbf{H}_t^{\text{ED}}(\theta_\star^t)^{-1}} &\leq \left\| \sum_{s=t-D}^{t-1} \gamma^{-s} \eta_{s+1} a_s \right\|_{(\widetilde{\mathbf{H}}_t^{\text{ED}})^{-1}} \\
 &\leq \frac{\sqrt{\lambda_t} \gamma^{-2(t-1)}}{2\sigma \gamma^{-(t-1)}} + \frac{2\sigma \gamma^{-(t-1)}}{\sqrt{\lambda_t} \gamma^{-2(t-1)}} \log \left( \frac{2^d \det(\widetilde{\mathbf{H}}_t^{\text{ED}})^{1/2}}{\delta (\lambda_t \gamma^{-2(t-1)})^{d/2}} \right) \\
 &\leq \frac{\sqrt{\lambda_t}}{2\sigma} + \frac{\sigma d}{\sqrt{\lambda_t}} \log \left( 4(1 + \sigma^2 D) / (d \lambda_t) / \delta \right)
 \end{aligned}$$

with probability at least  $1 - \delta$ . This high probability bound is obtained by directly applying [Theorem 2.4.1](#) after the first inequality with the weights  $w_s = \gamma^{-s}$  (all the requirements of the theorem are met thanks to [Eqs. \(4.1\)](#) and [\(4.2\)](#)). The last inequality is obtained by some straightforward (and rather crude) upper-bounding and the application of [Lemma B.2](#). Assembling the above and [Eq. \(4.19\)](#) yields:

$$\left\| \sum_{s=1}^{t-1} \gamma^{-s} \eta_{s+1} a_s \right\|_{\gamma^{2(t-1)} \mathbf{H}_t^{\text{ED}}(\theta_\star^t)^{-1}} \leq \frac{\sqrt{\lambda_t}}{2\sigma} + \frac{\sigma d}{\sqrt{\lambda_t}} \log \left( 4(1 + \sigma^2 D) / (d \lambda_t) / \delta \right) + 2\sigma \lambda_t^{-1/2} \gamma^D / (1 - \gamma).$$

Combining the above with [Eq. \(4.18\)](#) and [Eq. \(4.17\)](#) yields the announced result.  $\blacksquare$

## Appendix 4.B Proof of [Eq. \(4.4\)](#)

We proved a similar result in the stationary case. The main tool is the self-concordance property which can easily be leveraged to extend the technical results of [Section 1.4.3](#) in the non-stationary case. Fix  $t$  and assume that  $\theta_\star^t \in \mathcal{C}_t^{\text{SW}}$ . The following set of inequalities hold:

$$\begin{aligned}
 \left\| \tilde{\theta}_t - \theta_\star^t \right\|_{\mathbf{H}_t^{\text{SW}}(\theta_\star^t)} &\leq \sqrt{1 + 2S} \left\| \tilde{\theta}_t - \theta_\star^t \right\|_{\mathbf{G}_t^{\text{SW}}(\theta_\star^t, \tilde{\theta}_t)} \\
 &= \sqrt{1 + 2S} \left\| g_t^{\text{SW}}(\tilde{\theta}_t) - g_t^{\text{SW}}(\theta_\star^t) \right\|_{\mathbf{G}_t^{\text{SW}}(\theta_\star^t, \tilde{\theta}_t)^{-1}} \\
 &\leq \sqrt{1 + 2S} \left( \left\| g_t^{\text{SW}}(\tilde{\theta}_t) - g_t^{\text{SW}}(\hat{\theta}_t^{\text{SW}}) \right\|_{\mathbf{G}_t^{\text{SW}}(\theta_\star^t, \tilde{\theta}_t)^{-1}} + \left\| g_t^{\text{SW}}(\theta_\star^t) - g_t^{\text{SW}}(\hat{\theta}_t^{\text{SW}}) \right\|_{\mathbf{G}_t^{\text{SW}}(\theta_\star^t, \tilde{\theta}_t)^{-1}} \right) \\
 &\leq (1 + 2S) \left( \left\| g_t^{\text{SW}}(\tilde{\theta}_t) - g_t^{\text{SW}}(\hat{\theta}_t^{\text{SW}}) \right\|_{\mathbf{H}_t^{\text{SW}}(\tilde{\theta}_t)^{-1}} + \left\| g_t^{\text{SW}}(\theta_\star^t) - g_t^{\text{SW}}(\hat{\theta}_t^{\text{SW}}) \right\|_{\mathbf{H}_t^{\text{SW}}(\theta_\star^t)^{-1}} \right) \\
 &2(1 + 2S) \nu_t(\delta),
 \end{aligned}$$

where we last used that  $\tilde{\theta}_t, \theta_\star^t \in \mathcal{C}_t^{\text{SW}}$ .

## Appendix 4.C Proof of Eq. (4.6)

The goal is here to obtain the bound:

$$\sum_{t=1}^T \|a_t\|_{(\mathbf{V}_t^{\text{sw}})^{-1}}^2 \mathbf{1}(t \in \mathcal{T}_D) \leq 2\lceil T/D \rceil d \log(\lambda_T + T\bar{\ell}_\mu/d)$$

The proof simply decomposes  $\mathcal{T}_D$  into at most  $\lceil T/D \rceil$  stationary blocks and applies the Elliptical Potential lemma on each such blocks. Formally we construct a partition  $\{\tilde{\mathcal{T}}_D^j\}_j$  of  $\mathcal{T}_D$  as follows; let  $\tilde{\mathcal{T}}_D^0 = \emptyset$  and for  $j \geq 1$  define:

$$t_0^j := \min \left( t \in \mathcal{T}_D \setminus \bigcup_{k=1}^{j-1} \tilde{\mathcal{T}}_D^k \right) \quad \text{and}$$

$$\tilde{\mathcal{T}}_D^j := [t_0^j; t_0^j + D] \cap \mathcal{T}_D,$$

until  $\bigcup_{k=1}^j \tilde{\mathcal{T}}_D^k = \mathcal{T}_D$ . Clearly  $\{\tilde{\mathcal{T}}_D^j\}_j$  is a partition of  $\mathcal{T}_D$  of size at most  $\lceil T/D \rceil$ . Furthermore for each  $j \in [1, \lceil T/D \rceil]$  we have:

$$\begin{aligned} \sum_{t \in \tilde{\mathcal{T}}_D^j} \|a_t\|_{(\mathbf{V}_t^{\text{sw}})^{-1}}^2 \mathbf{1}(t \in \mathcal{T}_D) &= \sum_{t \in \tilde{\mathcal{T}}_D^j} \|a_t\|_{(\mathbf{V}_t^{\text{sw}})^{-1}}^2 \\ &\leq \sum_{t=t_0^j}^{t_0^j+D} \|a_t\|_{(\mathbf{V}_t^{\text{sw}})^{-1}}^2 \\ &\leq \sum_{t=t_0^j}^{t_0^j+D} \|a_t\|_{(\mathbf{V}_{t_0^j:t}^{\text{sw}})^{-1}}^2 \end{aligned}$$

where  $\mathbf{V}_{t_0^j:t}^{\text{sw}} \leq \sum_{s=t_0^j}^t a_s a_s^\top + (\lambda_t/\bar{\ell}_\mu)\mathbf{I}_d$ . Applying the Elliptical Potential lemma on this final quantity along with some simple upper bounding yields:

$$\sum_{t \in \tilde{\mathcal{T}}_D^j} \|a_t\|_{(\mathbf{V}_t^{\text{sw}})^{-1}}^2 \mathbf{1}(t \in \mathcal{T}_D) \leq 2d \log(\lambda_T + T\bar{\ell}_\mu/d)$$

Therefore decomposing  $\mathcal{T}_D$  along the partition  $\{\tilde{\mathcal{T}}_D^j\}_j$  of size at most  $\lceil T/D \rceil$  yields the announced result.

## Appendix 4.D Proof of Eq. (4.7)

This result is simply obtained by repeating the proof of [Lemma 3.1.2](#) and resorting to the partition  $\{\tilde{\mathcal{T}}_D^j\}_j$  of  $\mathcal{T}_D$  from [Appendix 4.C](#) whenever applying the Elliptical Potential lemma (note on any block  $\tilde{\mathcal{T}}_D^j$  the environment is stationary and  $\theta_\star^t$  for  $t \in \tilde{\mathcal{T}}_D^j$  is a constant).

# Summary and Future Work

---

## Outline

---

5.1	Summary . . . . .	121
5.2	Remaining challenges and open questions . . . . .	123
5.2.1	Simultaneous statistical and computational efficiency . . . . .	123
5.2.2	Best-arm identification . . . . .	123
5.2.3	Open question: optimality of forgetting mechanisms . . . . .	124

---

## 5.1 Summary

Remember that our original goal was to study the effects of non-linearity in a simple yet broad family of parametric bandits: the Generalized Linear Bandits (GLBs). We now briefly sum up what we have learned on that topic in the previous chapters.

- We have seen in [Chapter 1](#) that GLBs offer two important advantages over the LB approach; from a theoretical perspective the study of GLBs stands as a first-step towards understanding rich rewards signals, by providing a minimalistic yet powerful extension to Linear Bandits. GLBs are also of important practical relevance as they cover a large range of reward distributions (binary, categorical, ..) frequently encountered in real-life situations. We covered the seminal work of [Filippi et al. \(2010\)](#) (which encompasses almost all existing work on GLBs when it comes to the treatment of non-linearity); its great feat is to show that the LB recipes can be successfully applied to non-linear systems. However their linearization approach shows several downsides; it leads to regret upper-bounds of the form:

$$\text{Regret}_{\theta_*}(T) = \tilde{O}\left(\bar{\kappa}_\mu d\sqrt{T}\right),$$

where  $\bar{\kappa}_\mu$  is a problem-dependent constant which embodies the level of non-linearity in the reward signal. The higher the non-linearity, the larger  $\bar{\kappa}_\mu$ , which comes with two disappointing effects; **(1)** for several highly non-linear models of practical importance (*e.g* the Logistic Bandit) this constant is particularly large, which questions the practical relevance of GLBs and **(2)** it suggests that non-linearity stands as a fundamental obstacle when dealing with complex reward structures. We set out to question and hopefully negate this conclusion. This required improvements that are simultaneously of algorithmic and analytical nature. This requires tighter confidence sets, sensitive to the level of non-linearity through the varying variance that it imposes

across the reward signal. It also requires developing an enhanced analysis, aware of the local effects of non-linearity in order to characterize the correct balance between the estimation and prediction aspect of the regret minimization task.

- In [Chapter 2](#) we focused on the construction of enhanced confidence sets. Guided by an asymptotical analysis, we leveraged the tools provided by the theory of self-normalized processes to achieve our goal. This led us to new confidence sets of the form:

$$\left\{ \theta, \left\| \theta - \hat{\theta}_t \right\|_{\mathbf{H}_t(\theta)}^2 \leq d \log(t/\delta) \right\},$$

which captures the effective level of non-linearity through its variance, efficiently measured by the matrix  $\mathbf{H}_t(\theta)$ . We saw that such confidence sets are much smaller than their predecessors and are much better fitted to measure distance between parameters in non-linear situations.

- We applied our new confidence sets to the design of improved GLB algorithms in [Chapter 3](#). Along with an analysis that leverages a smoothness property called self-concordance and allows a local treatment of the non-linearity, we prove that these algorithms enjoy regret upper-bounds of the form:

$$\text{Regret}_{\theta_*}(T) = \tilde{O} \left( d \sqrt{\dot{\mu}(a_*^\top \theta_*) T} + d^2 \bar{\kappa}_\mu \right).$$

This bound replaces the multiplicative dependency in  $\bar{\kappa}_\mu$  by the local sensitivity of the reward signal around the optimal action, which is typically *much* smaller. It also defers the effects of non-linearity to a dominated, second-order term of the regret which is tied to a transitory exploration phase during which the algorithm searches for highly rewarding arms. It therefore tells a more nuanced story about the effects of non-linearity: it mostly impacts the early phase of the interaction, after which the problem simply looks like a linear bandit with slope  $\dot{\mu}(a_*^\top \theta_*)$ . We refined our analysis for the Logistic Bandit by showing that the regret incurred during this transitory phase is tied to the geometry of the problem, and is small (independent of  $\bar{\kappa}_\mu$ ) for several configurations. In particular, when the available actions are  $\mathcal{A} = \mathcal{B}_2(0, 1)$  we showed that the regret bound becomes:

$$\text{Regret}_{\theta_*}(T) = \tilde{O} \left( d \sqrt{\exp(-\|\theta_*\|) T} \right).$$

This embodies the benefits of our approach as this bound is exponentially smaller than existing ones for this problem; it also shows that for some highly non-linear problems, the effects of non-linearity are inexistent and those problems are *easier* than their linear counterparts. Finally, we showed that the scaling displayed above is *minimax-optimal* which suggests that our approach yields the correct characterization of the non-linearity's effects.

- In [Chapter 4](#) we extended our findings to piece-wise stationary environments through the use of forgetting mechanisms. We arrived to the same conclusions as in the stationary case by proving the following regret upper-bound:

$$\text{Regret}_{\theta_*}(T) = \tilde{O} \left( T^{2/3} \Gamma_T^{1/3} \left( \sqrt{\sum_{t=1}^T \dot{\mu}(a_{*,t}^\top \theta_*^t) / T} + \bar{L}_\mu \right) + T^{1/3} \Gamma_T^{2/3} (\bar{\kappa}_\mu + \bar{L}_\mu^2) \right),$$

were  $\Gamma_T$  counts the number of reward switches across the horizon  $T$ . We then considered a more general non-stationary metric known as the budget-variation  $B_T$ . We saw that this setting is much more challenging to address, as even a naive linearization approach fails. We detailed a first effort towards its theoretical treatment by generalizing the linearization approach, leaving a fine treatment of non-linearity for future work.

## 5.2 Remaining challenges and open questions

We discuss here several directions for future research and remaining open questions.

### 5.2.1 Simultaneous statistical and computational efficiency

**Motivation.** We focused in this dissertation on the statistical efficiency of the different GLB algorithms we covered but did not address their *computational* efficiency. Over all those algorithms we saw that OFU-GLB-r is the only fully tractable one as it does not rely on non-convex optimization routine. It however suffers from a high computational load which originates from both its learning and planning mechanisms. **(1)** On the learning part, it requires at every round a *batch* convex optimization procedure to compute  $\hat{\theta}_t$  (up to sufficient precision). This can be quite challenging in practice as it brings the per-round computational cost of the learning aspect of the algorithm to roughly  $\tilde{O}(T)$ . **(2)** On the planning part, we saw that OFU-GLB-r required to solve at each round  $|\mathcal{A}|$  convex programs to find an optimistic pair  $(x_t, \tilde{\theta}_t)$ ; this yields a  $\tilde{O}(|\mathcal{A}|T)$  additional per-round computational cost, even more burdensome when  $|\mathcal{A}|$  is large. An exciting direction of research therefore consists in constructing algorithms that enjoy the improved statistical efficiency of our algorithms but with reduced computational load. This requires to reduce the costs of both the learning and planning mechanisms.

**Challenges.** **(1)** For the learning part the goal is to produce fully *on-line* (that is with  $\tilde{O}(1)$  cost) estimators and confidence sets for  $\theta_*$ . This can for instance be done by using the online convex optimization (OCO) to confidence set conversion introduced by Abbasi-Yadkori et al. (2012); this approach was actually already followed in the GLB literature by Zhang et al. (2016); Jun et al. (2017) but fails to achieve statistical efficiency as the radii of the resulting confidence sets scale linearly with  $\bar{\kappa}_\mu$ . Existing lower-bounds for the regret of OCO algorithms on logistic regression (Hazan et al., 2014) suggests that in all generality this cannot be improved. However such lower-bounds are derived under some reward misspecification and might therefore evade the theoretical settings of GLB algorithms; also, recent improvements (Jézéquel et al., 2020) in the online optimization community relying on *improper* algorithms might provide us with the correct tools to reach our goal - although extending this framework to the bandit setting is not straight-forward. **(2)** From the planning side the most natural solution calls for generalizing the frequentist linear Thompson Sampling of Abeille and Lazaric (2017) - it replaces the computation of an optimistic parameter by sampling, much less costly. The analysis of Abeille and Lazaric (2017) relies on the ellipsoidal shape of the confidence sets in the LB setting; generalizing it to arbitrary convex sets seems like the way to go - although easier said than done.

### 5.2.2 Best-arm identification

This part of the thesis focused on the regret minimization aspects of parametric bandits and left aside the equally important pure-exploration (or best arm identification) task. There exist an important literature on pure exploration in linear bandits (Soare et al., 2014; Tao et al., 2018; Xu et al., 2018; Fiez et al., 2019; Réda et al., 2021). The GLB part remains relatively under-explored but for the work of Kazerouni and Wein (2021) ; their approach however suffers from severe detrimental dependencies in the usual culprits - that is,  $\bar{\kappa}_\mu$ . A remarkable effort to reduce such dependencies in the Logistic Bandit pure exploration setting was recently conducted by Jun et al. (2021) who derived a fixed-design version of our concentration inequality which enjoys a reduced dependency w.r.t the dimension  $d$ . Their approach still suffers from dependencies in  $\bar{\kappa}_\mu$ ; they however show that this is unavoidable in this setting. We believe that an interesting direction of research lies in characterizing the sample complexity of finding “almost-optimal”

arms (for instance, sub-optimal by  $1/\bar{\kappa}_\mu$ ) as we suspect that for the Logistic Bandit this task might be made easier by the flatness of the reward signal around good arms.

### 5.2.3 Open question: optimality of forgetting mechanisms

Our treatment of GLBs in drifting environments raises several open questions. The first one concerns the nature of the difference between the LB and GLB regret upper-bounds. We postulate that this is an artefact of the proof and that an improved analysis should yield the same rates. Fixing it should constitute the first step before trying to extend our non-linear analysis to this challenging setting. The second question is not specific to GLBs, but regards the optimality of forgetting strategies in parametric bandits. Indeed, the only existing lower-bound for non-stationary parametric bandits was obtained by [Cheung et al. \(2019b\)](#) in the linear case, and scales as  $\Omega(B_T^{1/3}T^{2/3})$ . The observed gap with the upper-bounds obtained by a correct analysis could potentially be explained by a fundamental sub-optimality of the forgetting principle. We see several ways of answering this question: **(1)** by providing an improved analysis of forgetting strategies in the general case, matching the lower bound or **(2)** proving lower-bounds for forgetting policies which establish their sub-optimality. Finally, this raises the question of the true minimax rates behind the non-stationary parametric bandit problem. Indeed, we are not aware of existing methods matching the lower-bound of [Cheung et al. \(2019b\)](#)<sup>1</sup> in the general case (*i.e.* without any geometric assumption on the arm set). This might be explained by the nature of this lower-bound, which is obtained on a very specific problem instance (*i.e.* piece-wise stationary) and might be too specific to cover harder non-stationary problems.<sup>2</sup> We believe that establishing new lower-bounds under generic dynamic scenarios (*e.g.* where the ground truth evolves at every round) therefore stands as a crucial missing piece in the non-stationary parametric bandit literature.

---

<sup>1</sup>[\(Chen et al., 2019\)](#) do obtain the desired rates, however for a different adversary of the regret. It is not straight-forward to adapt their guarantees to the more challenging setting discussed here.

<sup>2</sup>Actually, if the environment is known to be piece-wise stationary, we saw in [Section 4.2](#) that the proof strategy can be adapted to avoid the difficulties of the drifting analysis.

## Part II

# Offline Bandits: Robust Policy Evaluation and Optimization

# Learning from Logged Bandit Feedback

---

The second part of the dissertation focus on the problem of learning from logged bandit feedback, which arise from considerations that can be considered as orthogonal to the first part. The goal of this first chapter is to motivate and introduce the learning problem. This is done by examining the case of recommender systems - a practical setting which highlights some shortcomings of purely online approaches. We describe the notion of *offline* policy evaluation, which goal is to forecast the performances (in terms of expected collected reward) of *any* strategy based on a static dataset obtained through the interactions with the environment of a reference strategy. Reaching this goal enables for offline policy selection: ranking candidate strategies before-hand and selecting the most promising one for deployment in the environment. It also enables for offline policy optimization, which reduces the selection problem to an optimization program and leverages past interactions in a data-driven approach to automatically discover better strategies. We present and discuss state-of-the-art approaches for policy evaluation and optimization in the bandit literature, as well as their ties to approaches from reinforcement learning. While an important part of the literature focuses on providing point estimate for policy evaluation, we argue for the high practical need for confidence intervals for this task. This leads us to present the *counterfactual risk minimization* principle of [Swaminathan and Joachims \(2015a\)](#), a *risk-averse* approach based on empirical Bernstein confidence bounds for policy evaluation. Albeit displaying enjoyable theoretical guarantees, it suffers from several drawbacks which limits its use in practical settings. We discuss such limitations, with the goal of circumventing them in the next chapter.

## Outline

---

6.1	Motivation and formalization . . . . .	127
6.1.1	The ad-placement case . . . . .	127
6.1.2	The learning problem . . . . .	128
6.2	Counterfactual estimation . . . . .	129
6.2.1	Counterfactual estimators . . . . .	129
6.2.2	Confidence intervals . . . . .	131
6.3	The Counterfactual Risk Minimization principle . . . . .	132
6.3.1	A variance-regularized objective . . . . .	133
6.3.2	Limitations . . . . .	135

---

## 6.1 Motivation and formalization

### 6.1.1 The ad-placement case

**Contextual decision-making in the real world.** Before giving a formal definition of the learning problem, we will motivate it by discussing a real-life instance of contextual decision-making, arising from recommender systems. In particular, we will consider the ad-placement task. The setting is the following: at a certain point in time, a given user arrives on a *publisher* website, which sells blank or available sections of its page for ad hosting. Information about the user (*e.g* recent browsing history) is communicated to a *seller* - a third party which is trying to advertise for its products. If interested in targeting that user, the seller can pay a small fee to the publisher and acquire some space on the publisher's webpage. The seller must then select a product to suggest to the user; the more relevant this recommendation, the higher the chances for the user to purchase the product and for the seller to make a profit. This game repeats over time with different users with different preferences, that the seller must learn to adapt its recommendation (this makes it a contextual decision-making problem) to increase its revenue. We discussed in [Chapter 1](#) some fundamental challenges behind this problem (*e.g* the exploration/exploitation dilemma) and principled approaches to address them. Unfortunately, they may be disqualified by constraints that decision-makers face in real-life situations, such as the ad-placement problem. For instance, we saw the importance of exploration for minimizing long-term metrics such as the regret. It turns out that in practice, some short-term metrics are more important. In the ad-placement problem, exploration involves recommending products that, given current estimates, are not likely to lead to a purchase. In such cases, the seller suffers a net loss corresponding to the fee paid to the publisher to buy the ad space. Therefore, too much exploration can *in the short-term* lead to a loss of revenue, potentially prohibitive (the seller could go bankrupt before seeing the positive effects of exploration).

**Risk-aversion and warm-starts.** To avoid such outcomes, decision-makers in the real-life tend to be rather risk-averse; unfortunately, this goes against mechanisms such as optimism, which are *risk-seeking* approaches. The cost of exploration is particularly high in the beginning of the experiment, during which the seller must try out most of its catalogue to learn a decent user to product mapping. Fortunately for the seller, it does not truly face a *cold-start* setting as most often a great deal of information is available before any interactions. For instance, the seller might have access to what is often referred to as organic data: a collection of products seen by users when browsing directly on the seller's platform. Leverage this data to learn which product might interest which user already provides a decent strategy for recommendation, which will generate more revenue than an exploratory risk-seeking alternative in the short-term.

**Learning from logged bandit feedback.** If a strategy that is based on extraneous data provides a good starting point, it surely is sub-optimal and we expect it to be improvable through data-driven approach. Note that by adding a small, and more importantly *controllable* amount of exploration (through an  $\epsilon$ -greedy mechanism) allows (at least conceptually) for the long-term discovery of better strategies while working on a short-term revenue constraint. For our ad-placement example, one single day of deployment usually generates tens of thousands of opportunities for the seller to advertise its products. That is so many bandit interactions (*e.g* context, actions and reward triples) that can easily be logged at little cost. Before directly deducing from the logged data an improved strategy, an intermediary step is to leverage this data to forecast the performance (in terms of expected revenue) of other policies. In other words, infer the performance of any system from the logs, as if it was taking the actions by itself. This task, referred to as *offline policy evaluation* requires the design of counterfactual estimators to

abstract out the bias towards actions favored by the logging strategy. Even more important than the design of such estimators is the construction of confidence intervals for offline policy evaluation, so that in a defensive move the risk-averse decision-maker can judge strategies based on their probable *worst-case* performance. The action of selecting policies for future deployment based on the logged data through counterfactual estimators is often referred to *offline policy evaluation*.

### 6.1.2 The learning problem

We will formally introduce counterfactual estimators and associated confidence intervals in [Section 6.2](#). Before, we remind and introduce some useful notations and formalize the learning objective. In the following, we will use  $x$  to denote a context and  $a \in [K]$  an action, where  $K$  denotes the number of fixed available actions. In this part of the dissertation, we will work under a standard distributional assumption on the context.

**Assumption 6.1.1** (Context distribution). *The contexts are drawn i.i.d according to a distribution  $\nu$ , whose support  $\mathcal{X}$  is a compact subset of  $\mathbb{R}^p$  for some  $p \in \mathbb{N}$ .*

**Remark 6.1.1** (About the context distribution assumption). *Assumption 6.1.1 is stronger than assuming that an oblivious adversary is picking the contexts, as we did in the first part of the dissertation. It is nonetheless quite a logical assumption to make in real-world situations; for the ad-placement example, users typically arrive independently of each other. If needed, this assumption could be lessened to fast-mixing distributions which will imply that the empirical distribution along a sequence of context is close to i.i.d - see [Duchi et al. \(2016\)](#).*

The strategy followed by a decision-maker is formalized by a policy  $\pi : \mathcal{X} \rightarrow \Delta_K$  - i.e a mapping from context to distributions over actions which quantifies how likely the decision-maker is to select an action when presented with a context. We will slightly overload this notation and denote  $\pi(x, a) = [\pi(x)]_a$  the probability of selecting action  $a$  when the context is  $x$ . Given a context  $x$ , each action is associated with a reward  $r(x, a)$  (with the convention that better actions have higher rewards) and where the reward function  $r(\cdot)$  is unknown. We make the following assumption on the reward function.

**Assumption 6.1.2** (Bounded reward). *The reward function is bounded, and such that:*

$$r(x, a) \in [0, 1] \quad \forall x \in \mathcal{X}, \forall a \in \mathcal{A}.$$

This technical assumption is made for ease of exposition. It can be explicitly enforced by re-scaling the reward function. The goal of the decision-maker is to find a policy of high performance; here, we measure performance in terms of *expected* reward or *value*. This is quantified by the value function:

$$\text{Value}(\pi) := \mathbb{E}_{x \sim \nu} \mathbb{E}_{a \sim \pi(x)} [r(x, a)] .$$

Equivalently, one can introduce the *cost* function  $c(x, a) = -r(x, a)$  and the risk function:

$$\text{Risk}(\pi) := \mathbb{E}_{x \sim \nu} \mathbb{E}_{a \sim \pi(x)} [c(x, a)] .$$

with the convention that the lower the risk, the better the policy. Naturally, minimizing the risk function is equivalent to maximizing the value function. To stay coherent with existing work in the bandit literature, we will now refer to the risk (rather than the value) as the performance measure for the rest of this chapter. As we discussed in the previous section, it is not reasonable in practical scenarios to expect having the luxury of testing out several policies in order to

estimate and compare their empirical risk, to retain whichever policy has the smallest. Instead, we wish to leverage logged data, obtained by deploying a *logging policy* that we denote  $\pi_0$ . More precisely, we assume the access of a static dataset  $\mathcal{D}_n$  of length  $n$ , defined as follows:

$$\mathcal{D}_n := \left\{ x_i \sim \nu, a_i \sim \pi_0(x_i), c_i := c(x_i, a_i), p_i^0 := \pi_0(x_i, a_i) \right\}_{i \in [n]} .$$

Given a policy  $\pi$ , offline policy evaluation (OPE) consists in building an estimator for the risk  $R(\pi)$  based only on  $\mathcal{D}_n$ ; informally laid-out:

$$\text{construct } \hat{R}_n(\pi) = f(\mathcal{D}_n, \pi) \overset{\text{proxy}}{\rightsquigarrow} \text{Risk}(\pi) . \quad (\text{OPE})$$

for some real-valued function  $f$ . Offline policy optimization (OPO) consists in using such estimator to search for a policy of lowest risk; again for informal exposition, with  $\text{penalty}(\cdot)$  being a user-defined regularization function:

$$\text{compute } \hat{\pi}_n \in \arg \min_{\pi} \hat{R}_n(\pi) + \text{penalty}(\pi) \overset{\text{proxy}}{\rightsquigarrow} \pi^* \in \arg \min_{\pi} \text{Risk}(\pi) , \quad (\text{OPO})$$

henceforth reducing policy learning to an optimization problem.

## 6.2 Counterfactual estimation

We introduce and discuss here counterfactual estimators as well as counterfactual confidence intervals for offline policy evaluation.

### 6.2.1 Counterfactual estimators

**Direct methods.** One approach to build estimators for  $R(\pi)$  from  $\mathcal{D}_n$  is to leverage the static dataset to learn a model  $\hat{c}_n(x, a)$  of the cost function  $c(\cdot)$  and estimate the risk from the approximated cost instead of the true cost; this is most often referred to as *direct methods* (DM) with estimators of the form:

$$\hat{R}_n^{\text{DM}}(\pi) := \frac{1}{n} \sum_{i=1}^n \sum_{a \in [K]} \hat{c}_n(x_i, a) \pi(x_i, a) . \quad (6.1)$$

Such estimators are unfortunately *biased* without a perfect model for the cost function, which is typically unavailable (and most often even unattainable) for real-world situations. In what follows, we discuss propensity re-weighting to obtain unbiased estimators without a perfect model of the cost function.

**Inverse Propensity Scoring.** One way to obtain an unbiased estimator of  $R(\pi)$  based only on  $\mathcal{D}_n$  is to leverage a mechanism called *inverse propensity scoring* (IPS) (Horvitz and Thompson, 1952; Rosenbaum and Rubin, 1983). The principal idea to remove the preference bias of  $\pi_0$  from  $\mathcal{D}_n$  by re-weighting samples based on the discrepancy between  $\pi_0$  and  $\pi$ . Formally, introduce the propensity weights  $\omega^\pi(x, a) := \frac{\pi(x, a)}{\pi_0(x, a)}$  and the IPS risk estimator

$$\hat{R}_n^{\text{IPS}}(\pi) = \frac{1}{n} \sum_{i=1}^n \omega^\pi(x_i, a_i) c(x_i, a_i) , \quad (6.2)$$

first considered by Bottou et al. (2013) for counterfactual estimation. Under the technical requirement for  $\pi$  to be absolutely continuous w.r.t  $\pi_0$  (denoted  $\pi \ll \pi_0$ ) so that propensity weights are well-defined, it is straight-forward to show that the IPS estimator is *unbiased*:

$$\mathbb{E} \left[ \hat{R}_n^{\text{IPS}}(\pi) \right] = R(\pi) ,$$

This is a crucial advantage over direct methods. However, the IPS estimator suffers from a large (even potentially infinite) variance. Such behavior is highly related to the values of the propensity weights. Indeed through elementary computations one can show that:

$$\text{Var}\left(\hat{R}_n^{\text{IPS}}(\pi)\right) \geq \left(\mathbb{E}_{x \sim \nu} \mathbb{E}_{a \sim \pi(x)} \left[\omega^\pi(x, a) c(x, a)^2\right] - 1\right) / n .$$

In other words, the variance of the IPS estimator grow linearly with the importance weights  $\omega^\pi$  - which can be unbounded! In particular, the importance weight  $\omega^\pi(x, a)$  is very large when given  $x$  the policy  $\pi$  assigns a high probability to action  $a$ , whereas it was very unlikely to be played by  $\pi_0$ . This examples highlight that the variance of  $\hat{R}_n^{\text{IPS}}(\pi)$  is highly dependent on  $\pi$  (and its discrepancy w.r.t  $\pi_0$ ) and can be extremely large.

**Variance reduction.** An important strand of research have focused on reducing the variance of the vanilla IPS estimator. Perhaps the simplest way to do so is to explicitly trade bias for variance, for instance by *clipping*<sup>1</sup> the propensity weights (Ionides, 2008);

$$\omega^\pi(x, a) = \min\left(M, \frac{\pi(x, a)}{\pi_0(x, a)}\right), \quad (6.3)$$

where  $M > 0$ . While this necessarily introduce bias, it forces the variance to be bounded. A more sophisticated approach involves the use of control variates to reduce the variance. A good candidate for an additive control variate is the estimation given by a direct method - cf. Eq. (6.1). This idea lead to so-called *doubly robust* (DR) risk estimators introduced by Dudík et al. (2014):

$$\hat{R}_n^{\text{DR}} := \frac{1}{n} \sum_{i \in [n]} \sum_{a \in [K]} \hat{c}(x_i, a) \pi(x_i, a) + \frac{1}{n} \sum_{i \in [n]} \omega^\pi(x_i, a_i) [c(x_i, a_i) - \hat{c}(x_i, a_i)]. \quad (6.4)$$

Such an estimator is *unbiased* and typically has lower variance than the original IPS. It however comes with some additional methodological burdens, typically to select a model for the reward estimator  $\hat{c}(\cdot)$  which will involve splitting  $\mathcal{D}_n$  for the train/validation/test procedure. The idea behind the DR estimator, presented here in its bluntest form for the sake of clarity has driven an important stream of research for offline policy evaluation, mostly focusing on deriving improved additive control variates - see for instance (Wang et al., 2017; Farajtabar et al., 2018; Vlassis et al., 2019). While the DR approach relies on additive control variates, other development have involved multiplicative control variates for the IPS (Hesterberg, 1995) leading to self-normalized IPS (SNIPS) risk estimators by Swaminathan and Joachims (2015b):

$$\hat{R}_n^{\text{SNIPS}} := \frac{1}{n} \sum_{i=1}^n \tilde{\omega}^\pi(x_i, a_i) c(x_i, a_i) \quad \text{where} \quad \tilde{\omega}^\pi(x, a) = \frac{\omega^\pi(x, a)}{\sum_{j=1}^n \omega^\pi(x_j, a_j)}. \quad (6.5)$$

This estimator is unfortunately biased (however with bias disappearing asymptotically), but has been shown to have smaller variance than the IPS. It can be easily combined with the doubly robust approach for greater variance reduction, by replacing the propensity weights  $\omega^\pi$  with their renormalized counterparts  $\tilde{\omega}_\pi$  in Eq. (6.4).

**Remark 6.2.1** (Related work from Reinforcement Learning). *We described counterfactual estimators for the contextual bandit setting. There exists a substantial literature on off-line policy evaluation in the Reinforcement Learning setting, for which the IPS approach received important attention (Jiang and Li, 2016; Thomas and Brunskill, 2016; Xie et al., 2019). It however suffers from even greater variance (Mandel et al., 2014) in this setting - especially for long-horizon problems.*

<sup>1</sup>In practice, weight-clipping or equivalent approaches are most often used for the numerical stability it provides.

## 6.2.2 Confidence intervals

**Motivation.** We mentioned that the variance of the naive IPS estimator can be large, and depends on which policy is being evaluated. This is quite unfortunate for the risk-averse decision-maker, as for a particular realization of the dataset  $\mathcal{D}_n$  the actual performance of a policy can be way *worse* than what was predicted by the IPS estimator. This stresses the paramount importance of obtaining confidence intervals for offline policy evaluation. For instance, it allows to consider what could be the worst-case outcome when deploying a new policy. It therefore allows for defensive positions when selecting future policies - *e.g.* select the policy with smallest confident upper-bound on the true risk. Note that this reasoning also applies to other estimators (*e.g.* doubly robust and self-normalized approaches); even if their variance is smaller, it unavoidably depends on the level of discrepancy between the policy  $\pi$  being evaluated and the logging policy  $\pi_0$ . A substantial proportion of the existing literature on confident policy evaluation focuses on the IPS estimator to build confidence intervals for the true risk, as it is unbiased and writes as a sum of *i.i.d* random variables (unlike, for instance, the SNIPS estimator).

**An asymptotic confidence interval.** Bottou et al. (2013) proposed relying on the central limit theorem (*cf.* Theorem A.1 and Lemma A.1) applied to the IPS estimator to derive approximate confidence intervals. Using for short-hand  $\sigma^2(\pi) = \text{Var}(\omega^\pi(x, a))$  and  $\Phi$  the cumulative distribution function of the standard Gaussian random variable, this states that for a fixed policy  $\pi$  and a confidence level  $\delta \in (0, 1]$ :

$$\mathcal{I}_n^{\text{CLT}}(\pi) := \left[ \hat{R}_n^{\text{IPS}}(\pi) + \frac{\sigma(\pi)}{\sqrt{n}} \Phi^{-1}(\delta/2), \hat{R}_n^{\text{IPS}}(\pi) - \frac{\sigma(\pi)}{\sqrt{n}} \Phi^{-1}(\delta/2) \right],$$

is an asymptotic  $(1 - \delta)$ -confidence interval for  $R(\pi)$ ; *i.e.*  $\lim_{n \rightarrow \infty} \mathbb{P}(R(\pi) \in \mathcal{I}_n^{\text{CLT}}(\pi)) \geq 1 - \delta$ . This approximated confidence interval has a particularly enjoyable feature for the problem at hand as it captures the policy-dependent variance of the estimators. We can therefore expect the confidence intervals for policies *close* to  $\pi_0$  to be thin (small variance) and large for policies *far* from  $\pi_0$  (large variance). The variance  $\sigma^2(\pi)$  is of course unknown but can be replaced by its empirical counterpart without impacting the validity of the asymptotic confidence interval - see Lemma A.2.

**Finite-time confidence intervals.** It is natural to investigate how several concentration inequalities can be used to design finite-time confidence intervals for the risk, such as Chernoff-Hoeffding's concentration inequality. Let  $\pi$  be fixed and  $b_\pi$  be an upper-bound on the counterfactual weights  $\omega^\pi$ . Then by applying the Chernoff-Hoeffding's tail-inequality (*cf.* Lemma A.4) we obtain that for a confidence level  $\delta \in (0, 1]$ :

$$\mathcal{I}_n^{\text{CH}}(\pi) := \left[ \hat{R}_n^{\text{IPS}}(\pi) - b_\pi \sqrt{\frac{\log(2/\delta)}{2n}}, \hat{R}_n^{\text{IPS}}(\pi) + b_\pi \sqrt{\frac{\log(2/\delta)}{2n}} \right], \quad (6.6)$$

is a  $(1 - \delta)$ -confidence interval for the risk; *i.e.*  $\mathbb{P}(\text{Risk}(\pi) \in \mathcal{I}_n^{\text{CH}}(\pi)) \geq 1 - \delta$ . Note that the width of  $\mathcal{I}_n^{\text{CH}}(\pi)$  shows a direct dependency with  $b_\pi$ , which in all generality can be prohibitively large. Computing a tight upper-bound for the counterfactual weights requires domain-specific knowledge. In general, we can only settle for a crude upper-bound that holds uniformly for all policies, if  $\pi_0$  was forced to have fat tails. For instance, if  $\pi_0(x, a) \geq \varepsilon$  for some small known  $\varepsilon > 0$  we obtain  $b_\pi = 1/\varepsilon$  for all  $\pi$ . While Chernoff-Hoeffding's bound only relies on the sample mean, the empirical Bernstein's inequality Maurer and Pontil (2009) also leverages the sample variance. If we denote:

$$s_n^2(\pi) := \frac{1}{n(n-1)} \sum_{1 \leq i < j \leq n} (\omega^\pi(x_i, a_i) c(x_i, a_i) - \omega^\pi(x_j, a_j) c(x_j, a_j))^2, \quad (6.7)$$

the empirical variance of the IPS estimator then the empirical Bernstein’s inequality (cf. (Maurer and Pontil, 2009, Theorem 3) or Lemma A.6) yields another confidence interval for the risk:

$$\mathcal{I}_n^{\text{EB}}(\pi) := [\hat{R}_n^{\text{IPS}}(\pi) \pm \left( \frac{7b_\pi}{3(n-1)} \log(4/\delta) + \sqrt{\frac{2 \log(4/\delta)}{n} s_n^2(\pi)} \right)], \quad (6.8)$$

with coverage at least  $1 - \delta$ . Notice the smaller effect of  $b_\pi$  which now comes with a quickly vanishing term with rate  $1/n$  instead of  $1/\sqrt{n}$  with Chernoff-Hoeffding’s bound. The dominant term, scaling as  $1/\sqrt{n}$  involves the empirical variance  $s_n^2(\pi)$  of the weights. As we discussed earlier with the central limit theorem confidence interval, this effectively measures the uncertainty of the IPS estimator, which we expect to be accurate for policies being close to  $\pi_0$  (small variance) and unreliable for policy which are far (high variance). In this context, this difference between the behaviors of  $\mathcal{I}_n^{\text{CH}}(\pi)$  and  $\mathcal{I}_n^{\text{EB}}(\pi)$  is quite important, as we will illustrate in Fig. 6.1 under the lens of the offline policy optimization task. It is worth mentioning that the dependency in  $b$  can be reduced further. (Thomas et al., 2015) proved a concentration inequality, similar to the empirical Bernstein but which doesn’t depend directly on  $b_\pi$  but rather on the observed values of the counterfactual weights.

**Other related work.** We will focus in this dissertation in IPS-based confidence intervals, such as the ones we presented above. It is however worth mentioning that recently Kuzborskij et al. (2020) investigated the design of confidence intervals based on the SNIPS estimator. This is a technically more challenging task, namely because this estimator cannot be written as a sum of *i.i.d* random variables after re-normalization. They derive non-asymptotic confidence intervals based on the SNIPS estimator, namely by controlling its bias via Harris’ inequality and proving finite-time concentration inequalities through a semi-empirical Efron-Stein tail-inequality. Another stream of related work (Dai et al., 2020; Karampatziakis et al., 2020) departs from classical confidence intervals which quantify the deviation of sample-average estimators from their mean but leverages ideas from the empirical likelihood approach (Owen, 2001). These works are concurrent and closely related to the results we will present in Chapter 7; we defer a dedicated discussion to Chapter 8.

### 6.3 The Counterfactual Risk Minimization principle

In this section we describe the counterfactual risk minimization (CRM) principle of Swaminathan and Joachims (2015a), a conservative offline policy optimization approach that leverages the IPS estimator and is inspired from the empirical Bernstein confidence interval from Eq. (6.8).

As previously discussed, ranking policies based only on the IPS risk estimator is rather risky. It is preferable for a risk-averse decision maker to rank them according to a upper confidence bound on the true risk, for instance by combining the IPS estimator with a confidence interval. The same principle holds for offline policy optimization, where one wishes to learn a new policy directly from  $\mathcal{D}_n$ . Bluntly returning  $\hat{\pi} \in \arg \min_{\pi} \hat{R}_n(\pi)$  is perilous as for a bad realization of the dataset  $\mathcal{D}_n$ , the estimator  $\hat{R}_n(\hat{\pi}_n)$  might have a large variance and be a very poor proxy for the true risk  $\text{Risk}(\hat{\pi}_n)$ . Instead, it is safer to leverage confidence intervals to construct a confident upper-bound  $\bar{R}_n$  on the true risk. Assuming  $b$  is a known uniform bound on the counterfactual weights, the upper-bound given by the Chernoff-Hoeffding’s confidence interval (Eq. (6.6)) will write:

$$\bar{R}_n^{\text{CH}}(\pi) := \hat{R}_n^{\text{IPS}}(\pi) + b \sqrt{\frac{\log(2/\delta)}{2n}}$$

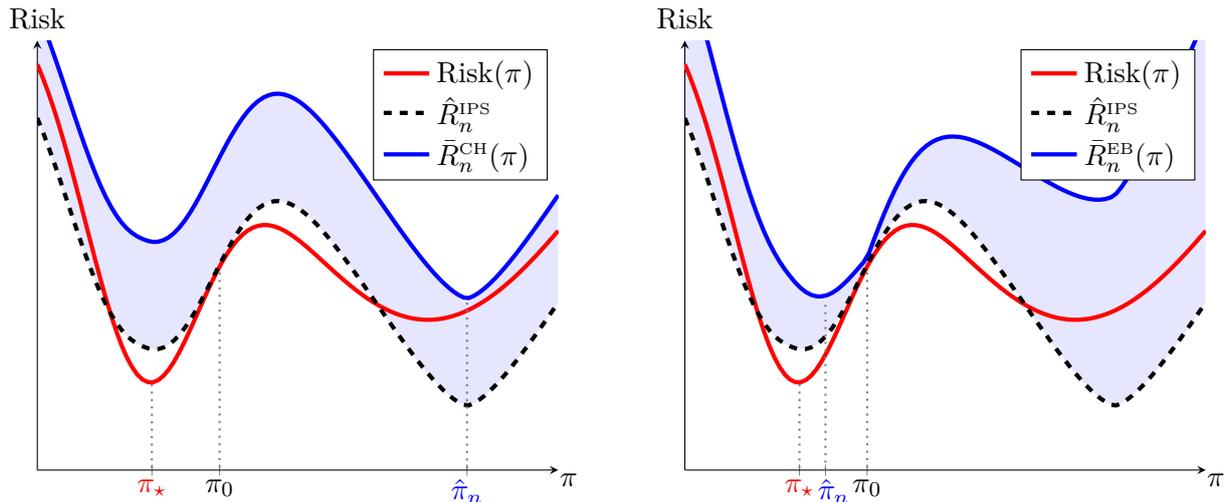


Figure 6.1: Conservative policy optimization based on Chernoff-Hoeffding (CH, left) and empirical Bernstein (EB, right) upper-confidence bound on the risk (smaller is better). The goal is to find a policy of small risk, approaching the optimal policy  $\pi_*$ . Without domain-specific knowledge to upper-bound the counterfactual weights in a policy-dependent way, the upper-bound by CH is uniform over all policies. The selected policy  $\hat{\pi}_n$  minimizing this upper-bound is also the minimizer of the IPS estimator, which here under-estimates the true risk of  $\hat{\pi}_n$ . The situation is different for EB, as it penalizes policies which are far from  $\pi_0$  and for which the IPS estimator is subject to high-variance. This allows for a conservative choice, and here a better choice of  $\hat{\pi}_n$ .

and is so that for every  $\pi$ ,  $\text{Risk}(\pi) \leq \bar{R}_n^{\text{CH}}(\pi)$  with high probability. This is unfortunately not enough; the resulting IPS regularization is policy independent, and therefore  $\arg \min_{\pi} \bar{R}_n^{\text{CH}}(\pi) = \arg \min_{\pi} \hat{R}_n(\pi)$ . A positive alternative is brought by the empirical Bernstein confidence interval (Eq. (6.8)), for which the confident upper-bound writes:

$$\bar{R}_n^{\text{EB}}(\pi) := \hat{R}_n^{\text{IPS}}(\pi) + \sqrt{\frac{2 \log(4/\delta)}{n} s_n^2(\pi)} + \frac{7b}{3(n-1)} \log(4/\delta).$$

Note that the regularizer now depends on  $\pi$  through the empirical variance  $s_n^2(\pi)$ . This effectively penalizes policies whose estimator have high variance and might be unreliable. The advantage for offline policy optimization of this upper-bound over the one inherited from the Chernoff-Hoeffding's concentration inequality is illustrated in Fig. 6.1, and is the leading idea behind the CRM principle.

### 6.3.1 A variance-regularized objective

Motivated by the variance-sensitive nature of the first term of the empirical Bernstein confident upper-bound, the counterfactual risk minimization principle of Swaminathan and Joachims (2015a) prescribes building a penalty term using the empirical variance as a *data-dependent* regularizer. We now give a formal definition of this learning principle.

**Definition** (Counterfactual Risk Minimization (Swaminathan and Joachims, 2015a)). *When learning with logged bandit feedback, the counterfactual risk minimization learning principle returns the following policy:*

$$\hat{\pi}_n \in \arg \min_{\pi} \left\{ \hat{R}_n^{\text{IPS}}(\pi) + \lambda \sqrt{\frac{s_n^2(\pi)}{n}} \right\}, \quad (\text{CRM})$$

where  $\lambda$  is hyper-parameter set by the user and  $s_n^2(\pi)$  is the empirical variance of the policy  $\pi$  as defined in Eq. (6.7).

Note the disappearance of the second-order term of  $\bar{R}_n^{\text{EB}}(\pi)$ , which is independent of the policy and therefore doesn't impact the minimizer.

**Remark 6.3.1** (Uniform validity of the confidence bound). *The empirical Bernstein upper-bound as we have stated it so far asserts that for any fixed policy  $\pi$ :*

$$\mathbb{P}\left(\text{Risk}(\pi) \leq \hat{R}_n^{\text{IPS}} + \sqrt{\frac{2 \log(4/\delta)}{n} s_n^2(\pi)} + \frac{7b}{3(n-1)} \log(4/\delta)\right) \geq 1 - \delta.$$

Such a point-wise statement is not enough for the reasoning (eg. minimize a confident upper-bound on the risk) to make sense. Instead the upper-bound should hold uniformly over all considered policies. In particular, it should hold at  $\hat{\pi}_n$  which is a random quantity that depends on  $\mathcal{D}_n$ . [Swaminathan and Joachims \(2015a\)](#) showed that this can be ensured by defining the right notion of capacity for a class of stochastic policies, and applying an union bound. For a given policy class  $\Pi$ , let  $\mathcal{N}_\infty(\varepsilon, \Pi, n)$  its  $\varepsilon$ -covering number (see ([Anthony and Bartlett, 2009](#); [Maurer and Pontil, 2009](#)) and references therein for a formal definition). Defining  $\mathcal{Q}_\infty(\Pi, \delta) := \log(\mathcal{N}_\infty(1/n, \Pi, 2n))$ , this leads to the following statement:

$$\mathbb{P}\left(\forall \pi \in \Pi, \quad \text{Risk}(\pi) \leq \hat{R}_n^{\text{IPS}}(\pi) + C \sqrt{\frac{\mathcal{Q}_\infty(\Pi, \delta)}{n} s_n^2(\pi)} + \frac{C \mathcal{Q}_\infty(\Pi, \delta)}{(n-1)}\right) \geq 1 - \delta,$$

where  $C$  is a universal constant. [Swaminathan and Joachims \(2015a\)](#) therefore obtain a uniform confident upper-bound, which justifies the intuitive motivation for the CRM principle. Indeed, while the capacity  $\mathcal{Q}_\infty(\Pi, \delta)$  of the class is hard to compute, it can in practice be absorbed into the hyper-parameter  $\lambda$ .

**Offline optimization of parametric policies.** When dealing with large context and action spaces, directly optimizing the policy (which can be seen as a  $|\mathcal{X}| \times |\mathcal{A}|$  matrix) is unfeasible (the context space  $\mathcal{X}$  can very well be infinite). Instead, it is natural to search for a good policy in a space of *parametric* policies. - eg. policies parametrized by neural networks. [Swaminathan and Joachims \(2015a\)](#) applied the CRM principle to exponentially parametrized policies; formally, let  $\Theta$  be a compact subset of  $\mathbb{R}^d$  and consider the following policy class:

$$\Pi_\theta^{\text{exp}} := \left\{ \pi_\theta(x, \cdot) = \eta_\theta \exp(\theta^\top \phi(x, \cdot)), \theta \in \Theta \right\} \quad (6.9)$$

where  $\phi : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}^d$  is a given joint feature-map and  $\eta_\theta$  is a normalization constant ensuring that  $\pi_\theta(x)$  is a valid probability distribution over  $\mathcal{A}$ . This gives rise to the POEM algorithm (Policy Optimizer for Exponential Model ([Swaminathan and Joachims, 2015a](#))) which minimizes the following objective:

$$\hat{\theta}_n \in \arg \min_{\Theta} \left\{ \hat{R}_n^{\text{IPS}}(\pi_\theta) + \lambda \sqrt{\frac{s_n^2(\pi_\theta)}{n}} \right\}, \quad (\text{POEM})$$

which can be solved, for instance, by projected plain gradient descent. For numerical stability, the counterfactuals weights are *clipped* by a constant  $M$  (Eq. (6.3)) also defined by the user. [Algorithm 8](#) provides the pseudo-code for such an approach. ([Swaminathan and Joachims, 2015a](#)) suggests tuning the pair of hyper-parameters  $(\lambda, M)$  by cross-validation. For instance, by splitting  $\mathcal{D}_n$  into two parts  $\mathcal{D}_n^{\text{train}}$  and  $\mathcal{D}_n^{\text{valid}}$ , applying the CRM principle on  $\mathcal{D}_n^{\text{train}}$  and select the best-values for  $M$  and  $\lambda$  according to the naive IPS estimator based on  $\mathcal{D}_n^{\text{valid}}$ . This approach stands

---

**Algorithm 8** POEM (Swaminathan and Joachims, 2015a) with projected gradient descent

---

**input:** Parameter space  $\Theta$ , regularization coefficient  $\lambda$ , clipping constant  $M$ , optimization horizon  $T$ , learning rate  $\alpha$ , initial parameter  $\theta_1$ , static bandit dataset  $\mathcal{D}_n = \{x_i, a_i, c_i, p_i^0\}_{i=1}^n$ .

**for**  $t \in [1, T - 1]$  **do**

  Compute likelihood scores:  $\{p_i(\theta_t) \leftarrow \exp(\theta_t^\top \phi(x_i, a_i)) / \sum_{a \in [K]} \exp(\theta_t^\top \phi(x_i, a_i))\}_{i=1}^n$ .

  Compute clipped propensity weights:  $\{\omega_i(\theta_t) \leftarrow \max(M, p_i(\theta_t)/p_i^0)\}_{i=1}^n$ .

  Compute the empirical mean:  $\hat{R}_n(\theta_t) = \frac{1}{n} \sum_{i=1}^n \omega_i(\theta_t) c_i$ .

  Compute the empirical variance:  $s_n^2(\theta_t) \leftarrow \frac{1}{n-1} \sum_{i=1}^n (\omega_i(\theta_t) c_i - \hat{R}_n(\theta_t))^2$ .

  Perform a gradient descent step:

$$\tilde{\theta}_{t+1} \leftarrow \theta_t - \alpha \nabla_{\theta_t} \left( \hat{R}_n(\theta) + \lambda \sqrt{\frac{s_n^2(\theta)}{n}} \right).$$

  Project back to  $\Theta$ :  $\theta_{t+1} \leftarrow \arg \min_{\theta \in \Theta} \|\theta - \tilde{\theta}_{t+1}\|^2$ .

**end for**

**return** policy  $\pi_{\theta_T}$ .

---

as state-of-the-art for offline policy optimization, and (Swaminathan and Joachims, 2015a) reported that empirically, the policies returned by the CRM principle have much lower risk than those obtained by bluntly minimizing the IPS objective.

**Beyond the IPS estimator.** This good empirical results of the CRM principle also hold for other estimator than the IPS - even though the empirical Bernstein bound does not apply for such estimators. For instance, Swaminathan and Joachims (2015b) reported that empirical variance penalization also improved the offline policy optimization when it relies on the SNIPS estimator (Eq. (6.5)) - although this estimator typically has smaller variance than the IPS. This hints that even if the variance of a risk estimator is reasonable, its policy-dependent nature speaks in favor of empirical variance regularization to discover better policies.

### 6.3.2 Limitations

The CRM principle however suffers from two important limitations, inherited from the empirical-variance regularizer.

**Non-convexity.** Contrary to the IPS objective which is linear (and therefore convex) in  $\pi$ , the CRM objective adds a square-root variance penalization;

$$s_n(\pi) = \sqrt{\frac{1}{n-1} \sum_{i=1}^n \left[ \omega^\pi(x_i, a_i) c(x_i, a_i) - \frac{1}{n} \sum_{j=1}^n \omega^\pi(x_j, a_j) c(x_j, a_j) \right]^2}, \quad (6.10)$$

which breaks this convexity. This results in ill-posed optimization programs, for which classical optimization techniques might fail. In practice, this may potentially hinder the statistical benefits brought by the regularizer, since a good minimizer of the CRM objective can therefore be challenging to discover. We illustrate the non-convexity of the empirical variance term on a toy example in Fig. 6.2. Providing an alternative statistical principle for offline policy optimization

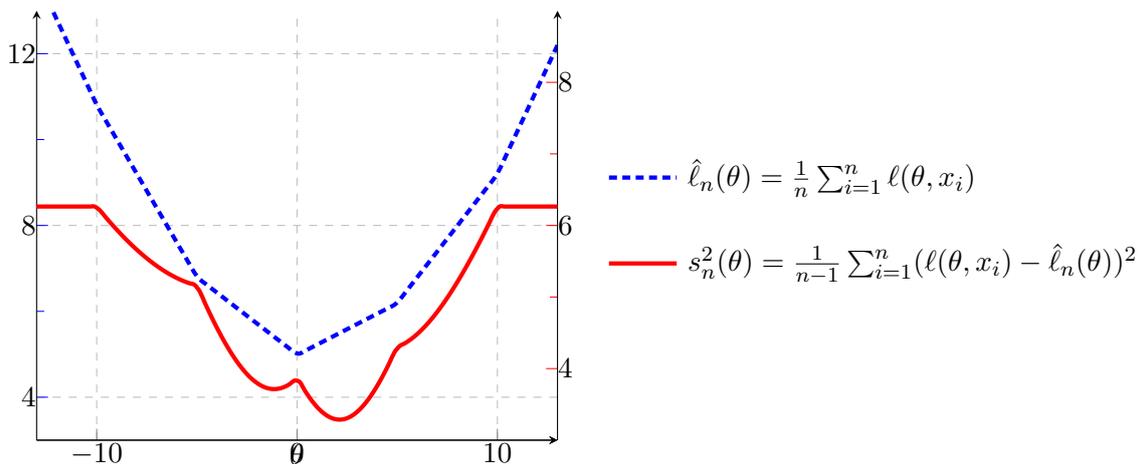


Figure 6.2: An illustration of the non-convex nature of the empirical variance on a toy example. The quantity to estimate is  $\ell(\theta) = \mathbb{E}_x[\ell(\theta, x)]$  where  $\ell(\theta, x) = |\theta - x|$  and  $x$  is sampled according to a mixture of Gaussian. We draw  $n = 100$  samples  $\{x_i\}_{i \in [n]}$  and plot the resulting sample-mean estimator  $\hat{\ell}_n(\theta)$  and empirical variance  $s_n^2(\theta)$ . While both  $\ell(\theta)$  and  $\hat{\ell}_n(\theta)$  are convex in  $\theta$ , the empirical variance  $s_n^2(\theta)$  is a non-convex function of  $\theta$ .

which would translate into a *convex* objective is therefore highly desirable for practical applications. Indeed, standard optimizers could then be deployed safely to optimize this objective, and provably discover good minimizers with arbitrary precision.

**Remark 6.3.2** (Parametrized policies). *One could argue that even though the IPS estimator is convex w.r.t  $\pi$ , it might become non-convex once the policy is parametrized and the optimized variable becomes the parameter  $\theta$  - e.g for exponentially parametrized class of policies  $\Pi_\theta^{\text{exp}}$ . For this reason, trying to find convex alternatives for the empirical variance could be a vain attempt to obtain well-posed optimization for parametric problems. In [Chapter 7](#) we will provide empirical evidence that even for parametrized policies the empirical variance is the main source of optimization challenges, and that convex alternatives give rise to better optimization behaviors.*

**Scalability.** Another important limitation of the CRM principle is its scalability to large logged datasets  $\mathcal{D}_n$ . Computing the gradient of the CRM objective requires going through the entire dataset and therefore has a computation cost and memory cost  $\mathcal{O}(n)$ . In practice,  $n$  is extremely large and such a cost is prohibitively large - consider for instance our recommender system example, where at least tens of thousands of interactions can be logged everyday. This issue is common in many machine learning problems and is usually solved by resorting to stochastic optimization algorithm, which requires only access to *unbiased* stochastic gradients of the objective. Those are particularly easy to obtain when the objective is *composite* (i.e it writes as a finite sum over the dataset's entries) as sub-sampling the dataset is enough to obtain an unbiased gradient. Unfortunately, this is not the case for the CRM objective - again because of the empirical variance term which does not writes as a sum - cf. [Eq. \(6.10\)](#). It is therefore not well-suited for stochastic optimization, as obtaining unbiased stochastic gradients of their related objectives is not straightforward. [Swaminathan and Joachims \(2015a\)](#) proposed a relaxation of the CRM objective amenable to stochastic gradients, however only applicable in the case of exponential policies. Their approach consists in a majorization-minimization strategy, which still requires going through the whole logged dataset once in a while.

**Hyper-parameter selection.** Finally, the CRM principle comes with two hyper-parameters: the clipping threshold  $M$  and the regularization amplitude  $\lambda$ . While good values of  $M$  can be deduced from the empirical repartitions of the weights, the constant  $\lambda$  requires careful tuning as its choice drastically impacts the performance of the obtained policy. For good performance,  $\lambda$  should be cross-validated over a relatively fine grid, which adds to the computational complexity of the algorithm.

These shortcomings of the CRM principle limits its applicability in real-life situations. We will present in [Chapter 7](#) an alternative formulation through distributionally robust optimization, which circumvents or at least mitigates such limitations altogether, without sacrificing statistical guarantees or empirical performances.

# Distributionally Robust Policy Evaluation and Optimization

---

In this chapter we present an alternative formulation to the CRM principle by resorting to the distributionally robust optimization (DRO) framework, a generalization of the empirical likelihood (EL) approach. We begin by a brief presentation of generalized empirical likelihood approaches and remind some important asymptotic properties. In particular, it provides an alternative way to compute variance-sensitive confident upper-bounds over unknown quantities. We apply this principle to the problem of offline policy evaluation and optimization. For policy evaluation, we show empirically that the resulting (asymptotic) confidence intervals are tighter than the baseline methods, while still providing sufficient coverage. We also show that when applied to policy optimization, the DRO principle leads to *convex* objectives (w.r.t the policy) that are amenable to stochastic optimization, henceforth circumventing the limitations of the CRM objective. We discuss efficient implementations of the resulting algorithms, and their (approximate) automatic calibrations by resorting to asymptotic arguments. Finally, we display promising numerical experiments, validating the benefits of this alternative to the original CRM objective.

## Outline

---

7.1	Distributionally Robust Optimization . . . . .	139
7.1.1	High-level presentation . . . . .	139
7.1.2	Asymptotic guarantees of generalized empirical likelihood estimators . . .	141
7.2	Application to offline policy evaluation . . . . .	142
7.2.1	Asymptotic confidence interval on policy risk . . . . .	142
7.2.2	Computing the confidence interval . . . . .	142
7.2.3	Numerical simulations . . . . .	144
7.3	Distributionally robust policy optimization algorithms . . . . .	145
7.3.1	An approximation for Kullback-Leibler ambiguity sets . . . . .	146
7.3.2	Robust policy optimization for coherent $f$ -divergences . . . . .	150
7.3.3	Extensions . . . . .	152

---

## 7.1 Distributionally Robust Optimization

In this section we temporarily abandon our notations and discussions from the previous chapter in order to provide a brief introduction to the DRO framework. This is a vast topic which has generated a wide stream of research; covering it exhaustively is out of the scope of the section and we therefore focus on a few principles that we will apply to the offline policy evaluation and optimization tasks. We refer the interested reader to (Rahimian and Mehrotra, 2019) for a detailed survey on DRO.

Throughout this entire section, we consider  $z$  a real-valued random-variable following a distribution  $P_\star$  and such that  $\zeta = \mathbb{E}_{P_\star}[z]$ . Further, we let  $(z_1, \dots, z_n)$  be  $n$  independent realizations of  $z$  and denote  $\hat{P}_n$  their empirical distribution - *i.e.* such that  $\hat{\zeta}_n = \frac{1}{n} \sum_{i=1}^n z_i$  can be rewritten as  $\hat{\zeta}_n = \mathbb{E}_{\hat{P}_n}[z]$ .

### 7.1.1 High-level presentation

**Empirical distribution and ambiguity sets.** Being a sum of *i.i.d* the sample-average estimator  $\hat{\zeta}_n$  fits the usual requirements for the development of “classical” confidence intervals, such as the ones we discussed in the last chapter. One can follow a different rationale and instead try to construct uncertainty sets directly over the data generating process, centered around the empirical distribution  $\hat{P}_n$ . While  $\hat{P}_n$  might provide only an *incomplete* description of  $P_\star$  in the finite sample regime, we can expect it to converge (in some sense) to  $P_\star$ . In particular, for a large enough number of samples we can expect both distribution to be close; if we can quantify how close, then we can design a confidence region for  $P_\star$ . Therefore, by treating the empirical distribution with *skepticism*, we can consider a whole set of plausible distributions, which may contain  $P_\star$ . Informally, let  $d(\cdot|\cdot)$  be a measure of distance between probability distributions and define the *ambiguity* set at level  $\varepsilon > 0$ :

$$\mathcal{U}_\varepsilon(\hat{P}_n) := \left\{ P \text{ s.t. } d(P|\hat{P}_n) \leq \varepsilon \right\} .$$

Through these ambiguity sets we can design a new family of estimators for  $\zeta$ , by replacing  $\hat{P}_n$  by another plausible distribution  $P \in \mathcal{U}_\varepsilon(\hat{P}_n)$ . In particular, we can consider extremal plausible distributions to obtain optimistic or pessimistic estimators for  $\zeta$ . This will lead to the definition of so-called *generalized empirical likelihood* confidence regions for  $\zeta$ .

**Robustifying uncertain optimization problem.** This general idea can be used to *robustify* any optimization problem which involves randomness. Indeed, consider the population objective: minimize  $\theta \in \Theta \mathbb{E}_{P_\star}[\ell(\theta, z)]$  where  $\ell : \Theta \times \mathbb{R} \rightarrow \mathbb{R}$ . Its sample-average version can be robustified using ambiguity sets. This is the main intuition behind the DRO principle, which advocates for solving instead:

$$\underset{\theta \in \Theta}{\text{minimize}} \quad \max_{P \in \mathcal{U}_\varepsilon(\hat{P}_n)} \mathbb{E}_P[\ell(\theta, z)] . \quad (\text{DRO})$$

The main motivation for this objective is to ensure the robustness of the returned solution with respect to the randomness in the empirical distribution. In particular, as soon as  $P_\star \in \mathcal{U}_\varepsilon(\hat{P}_n)$  then the robust estimator provides a *performance-certificate* for the true performance. Indeed, in this case the robust objective provides an upper-bound on the true performance;

$$P_\star \in \mathcal{U}_\varepsilon(\hat{P}_n) \implies \max_{P \in \mathcal{U}_\varepsilon(\hat{P}_n)} \mathbb{E}_P[\ell(\theta, z)] \geq \mathbb{E}_{P_\star}[\ell(\theta, z)] .$$

Therefore, the main goal behind the design of the uncertainty sets  $\mathcal{U}_\varepsilon(\hat{P}_n)$  is to ensure that it contains  $P_\star$  (with high probability) while still being tight enough to provide a meaningful performance-certificate.

Divergence	$\varphi(t)$	$d_\varphi(Q\ P)$	$\varphi^*(s)$
Kullback-Leibler	$\varphi_1(t) = t \log t - t + 1$	$\sum_{i=1}^n q_i \log(q_i/p_i)$	$e^s - 1$
Chi-Square	$\varphi_2(t) = (t - 1)^2$	$\sum_{i=1}^n \frac{(q_i - p_i)^2}{p_i}$	$\begin{cases} s + s^2/4 & s \geq -2 \\ -1 & s \leq -2 \end{cases}$
Burg entropy	$\varphi_3(t) = -\log t + t - 1$	$\sum_{i=1}^n p_i \log(p_i/q_i)$	$-\log(1 - s), s < 1$
Hellinger distance	$\varphi_4(t) = (\sqrt{t} - 1)^2$	$\sum_{i=1}^n (\sqrt{p_i} - \sqrt{q_i})^2$	$\frac{s}{1-s}, s \leq 1$

Table 7.1: Some coherent  $\varphi$ -divergences and their characterizations. In this table we assume that  $Q$  and  $P$  are distributions on the  $n$ -dimensional simplex  $\Delta_n$  and can therefore be associated with vectors  $q, p \in [0, 1]^n$ , respectively. The notation  $\varphi^*$  refers to the convex conjugate of  $\varphi$ .

**$\varphi$ -divergence ambiguity sets.** The design of the uncertainty sets is imposed by the discrepancy function  $d(\cdot\|\cdot)$  used to measure distances between probability distributions. An important part of the DRO literature has focused on Wasserstein distances for generating ambiguity sets - see (Kuhn et al., 2019) for a recent survey. In this work we will focus on another family of ambiguity sets, generated by  $\varphi$ -divergences. This choice is motivated by the desirable theoretical guarantees of such sets, which will be detailed in the following section. We remind below the definition of  $\varphi$ -divergences.

**Definition 7.1.1** ( $\varphi$ -divergence). *Let  $\varphi$  be a real-valued, convex function such that  $\varphi(1) = 0$ . Let  $P$  and  $Q$  be two probability distributions over a space  $\Omega$ . The divergence of  $Q$  with respect to  $P$  is defined by:*

$$d_\varphi(Q\|P) \triangleq \begin{cases} \int_\Omega \varphi(dQ/dP)dP, & \text{if } Q \ll P, \\ +\infty, & \text{else.} \end{cases}$$

In particular, we will focus on  $\varphi$ -divergences which satisfy so-called *coherence* properties, studied extensively by (Rockafellar, 2017). To this end, we make the following smoothness assumption on  $\varphi$  in order to narrow the space of information divergences we consider.

**Assumption 7.1.1** (Coherence).  *$\varphi$  is a real-valued function satisfying:*

- $\varphi$  is convex and lower-semi-continuous.
- For all  $t > 0$  we have  $\varphi(t) \geq \varphi(1) = 0$ .
- $\varphi$  is twice continuously differentiable at  $t = 1$  with  $\dot{\varphi}(1) = 0$  and  $\ddot{\varphi}(1) > 0$ .

Common divergences associated to coherent functions includes for instance the Kullback-Leibler and Chi-Square divergences, the Burg entropy and the squared Hellinger distance. We provide in Table 1 their associated function  $\varphi$  and closed-form expressions for the divergences on the simplex. Equipped with such definition, we will use the affiliated ambiguity sets:

$$\mathcal{U}_\varepsilon^\varphi(\hat{P}_n) := \left\{ P \text{ s.t. } d_\varphi(P\|\hat{P}_n) \leq \varepsilon \right\}.$$

One can note that a consequence of **Definition 7.1.1** is that any  $P \in \mathcal{U}_\varepsilon^\varphi(\hat{P}_n)$  must be absolutely continuous w.r.t  $\hat{P}_n$ . Because  $\hat{P}_n$  has a finite support (it is supported by at most  $n$  atoms) this directly implies that  $P$  has the same finite support - in others words both  $\hat{P}_n$  and any  $P \in \mathcal{U}_\varepsilon^\varphi(\hat{P}_n)$  live in  $\Delta_n$ , the  $n$ -dimensional simplex. In the following, we will therefore confuse  $\hat{P}_n$  and  $P$  with their associated vectors in  $[0, 1]^n$  which will be denoted by lower-case symbols - for instance,  $p_n := (1/n)1_n$  and  $p \in \Delta_n$ .

### 7.1.2 Asymptotic guarantees of generalized empirical likelihood estimators

We hereinafter present some asymptotic guarantees of mean estimators based on coherent  $\varphi$ -divergences ambiguity sets. We will for now limit ourselves to blunt statements and their immediate consequences; we leave the discussion to the following section, where we will investigate their meanings under the lens of our original problem - *i.e* offline policy evaluation. All results presented in this section are from [Duchi et al. \(2016\)](#).

**Generalized empirical likelihood confidence interval.** The first result concerns the design of confidence regions for means based on generalized empirical likelihood estimators.

**Proposition 7.1.1** (Proposition 1 of [Duchi et al. \(2016\)](#)). *Let  $\varphi$  a function satisfying [Assumption 7.1.1](#) and assume that  $z$  has a finite second-order moment. Let  $\rho > 0$ ,  $\varepsilon = \check{\varphi}(1)\rho/(2n)$  and  $y$  be a standard normal random variable. Then:*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left( \zeta \in \left\{ \mathbb{E}_P[z], P \in \mathcal{U}_\varepsilon^\varphi(\hat{P}_n) \right\} \right) = \mathbb{P} \left( y^2 \leq \rho \right) .$$

[Proposition 7.1.1](#) namely asserts the following; if  $\delta \in (0, 1]$  then by setting the ambiguity level to  $\varepsilon = \rho_\delta/n$  where  $\rho_\delta := \chi_{1,1-\delta}^2$  is the  $1 - \delta$  chi-square quantile, we obtain an asymptotic confidence interval for  $\zeta$  with exact coverage  $1 - \delta$ :

$$\lim_{n \rightarrow \infty} \mathbb{P} \left( \zeta \in \left[ \inf_{\mathcal{U}_\varepsilon(\hat{P}_n)} \mathbb{E}_P[z], \sup_{\mathcal{U}_\varepsilon(\hat{P}_n)} \mathbb{E}_P[z] \right] \right) = 1 - \delta .$$

**Remark** (Alternative approach). While [Proposition 7.1.1](#) aims to directly give a confidence region for the mean  $\zeta$  based on  $\mathcal{U}_\varepsilon^\varphi(\hat{P}_n)$ , an alternative strategy could be to control the probability of the event  $P_\star \in \mathcal{U}_\varepsilon^\varphi(\hat{P}_n)$  which can be done under mild assumptions. For instance, under the assumption that  $P_\star$  has a finite support of cardinality  $m$  this can be done thanks to ([Pardo, 2018](#), Corrolary 3.1) which asserts that the normalized empirical  $\varphi$ -divergence  $\frac{2n}{\check{\varphi}(1)} d_\varphi(P_\star \| \hat{P}_n)$  asymptotically follows a  $\chi_{m-1}^2$  distribution. This would result in the asymptotic following  $1 - \delta$  confidence region for  $P_\star$ :

$$\left\{ P \text{ s.t. } d_\varphi(P \| \hat{P}_n) \leq \frac{\check{\varphi}(1)}{2n} \chi_{m-1,1-\delta}^2 \right\}$$

This yields an asymptotic confidence interval similar to the one presented in [Proposition 7.1.1](#), yet much larger and more restrictive because of the added dependency in  $m$ .

**Variance sensitivity.** The second result details the asymptotic behavior of the generalized empirical likelihood confidence intervals, and ties it to the empirical variance of the sample-average estimator.

**Proposition 7.1.2** (Lemma 1 of [Duchi et al. \(2016\)](#)). *Let  $\varphi$  a function satisfying [Assumption 7.1.1](#) and assume that  $z$  has a finite second-order moment. Let  $\rho > 0$ ,  $\varepsilon = \rho/n$  and denote  $s_n^2 := \mathbb{E}_{\hat{P}_n}[z^2] - \mathbb{E}_{\hat{P}_n}[z]^2$ . Then:*

$$\begin{aligned} \sup_{P \in \mathcal{U}_\varepsilon^\varphi(\hat{P}_n)} \mathbb{E}_P[z] &= \hat{\zeta}_n + \sqrt{\frac{\rho}{n} s_n^2} + \frac{\varepsilon_n^+}{\sqrt{n}} , \\ \inf_{P \in \mathcal{U}_\varepsilon^\varphi(\hat{P}_n)} \mathbb{E}_P[z] &= \hat{\zeta}_n - \sqrt{\frac{\rho}{n} s_n^2} + \frac{\varepsilon_n^-}{\sqrt{n}} . \end{aligned}$$

where  $\lim_{n \rightarrow \infty} \varepsilon_n^+ = \lim_{n \rightarrow \infty} \varepsilon_n^- = 0$  almost surely.

## 7.2 Application to offline policy evaluation

We will now discuss how such guarantees can be applied to the offline policy evaluation task and why they constitute good alternatives over baseline approaches. In all the following we let  $\hat{p}_n$  represents the empirical distribution of the observed counterfactual costs  $\{\omega^\pi(x_i, a_i)c(x_i, a_i)\}_{i=1}^n$ . As discussed earlier, it will be confused with the  $n$ -dimensional vector  $(1/n)\mathbf{1}_n \in \Delta_n$ .

### 7.2.1 Asymptotic confidence interval on policy risk

Let us fix a policy  $\pi$  and let  $\varphi$  be a function satisfying [Assumption 7.1.1](#). Define the upper  $\varepsilon$ -distributionally robust risk of  $\pi$  as follows:

$$\text{RobustRisk}_n^{\varphi, \text{up}}(\pi, \varepsilon) := \sup_{p \in \Delta_n} \left\{ \sum_{i=1}^n p_i \omega^\pi(x_i, a_i) c(x_i, a_i) \quad \text{s.t.} \quad d_\varphi(p, \hat{p}_n) \leq \varepsilon \right\}, \quad (7.1)$$

as well as its lower counterpart;

$$\text{RobustRisk}_n^{\varphi, \text{low}}(\pi, \varepsilon) := \inf_{p \in \Delta_n} \left\{ \sum_{i=1}^n p_i \omega^\pi(x_i, a_i) c(x_i, a_i) \quad \text{s.t.} \quad d_\varphi(p, \hat{p}_n) \leq \varepsilon \right\}. \quad (7.2)$$

Note that in both [Eqs. \(7.1\) and \(7.2\)](#) the quantity  $\sum_{i=1}^n p_i \omega^\pi(x_i, a_i) c(x_i, a_i)$  is an expectation w.r.t a distribution  $P$ , absolutely continuous w.r.t  $\hat{P}_n$  and therefore represented by its corresponding vector  $p \in \Delta_n$ . Under the assumption that the counterfactual weights are bounded, [Proposition 7.1.1](#) applies to the random variable  $\omega^\pi(x, a)c(x, a)$  and asserts that the following interval is an asymptotic confidence region with coverage  $1 - \delta$ :

$$\mathcal{I}_n^\varphi(\pi) := \left[ \text{RobustRisk}_n^{\varphi, \text{low}}\left(\pi, \frac{\varepsilon_\delta}{n}\right), \text{RobustRisk}_n^{\varphi, \text{up}}\left(\pi, \frac{\varepsilon_\delta}{n}\right) \right],$$

where  $\varepsilon_\delta = \check{\varphi}(1)\chi_{1-\delta}^2/2$ . Furthermore, [Proposition 7.1.2](#) informs us that this confidence interval is sensitive to the empirical variance behind the policy  $\pi$ ; for instance:

$$\text{RobustRisk}_n^{\varphi, \text{up}}(\pi, \varepsilon) = \hat{R}_n^{\text{IPS}}(\pi) + \sqrt{\frac{\rho s_n^2(\pi)}{n}} + o\left(\frac{1}{\sqrt{n}}\right). \quad (7.3)$$

As we discussed earlier, this is a particularly enjoyable feature for offline policy evaluation - this alone was enough to motivate the use of empirical Bernstein based confidence intervals for the CRM principle. We will investigate the use of  $\mathcal{I}_n^\varphi$  for policy optimization in [Section 7.3](#). In the rest of this section, we focus on the computation and the finite-sample performances of this asymptotic confidence interval.

### 7.2.2 Computing the confidence interval

We now turn to the computation of the confidence interval  $\mathcal{I}_n^\varphi(\pi)$ . We first focus on the upper-bound  $\text{RobustRisk}_n^{\varphi, \text{up}}(\pi, \varepsilon)$  which requires finding the most *pessimistic* distribution inside the ambiguity set:

$$\arg \max_{p \in \Delta_n} \left\{ \sum_{i=1}^n p_i \omega^\pi(x_i, a_i) c(x_i, a_i) \quad \text{s.t.} \quad d_\varphi(p, \hat{p}_n) \leq \varepsilon \right\} \quad (7.4)$$

Note that the objective is linear and therefore convex with respect to the optimization variable  $p$ . Furthermore, the constraint set  $\{p \in \Delta_n, d_\varphi(p, \hat{p}_n)\}$  is convex. This makes [Eq. \(7.4\)](#) a *convex*

program which can therefore be solved efficiently. The precise nature of the primal formulation (e.g. quadratic programming) depends on which function  $\varphi$  is used. Note that the primal can be challenging to solve directly; its dimension is  $n$ , the size of the historical dataset  $\mathcal{D}_n$  which is likely to be prohibitively large. Therefore, we will essentially rely on the dual formulation; it is easily solvable, even in regimes where  $n$ , the amount of data, is large. Formally, we rely on the following result to characterize the robust risk. The notation  $\varphi^*$  refers to the convex conjugate of  $\varphi$ ; it is formally defined  $\varphi^*(s) := \sup_{t \in \mathbb{R}} (st - \varphi(t))$ . It can be computed in closed form for many of the functions  $\varphi$  we consider - cf. Table 7.1. This robust program characterization can be extracted from more general results - see for instance (Ben-Tal et al., 2013, Section 4).

**Proposition 7.2.1.** [Dual program for computing the robust risk] Define:

$$g_\varphi^\pi(\beta, \gamma, \varepsilon) := \beta + \gamma\varepsilon + \frac{1}{n} \sum_{s=1}^n (\gamma\varphi)^*(\omega^\pi(x_i, a_i)c(x_i, a_i) - \beta), \quad (7.5)$$

where  $(\gamma\varphi)^*(s) = \gamma\varphi^*(s/\gamma)$ , with the convention that  $(0\varphi)^*(s) = +\infty$  if  $s > 0$  and 0 otherwise. For any  $\pi$  and  $\varepsilon \geq 0$  the function  $(\beta, \gamma) \mapsto g_\varphi^\pi(\beta, \gamma, \varepsilon)$  is convex and:

$$\text{RobustRisk}_n^{\varphi, \text{up}}(\pi, \varepsilon) = \inf_{\beta, \gamma \geq 0} g_\varphi^\pi(\beta, \gamma, \varepsilon). \quad (7.6)$$

*Proof.* From the (upper) robust risk definition we obtain:

$$\text{RobustRisk}_n^{\varphi, \text{up}}(\pi, \varepsilon) = \sup_{p \in \mathcal{Q}_\varepsilon} \left\{ \sum_{i=1}^n p_i \omega^\pi(x_i, a_i) c(x_i, a_i) \right\}$$

where the set of constraints is:

$$\mathcal{Q}_\varepsilon := \left\{ p \in \mathbb{R}^n \text{ s.t. } \sum_{i=1}^n p_i = 1, \frac{1}{n} \sum_{i=1}^n \varphi(np_i) \leq \varepsilon, p_i \geq 0 \text{ for all } i \in [n] \right\} \quad (7.7)$$

This program is convex; the objective is linear (henceforth convex) and the constraint set  $\mathcal{Q}_\varepsilon$  is convex by convexity of  $\varphi$ . Furthermore if  $\varepsilon > 0$  then  $\hat{p}_n = (1/n)1_n$  is strictly feasible. Therefore Slater's condition hold and the program enjoys strong duality. Writing down its Lagrangian, we obtain the following equivalence:

$$\begin{aligned} \text{RobustRisk}_n^{\varphi, \text{up}}(\pi, \varepsilon) &= \sup_{p_i \geq 0} \inf_{\beta, \gamma \geq 0} \sum_{i=1}^n p_i \omega^\pi(x_i, a_i) c(x_i, a_i) + \beta \left( 1 - \sum_{i=1}^n p_i \right) + \gamma \left( \varepsilon - \frac{1}{n} \sum_{i=1}^n \varphi(np_i) \right) \\ &= \inf_{\beta, \gamma \geq 0} \sup_{p_i \geq 0} \sum_{i=1}^n p_i \omega^\pi(x_i, a_i) c(x_i, a_i) + \beta \left( 1 - \sum_{i=1}^n p_i \right) + \gamma \left( \varepsilon - \frac{1}{n} \sum_{i=1}^n \varphi(np_i) \right) \\ &= \inf_{\beta, \gamma \geq 0} \beta + \gamma\varepsilon + \frac{1}{n} \sum_{i=1}^n \sup_{p_i \geq 0} \{ (np_i) (\omega^\pi(x_i, a_i) c(x_i, a_i) - \beta) - \gamma\varphi(np_i) \} \end{aligned} \quad (7.8)$$

where the first equality is a consequence of strong duality, and the second is obtained through simple re-arranging. If  $\gamma \neq 0$ , factorizing with  $\gamma$  and the change of variable  $p_i \leftarrow np_i$  lead to:

$$\begin{aligned} \text{RobustRisk}_n^{\varphi, \text{up}}(\pi, \varepsilon) &= \inf_{\beta, \gamma \geq 0} \beta + \gamma\varepsilon + \frac{\gamma}{n} \sum_{i=1}^n \sup_{p_i \geq 0} \left\{ p_i \frac{\omega_\pi(x_i, a_i) c(x_i, a_i) - \beta}{\gamma} - \varphi(p_i) \right\} \\ &= \inf_{\beta, \gamma \geq 0} \beta + \gamma\varepsilon + \frac{\gamma}{n} \sum_{i=1}^n \varphi^* \left( \frac{\omega_\pi(x_i, a_i) c(x_i, a_i) - \beta}{\gamma} \right) \end{aligned}$$

by using the definition of  $\varphi^*$ . The precise limit conditions announced in [Proposition 7.2.1](#) are easily checked by computing the dual function when  $\gamma = 0$ . We therefore obtain the equality announced by using the definition of  $g_\varepsilon^\pi$ :

$$\text{RobustRisk}_n^{\varphi, \text{up}}(\pi, \epsilon) = \inf_{\beta, \gamma \geq 0} g_\varepsilon^\pi(\beta, \gamma)$$

The convexity of  $g_\varepsilon^\pi$  can be obtained two ways; by noticing that  $g_\varepsilon^\pi$  is obtained through convexity-transforming transformations of a *perspective* function ([Combettes, 2018](#)), or by noticing thanks to [Eq. \(7.8\)](#) that  $(\beta, \gamma) \mapsto g_\varepsilon^\pi(\beta, \gamma)$  is convex as a sum of supremum of linear (and hence convex) functions. ■

The main implication of [Proposition 7.2.1](#) is that the robust risk can be efficiently computed by solving a two-dimensional convex program. When  $n$  is reasonably small (*e.g.* when the dataset  $\mathcal{D}_n$  fits in memory), coordinate descent (with exact line search) or two-dimensional bisection provide efficient, principled tools for computing the robust risk. The program presented in [Eq. \(7.6\)](#) is also well-suited for the large-data regime (*e.g.* large  $n$ ) as it naturally adapts to stochastic optimization. Indeed, the function  $g_\varphi^\pi$  is *composite* and unbiased gradients of this objective are easily obtainable. Stochastic gradient descent methods therefore provide efficient and flexible solutions for this problem (up to some mild modifications to account for the fact that the  $g$  is not smooth for  $\gamma$  in a neighborhood of 0). A similar characterization of course holds for  $\text{RobustRisk}_n^{\varphi, \text{low}}(\pi, \epsilon)$  - the objective in [Eq. \(7.2\)](#) is a linear function of  $p$  and after flipping the signs of the counterfactual weights the exact same reasoning can be conducted.

### 7.2.3 Numerical simulations

The generalized empirical likelihood confidence interval  $\mathcal{I}_n^\varphi$  are *asymptotic* and therefore might fail in real-life scenarios where  $n$  is finite. In this section, we compare both its coverage and width with the other risk confidence intervals we detailed in [Section 6.2.2](#). For this empirical evaluation we will work with each of the four  $\varphi$ -divergence presented in [Table 7.1](#).

**Methodology.** To create a bandit dataset  $\mathcal{D}_n$ , we follow [Swaminathan and Joachims \(2015a\)](#) and employ the classical supervised to bandit conversion of [Agarwal et al. \(2014\)](#). Formally, denote  $\mathcal{D}_{\text{full}} = \{(x_1, y_1), \dots, (x_M, y_M)\}$  a given *supervised* multi-label dataset (*i.e.* with full information feedback) where  $x \in \mathcal{X}$  is a feature vector and  $y \in \{0, 1\}^L$  its associate label. The logging policy  $\pi_0$  is obtained by training a linear softmax model on a fraction of  $\mathcal{D}_{\text{full}}$ . We then create the historic data  $\mathcal{D}_n$  by repeating  $P$  (the *replay count*) times the following procedure: for every  $(x_i, y_i) \in \mathcal{D}_{\text{full}}$ , sample  $a_i \sim \pi_0(x)$  and log the cost  $c(x_i, a_i) = \|a_i - y_i\|_1$ . The effective size of the bandit dataset is therefore  $n = P|\mathcal{D}_{\text{full}}|$ . The policies which are up for evaluation are also based on a linear softmax model, and are obtained by training over a random subset of  $\mathcal{D}_{\text{full}}$  - however different from the subset used to obtain the initial  $\pi_0$ . The confidence intervals are computed by solving the dual formulation of [Proposition 7.2.1](#) using `scipy`<sup>1</sup> off-the-shelf optimizer through the `minimize` procedure.

**Results.** We present in [Figs. 7.1](#) and [7.2](#) the empirical coverage and width results when applying this methodology to two multi-label supervised dataset: the Yeast dataset and the Scene dataset, both taken from the LibSVM repository and standard for the policy optimization task ([Swaminathan and Joachims, 2015a,b](#)). The empirical coverage and confidence interval width are reported for increasing values of the replay count  $P$  (or equivalently for increasing values of the historic data size  $n$ ) and aggregated over different random realizations of the bandit dataset

<sup>1</sup><https://www.scipy.org/index.html>

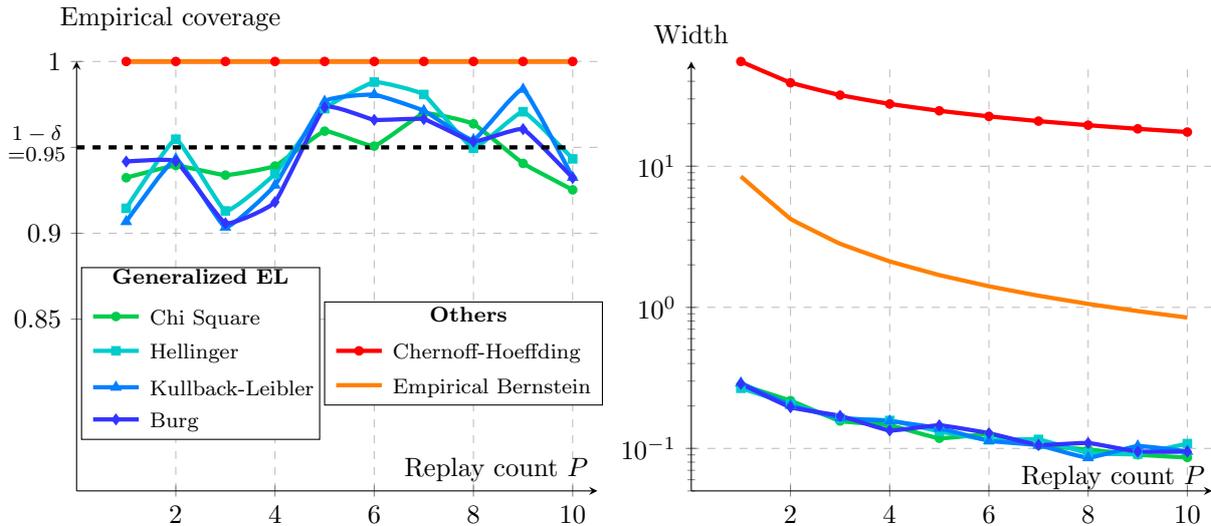


Figure 7.1: Empirical coverage and confidence interval width after a bandit conversion from the Yeast dataset, as a function of the replay count  $P = n/|\mathcal{D}_{\text{full}}|$ . The generalized EL confidence intervals  $\mathcal{I}_n^\varphi$  provide satisfactory coverage, close to the required  $1 - \delta = 0.95$  confidence level required here. As often, the worst-case finite-time confidence intervals  $\mathcal{I}_n^{\text{CH}}$  and  $\mathcal{I}_n^{\text{EB}}$  over-cover and are much wider than asymptotic counterparts.

$\mathcal{D}_n$ . We observe that for the policy evaluation task, the generalized empirical likelihood confidence intervals  $\mathcal{I}_n^\varphi$  provide almost exact  $(1 - \delta)$  coverage and are therefore safe to use, even in the small data regime. As a side comment, we observe that all four considered  $\varphi$ -divergence lead to very similar empirical results. Finally, we can check experimentally that as expected, those asymptotic confidence intervals are by orders of magnitude smaller than the worst-case, finite-time ones such as  $\mathcal{I}_n^{\text{CH}}$  and  $\mathcal{I}_n^{\text{EB}}$ .

### 7.3 Distributionally robust policy optimization algorithms

We saw in the last section that generalized empirical likelihood confidence intervals provides a trust-worthy alternative for confident policy evaluation. Also, one of their salient feature is to produce a *variance-sensitive* confident upper-bound on the risk - the same characteristics which motivated the use of the empirical Bernstein upper-bound for the CRM principle. With the same rationale, we can therefore use these confident intervals to design a new, distributionally robust, offline policy optimization objective:

$$\hat{\pi}_n \in \arg \min_{\pi} \text{RobustRisk}_n^{\varphi, \text{up}}(\pi, \varepsilon), \quad (\text{DRO-CRM})$$

This section investigates the empirical benefits behind this alternative approach. First, we show in [Section 7.3.1](#) how it leads to obtain an exponential-weights version of the CRM objective, by relying on an approximate closed form of the robust risk when defined by the Kullback-Leibler divergence. This new objective displays improved empirical performances over the original CRM objective; yet, it suffers from similar downsides (*e.g.* non-convex, not suited for stochastic optimization). We tackle such limitations in [Section 7.3.2](#) by relying on the robust risk's dual formulation from [Proposition 7.2.1](#). This gives rises to better behaved optimization objectives and empirically superior policy optimization algorithms.

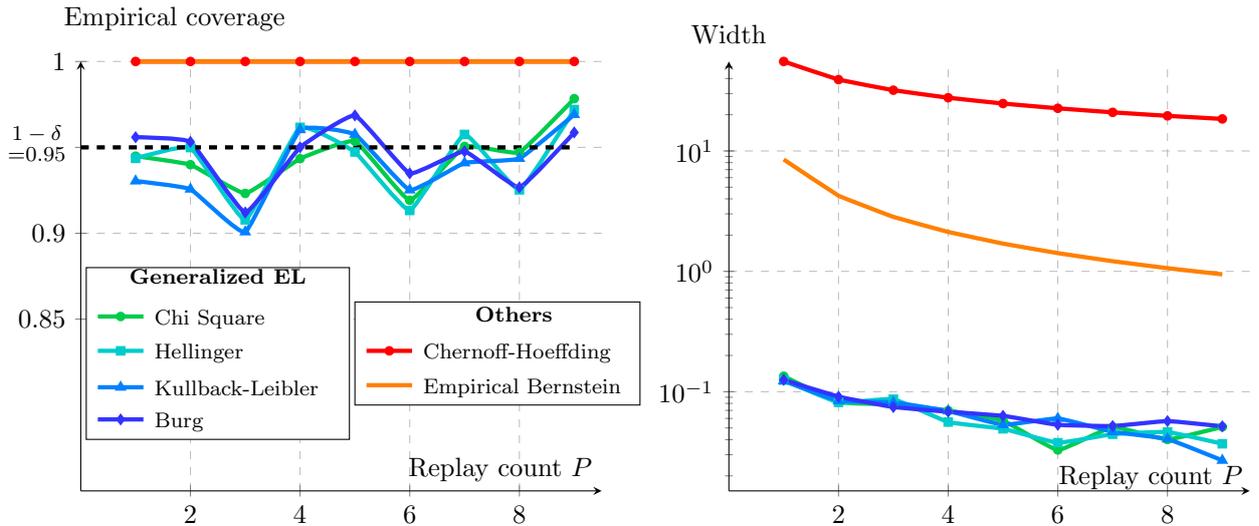


Figure 7.2: Empirical coverage and confidence interval width after a bandit conversion from the Scene dataset, as a function of the replay count  $P = n/|\mathcal{D}_{\text{full}}|$ . The generalized EL confidence intervals  $\mathcal{I}_n^\varphi$  provide satisfactory coverage, close to the required  $1 - \delta = 0.95$  confidence level required here. As often, the worst-case finite-time confidence intervals  $\mathcal{I}_n^{\text{CH}}$  and  $\mathcal{I}_n^{\text{EB}}$  over-cover and are much wider than asymptotic counterparts.

**Remark** (Equivalence with the original CRM objective). *The asymptotic expansion of Eq. (7.3) ensures that (DRO-CRM) and (CRM) are asymptotically equivalent, for any  $\varphi$  satisfying Assumption 7.1.1. Actually, this equivalence holds even non-asymptotically. Under mild conditions, the robust risk generated by the Chi-Square divergence yields exactly the CRM objective (Fauray et al., 2020b, Lemma 3). Therefore, (DRO-CRM) can be understood as a strict generalization of the (CRM) approach.*

### 7.3.1 An approximation for Kullback-Leibler ambiguity sets

**Kullback-Leibler robust risk.** In this section, we are interested in the robust risk generated by Kullback-Leibler divergence ambiguity sets, that is with  $\varphi_1(t) = t \log(t) - t + 1$ . For such ambiguity sets, alternative characterization of the robust risk can easily be derived, which subsequently yields a closed-form objective for offline policy optimization. Such a characterization is made explicit in the following proposition, through exponential weights which re-balance the original IPS objective.

**Proposition 7.3.1** (Kullback-Leibler robust risk). *If  $\varphi = \varphi_1$  then the robust risk writes:*

$$\text{RobustRisk}_n^{\varphi, \text{up}}(\pi, \varepsilon) = \sum_{i=1}^n \omega^\pi(x_i, a_i) c(x_i, a_i) \frac{\exp(\omega^\pi(x_i, a_i) c(x_i, a_i) / \gamma_\varepsilon)}{\sum_{j=1}^n \exp(\omega^\pi(x_j, a_j) c(x_j, a_j) / \gamma_\varepsilon)}, \quad (7.9)$$

where the temperature  $\gamma_\varepsilon$  is a solution of the fixed-point equation:

$$\gamma \sum_{i=1}^n \exp(\omega^\pi(x_i, a_i) c(x_i, a_i) / \gamma) = \frac{\sum_{i=1}^n \omega^\pi(x_i, a_i) c(x_i, a_i) \exp(\omega^\pi(x_i, a_i) c(x_i, a_i) / \gamma)}{\varepsilon + \log(\sum_{i=1}^n \exp(\omega^\pi(x_i, a_i) c(x_i, a_i) / \gamma) / n)}.$$

*Proof.* From Proposition 7.2.1 we have that:

$$\text{RobustRisk}_n^{\varphi_1, \text{up}}(\pi, \varepsilon) = \inf_{\gamma \geq 0} \left\{ \gamma \varepsilon + \inf_{\beta} \left( \beta + \frac{1}{n} \sum_{i=1}^n (\gamma \varphi^*)(\omega^\pi(x_i, a_i) c(x_i, a_i) - \beta) \right) \right\}.$$

Straight-forward computation easily yield that  $\varphi_1^*(s) = e^s - 1$ . Let us assume for now that  $\gamma > 0$ . We then obtain:

$$\text{RobustRisk}_n^{\varphi_1, \text{up}}(\pi, \varepsilon) = \inf_{\gamma \geq 0} \left\{ \gamma \varepsilon + \inf_{\beta} \left( \beta + \frac{\gamma}{n} \sum_{i=1}^n \varphi^* \left( \frac{\omega^\pi(x_i, a_i) c(x_i, a_i) - \beta}{\gamma} \right) \right) \right\}.$$

Solving the inner minimization yields:

$$\text{RobustRisk}_n^{\varphi_1, \text{up}}(\pi, \varepsilon) = \inf_{\gamma \geq 0} \left\{ \gamma \varepsilon + \gamma \log \left( \frac{1}{n} \sum_{i=1}^n \exp(\omega^\pi(x_i, a_i) c(x_i, a_i) / \gamma) \right) \right\}.$$

Solving this equation by differentiating the r.h.s and setting it to 0 indicates that the optimum is a solution of the fixed-point equation stated in [Proposition 7.3.1](#). Replacing it in the previous line gives the announced closed form formula for the robust risk when  $\varphi = \varphi_1$ . To conclude the proof, we need to rule out the case  $\gamma = 0$ ; this is done by computing the optimal value of the objective in this case, which yields a larger quantity than when  $\gamma > 0$ . ■

The identity given by [Proposition 7.3.1](#) suggests the idea of an exponentially-weighted version of the original CRM principle. Indeed one can consider the following policy optimization strategy:

$$\hat{\pi}_n \in \arg \min_{\pi} \left\{ \sum_{i=1}^n \omega^\pi(x_i, a_i) c(x_i, a_i) \frac{\exp(\omega^\pi(x_i, a_i) c(x_i, a_i) / \gamma)}{\sum_{j=1}^n \exp(\omega^\pi(x_j, a_j) c(x_j, a_j) / \gamma)} \right\} \quad (\text{KL-DRO-CRM})$$

where the temperature  $\gamma$  is treated as an hyper-parameter, which optimal value is to be determined through cross-validation. This strategy was followed in [Fauray et al. \(2020b\)](#) and proves to be competitive with the CRM objective. It however can be improved, by leveraging the fact that one can obtain a good approximation of the optimal value  $\gamma_\varepsilon$ , provided in the following Lemma.

**Lemma 7.3.1.** *The optimal value  $\gamma_\varepsilon$  of the temperature can be approximated as follows:*

$$\gamma_\varepsilon \approx \sqrt{\frac{s_n^2(\pi)}{2\varepsilon}}$$

*Proof.* It can be extracted from the proof of [Proposition 7.3.1](#) that:

$$\text{RobustRisk}_n^{\varphi_1, \text{up}}(\pi, \varepsilon) = \inf_{\gamma \geq 0} \left\{ \gamma \varepsilon + \gamma \log \left( \frac{1}{n} \sum_{i=1}^n \exp(\omega^\pi(x_i, a_i) c(x_i, a_i) / \gamma) \right) \right\}.$$

A second order asymptotic expansion around  $\gamma \rightarrow \infty$  yields:

$$\text{RobustRisk}_n^{\varphi_1, \text{up}}(\pi, \varepsilon) \approx \inf_{\gamma \geq 0} \left\{ \gamma \varepsilon + \hat{R}_n^{\text{IPS}} + \frac{s_n^2(\pi)}{2\gamma} \right\}.$$

Solving this quadratic minimization yields the announced result. ■

**Algorithm.** In light of [Proposition 7.3.1](#), we propose to minimize the (KL-DRO-CRM) objective (*e.g* by an iterative algorithm) while maintaining a decent approximation of the optimal temperature  $\gamma_\varepsilon$  - as provided by [Lemma 7.3.1](#). More precisely, at every step taken by an iterative algorithm, the current approximation for  $\gamma_\varepsilon$  is updated based on the current policy. We illustrate this idea in [Algorithm 9](#) with the Kullback-Leibler DRO equivalent of the POEM algorithm with projected gradient descent. It remains to set the value of  $\varepsilon$ . For now, we treat as an hyper-parameter, either set by the practitioner or determined through cross-validation.

---

**Algorithm 9** KL-DRO-CRM for exponential distribution with projected gradient descent

---

**input:** Parameter space  $\Theta$ , ambiguity level  $\varepsilon$ , clipping constant  $M$ , optimization horizon  $T$ , learning rate  $\alpha$ , initial parameter  $\theta_1$ , static bandit dataset  $\mathcal{D}_n = \{x_i, a_i, c_i, p_i^0\}_{i=1}^n$ .

**for**  $t \in [1, T - 1]$  **do**

    Compute likelihood scores:  $\{p_i(\theta_t) \leftarrow \exp(\theta_t^\top \phi(x_i, a_i)) / \sum_{a \in [K]} \exp(\theta_t^\top \phi(x_i, a_i))\}_{i=1}^n$ .

    Compute clipped propensity weights:  $\{\omega_i(\theta_t) \leftarrow \max(M, p_i(\theta_t) / p_i^0)\}_{i=1}^n$ .

    Compute the empirical variance:  $s_n^2(\theta_t) \leftarrow \frac{1}{n(n-1)} \sum_{i=1}^n (\omega_i(\theta_t) c_i - \omega_j(\theta_t) c_j)^2$ .

    Compute the current temperature  $\gamma_\varepsilon = \sqrt{s_n^2(\theta_t) / (2\varepsilon)}$ .

    Compute the KL robust risk:

$$\text{RobustRisk}_n^{\varphi^1, \text{up}}(\theta_t, \varepsilon) = \sum_{i=1}^n \omega_i(\theta_t) c_i \frac{\exp(\omega_i(\theta_t) c_i / \gamma_\varepsilon)}{\sum_{j=1}^n \exp(\omega_j(\theta_t) c_j / \gamma_\varepsilon)}$$

    Perform a gradient descent step:

$$\tilde{\theta}_{t+1} \leftarrow \theta_t - \alpha \nabla_{\theta_t} (\text{RobustRisk}_n^{\varphi^1, \text{up}}(\theta_t, \varepsilon)) .$$

    Project back to  $\Theta$ :  $\theta_{t+1} \leftarrow \arg \min_{\theta \in \Theta} \|\theta - \tilde{\theta}_{t+1}\|^2$ .

**end for**

**return** policy  $\pi_{\theta_T}$ .

---

**Empirical evaluation.** The methodology used to evaluate KL-DRO-CRM is similar to the one detailed in [Section 7.2.3](#) and follows the experimental procedure introduced in [Swaminathan and Joachims \(2015a\)](#).<sup>2</sup> A supervised multi-label dataset  $\mathcal{D}_{\text{full}} = \{(x_1, y_1), \dots, (x_M, y_M)\}$  where  $y \in \{0, 1\}^L$  is split into two parts  $\mathcal{D}_{\text{full}}^{\text{train}}$  and  $\mathcal{D}_{\text{full}}^{\text{valid}}$  as follows: 75% goes to  $\mathcal{D}_{\text{full}}^{\text{train}}$  and 25% to  $\mathcal{D}_{\text{full}}^{\text{valid}}$ . We also assume that some test data  $\mathcal{D}_{\text{full}}^{\text{test}}$  is provided; it will serve for evaluation. Following [Swaminathan and Joachims \(2015a\)](#), we use joint features maps  $\phi(x, y) = xy^\top$  and train a Conditional Random Field ([Lafferty et al., 2001](#)) (CRF) on a fraction (5%, randomly constituted) of  $\mathcal{D}_{\text{full}}^{\text{train}}$ . This CRF has access to the full supervised feedback and plays the role of the logging policy  $\pi_0$ . That is, for every  $x \in \mathcal{D}_{\text{full}}^{\text{train}}$ , a label prediction  $a \in \{0, 1\}^L$  is sampled from the CRF with probability  $\pi_0(x, a)$ . The quality of this prediction is measured through the cost  $c = \|a - y\|_1$ . The logged bandit dataset is generated by running  $\pi_0$  through  $\mathcal{D}_{\text{full}}^{\text{train}}$  for  $P$  times. We train exponential policies with features  $\phi(x, y) = xy^\top$ . After training, the performances of the policy  $\hat{\pi}_n$  returned by the different algorithms are reported as their expected Hamming loss on the held-out set  $\mathcal{D}_{\text{test}}^*$ . Every experiment is run 20 times with a different random seed (which controls the random training fraction for the logging policy and the creation of the bandit dataset). For each dataset we consider (Scene, Yeast, RCV1-Topics and TMC2009, all taken from the LibSVM library) we compare our algorithm with the naive IPS-based approach and the POEM algorithm. For all algorithms, the numerical optimization routine is deferred to the L-BFGS algorithm, which we found to work best in practice. The clipping constant  $M$  is systematically set to the ratio of the 90%ile to the 10%ile of the propensity scores observed in the logged dataset  $\mathcal{D}_n$ . Remaining hyper-parameters ( $\lambda$  for POEM and  $\varepsilon$  for KL-DRO-CRM) are selected by cross-validation based on the smallest value IPS estimator on  $\mathcal{D}_{\text{full}}^{\text{valid}}$ . We also report the performance of the logging policy  $\pi_0$  on the test set as an indicative baseline measure,

---

<sup>2</sup>It is well known that experiments in the field of counterfactual reasoning are highly sensitive to differences in datasets and implementations. To ensure fair comparison, we use the code provided by [Swaminathan and Joachims \(2015a\)](#), available on the authors' website.

	Scene	Yeast	RCV1-Topics	TMC2009
$\pi_0$	1.529	5.542	1.462	3.435
IPS	1.163	4.658	0.930	2.776
POEM	1.157	<b>4.535</b>	0.918	2.191
KL-DRO-CRM	<b>1.128</b>	<b>4.553</b>	<b>0.783</b>	<b>2.126</b>
CRF	0.646	2.817	0.341	1.187

Table 7.2: Expected risk of the policy  $\hat{\pi}_n$  returned by the different algorithms, evaluated on  $\mathcal{D}_{\text{full}}^{\text{test}}$  and averaged over 20 independent runs. Bold font indicates that one or several algorithms are statistically better than the rest, according to a one-tailed paired difference t-test at significance level of 0.95. KL-DRO-CRM is significantly better than its competitors on three out of four datasets, and is competitive with POEM on the last dataset.

and the performance of a skyline CRF trained on the whole supervised dataset, despite its unfair advantage. In Table 7.2 we report performances in terms of the risk of the returned policies, evaluated on the test dataset  $\mathcal{D}_{\text{full}}^{\text{test}}$ . These results highlights that KL-DRO-CRM is a valuable alternative to POEM. For further evaluation, we also report in Table 7.3 the risk of the *greedy* version of the policies  $\hat{\pi}_n$ ; that is:

$$\hat{\pi}_n^\infty(x) := \arg \max_{a \in [K]} \left\{ \exp \left( \hat{\theta}_n^\top \phi(x, a) \right) \right\}$$

The reason for this evaluation is that  $\hat{\pi}_n^\infty$  is much easier to deploy in a real-life situation, as it doesn't require the evaluation of the normalizing constant in Eq. (6.9). The superiority of KL-DRO-CRM is also confirmed in this context, with a greater performance gap over POEM than in the non-greedy evaluation of Table 7.2.

	Scene	Yeast	RCV1-Topics	TMC2009
IPS	1.163	4.369	0.929	2.774
POEM	1.157	<b>4.261</b>	0.918	2.190
KL-DRO-CRM	<b>1.128</b>	<b>4.271</b>	<b>0.779</b>	<b>2.034</b>

Table 7.3: Expected risk of the policy  $\hat{\pi}_n^\infty$  returned by the different algorithms, evaluated on  $\mathcal{D}_{\text{full}}^{\text{test}}$  and averaged over 20 independent runs. Bold font indicates that one or several algorithms are statistically better than the rest, according to a one-tailed paired difference t-test at significance level of 0.95. KL-DRO-CRM is significantly better than its competitors on three out of four datasets, and is competitive with POEM on the last dataset.

Further examination of the performance of KL-DRO-CRM requires inspecting its sample efficiency. In Fig. 7.3 we display the expected risk of  $\hat{\pi}_n$  (returned by either KL-DRO-CRM or POEM) as a function of the replay count  $P$ , which directly controls the size of the bandit dataset  $\mathcal{D}_n$ . As expected given the asymptotical equivalence of the two approaches, KL-DRO-CRM and POEM are equivalent for large values of  $P$  (*i.e*  $n \gg 1$ ). For smaller  $P$  and smaller bandit datasets, we notice that KL-DRO-CRM outperforms POEM by a significant margin.

**Limitations.** Despite displaying enjoyable empirical performances, this approach suffers from similar limitation as the original CRM approach. Indeed, the (DRO-CRM) objective is also non-convex w.r.t the policy  $\pi$  and is not composite (*i.e* does not undergoes classical stochastic optimization strategies). Overcoming such downsides is the goal of the following section.

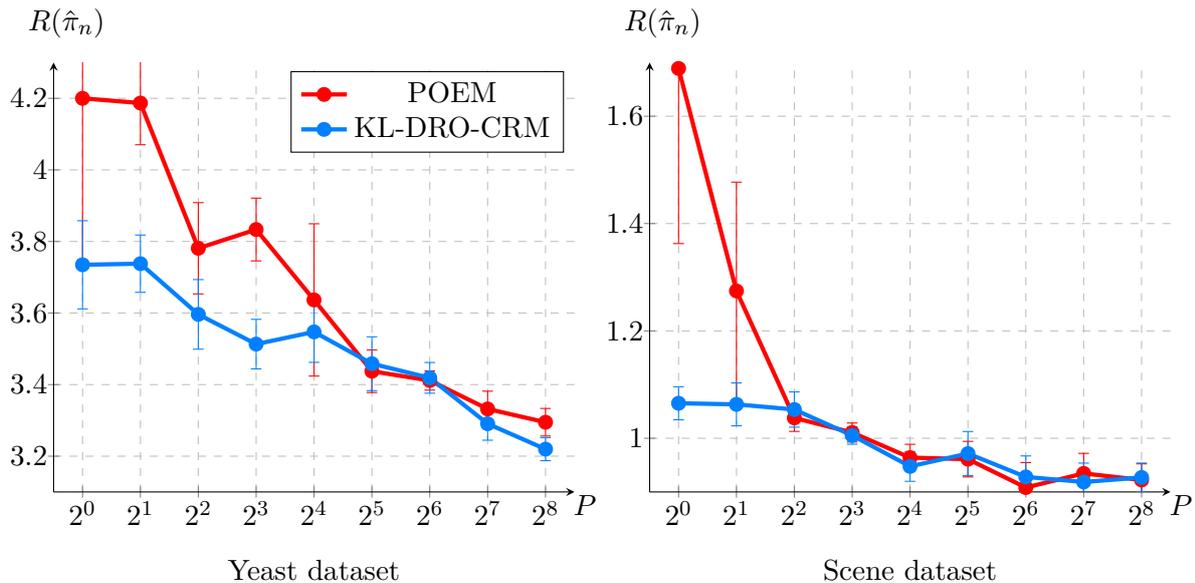


Figure 7.3: Risk of the policy  $\hat{\pi}_n$  returned by KL-DRO-CRM and POEM, as a function of the replay count  $P$ , evaluated on  $\mathcal{D}_{full}^{train}$ . Error bars represent one standard deviation of the risk aggregated over 20 random realization of the logging dataset. The replay count directly impacts the size  $n$  of the bandit dataset  $\mathcal{D}_n$  since  $n = P \times |\mathcal{D}_{full}^{train}|$ . As predicted by [Proposition 7.1.2](#), the two approaches are nearly equivalent for large values of  $n$ ; however, KL-DRO-CRM seems to over-perform POEM for small logged dataset, by yielding policies of much smaller risk.

### 7.3.2 Robust policy optimization for coherent $f$ -divergences

**Training objective.** The leading idea to obtain better-behaved policy optimization objectives is to leverage the general characterization of the robust risk from [Proposition 7.2.1](#), and to solve for the confident upper-bound and the optimal policy *jointly*. In other words, we optimize jointly the following objective over the policy  $\pi$  and the dual variables  $\beta, \gamma$ :

$$\hat{\pi}_n \in \arg \min_{\pi, \beta, \gamma \geq 0} \left\{ g_{\varphi}^{\pi}(\beta, \gamma, \varepsilon) = \beta + \gamma \varepsilon + \frac{1}{n} \sum_{s=1}^n (\gamma \varphi)^* (\omega^{\pi}(x_i, a_i) c(x_i, a_i) - \beta) \right\}. \quad (\varphi\text{-DRO-CRM})$$

Note that this objective is composite and its stochastic gradients can easily be obtained; furthermore, it is convex in  $\pi, \beta$  and  $\gamma$  jointly.

**Proposition 7.3.2.** *The function  $(\pi, \beta, \gamma) \rightarrow g_{\varphi}^{\pi}(\beta, \gamma, \varepsilon)$  is convex.*

*Proof.* From the proof of [Proposition 7.2.1](#) it can be extracted that:

$$g_{\varphi}^{\pi}(\beta, \gamma, \varepsilon) = \beta + \gamma \varepsilon + \frac{1}{n} \sum_{i=1}^n \sup_{p_i \geq 0} \{ (np_i) \omega^{\pi}(x_i, a_i) c(x_i, a_i) - \gamma \varphi(np_i) \}.$$

Remembering that  $\omega^{\pi}$  is linear in  $\pi$  yields that  $g_{\varphi}^{\pi}(\beta, \gamma, \varepsilon)$  is a sum of a linear function and a supremum over linear functions - and therefore is convex.  $\blacksquare$

The takeaway from [Proposition 7.3.2](#) is that the  $(\varphi\text{-DRO-CRM})$  objective can be minimized in principled ways - while enjoying similar guarantees as the original CRM objective, thanks to [Propositions 7.1.1](#) and [7.1.2](#). Intuitively, one can expect such an important transformation of the optimization properties of the policy improvement objective to lead to greater practical performances. The most natural way to solve this new policy improvement objective

is through plain gradient descent. Indeed, a valid strategy consists in feeding gradients of the function  $(\beta, \gamma, \pi) \rightarrow g_\pi(\beta, \gamma)$  to a gradient optimizer. As for the policy evaluation, the  $(\varphi$ -DRO-CRM) objective is particularly adapted when the historic data is large (i.e  $n \gg 1$ ) and only stochastic gradients can be processed. Note that stochastic gradients methods might encounter some issues (linked to the possible unboundedness of the gradients) which can be alleviated thanks to specialized methods based on mirror descent (Namkoong and Duchi, 2016). We found in our experiments that such problem did not arise in practice.

**Remark 7.3.1** (Loss of convexity for parametrized policies). *As anticipated in Remark 6.3.2, we rarely directly optimize over all policies; rather, it is more usual to tie the space of candidate policy to a given parametrization  $\pi_\theta$ , and optimize over the parameter  $\theta$ . Unfortunately, this might break the convexity of the  $(\varphi$ -DRO-CRM) objective, as  $\theta \rightarrow c(x, a)\omega^{\pi_\theta}(x, a)$  is rarely a convex function of  $\theta$ . We argue that the resulting optimization objective still improves over POEM, which was already highly non-convex w.r.t  $\pi$  because of the square root empirical variance term. The empirical results to come confirms this intuition. We will also discuss in Section 7.3.3 how to achieve fully convex off-line policy optimization objectives for policies which are log-concave w.r.t their parametrization  $\theta$ .*

**Hyper-parameters.** Through its desirable optimization properties, the  $(\varphi$ -DRO-CRM) objective solves two major limitations of the original (CRM) objective. Remains the issue of hyper-parameter tuning; the ambiguity size  $\varepsilon$  remains to be determined, either by cross-validation or by resorting to an heuristic rule. In an attempt to circumvent all of POEM's limitation altogether, one can use the value recommended by the asymptotical analysis:  $\varepsilon = \check{\varphi}(1)\chi_{1,1-\delta}^2/(2n)$  for a given value of the failure level  $\delta$ . We will follow this idea in the experimental section to come.

**Empirical evaluation.** We repeat the same methodology as in Section 7.3.1, and report experiments illustrating the appealing empirical performances of  $\varphi$ -DRO-CRM algorithms<sup>3</sup>. We report in Fig. 7.4 the expected risk of the policies  $\hat{\pi}_n$  returned by  $\varphi$ -DRO-CRM, as well as their greedy versions  $\hat{\pi}_n^\infty$  for the four  $f$ -divergences listed in Table 7.1. As anticipated in the previous paragraph, one major difference here is that we no longer use cross-validation to set  $\varepsilon$ , but rather use the value recommended by the asymptotic analysis. Therefore,  $\mathcal{D}_{\text{full}}^{\text{valid}}$  is not used by  $\varphi$ -DRO-CRM, but only by POEM to select its parameter  $\lambda$  by cross-validation. While this gives a *de-facto* advantage to POEM, we found this strategy to work quite well and still allows for  $\varphi$ -DRO-CRM to be competitive with POEM. We report result for batch (suffix -b) and stochastic (suffix -s) implementation of the algorithms. For batch implementations, we defer the optimization routine to L-BFGS. For stochastic implementations, we use the Adam optimizer (Kingma and Ba, 2014) with default configuration and a batch size of 32 samples. In the batch case, one can notice that DRO-based methods provide either similar or better empirical results than POEM on all considered datasets, while being hyper-parameter free (which, again, is not the case of POEM). On the Yeast dataset, the improvement is quite significative for two of the four  $f$ -divergence (Burg and Hellinger). On the negative side, it seems there is no consistency in the relative performance of the different divergences. This is quite troublesome in practice, as to the best of our knowledge there is no obvious nor preferable choice of divergences given a dataset. A solution to this problem is probably to cross-validate this choice, potentially over a continuous parametrization of the divergence considered here (such as the parameter of a Cressie-Read divergence). Finally, we note that POEM-s dominates among the all stochastic

<sup>3</sup>The results for POEM are different than those listed in Tables 7.2 and 7.3 because both experiments were run with different logging policies.

Algorithm	Risk( $\hat{\pi}_n$ )	Risk( $\hat{\pi}_n^\infty$ )	Algorithm	Risk( $\hat{\pi}_n$ )	Risk( $\hat{\pi}_n^\infty$ )
POEM-b	0.93 (0.06)	0.91 (0.06)	POEM	5.15 (0.07)	4.34 (0.13)
$\varphi$ -DRO-CRM-b			$\varphi$ -DRO-CRM-b		
$\varphi_1$	0.90 (0.06)	0.88 (0.06)	$\varphi_1$	5.32 (0.04)	5.29 (0.11)
$\varphi_2$	<b>0.89</b> (0.06)	0.87 (0.06)	$\varphi_2$	5.17 (0.06)	4.71 (0.11)
$\varphi_3$	1.06 (0.06)	<b>0.85</b> (0.05)	$\varphi_3$	<b>5.07</b> (0.07)	<b>4.24</b> (0.13)
$\varphi_4$	1.06 (0.06)	0.85 (0.05)	$\varphi_4$	5.09 (0.07)	4.27 (0.13)
POEM-s	<b>1.0</b> (0.05)	<b>0.97</b> (0.05)	POEM-s	<b>5.16</b> (0.05)	<b>4.62</b> (0.1)
$\varphi$ -DRO-CRM-s			$\varphi$ -DRO-CRM-s		
$\varphi_1$	1.06 (0.07)	1.04 (0.07)	$\varphi_1$	5.17 (0.06)	4.72 (0.12)
$\varphi_2$	1.05 (0.06)	1.02 (0.06)	$\varphi_2$	5.17 (0.06)	4.71 (0.11)
$\varphi_3$	1.12 (0.05)	1.08 (0.05)	$\varphi_3$	5.17 (0.06)	4.72 (0.1)
$\varphi_4$	1.3 (0.08)	1.18 (0.07)	$\varphi_4$	5.27 (0.06)	4.71 (0.1)

(a) Scene dataset.

(b) Yeast dataset.

Figure 7.4: Empirical comparison between POEM and  $\varphi$ -DRO-CRM on the Yeast and Scene datasets. The suffix **-b** (resp. **-s**) refer to batch implementation (resp. stochastic). We evaluate  $\varphi$ -DRO-CRM for the four  $f$ -divergence listed in Table 7.1. On both datasets,  $\varphi$ -DRO-CRM-b matches or improves POEM-b, while being hyper-parameter free. The stochastic version manages to be competitive with POEM-s, despite the latter being not a fully stochastic algorithm - it needs to periodically go through the entire bandit dataset.

algorithms considered. This is however to be nuanced, as this algorithm still needs to load in memory the entire dataset at every epoch (*e.g.* every time an upper-bound on the true objective is constructed). This is not the case for DRO-based algorithms. We also postulate that the nonetheless good performances reported here for stochastic DRO algorithms can be decisive when considering more complex policies (*e.g.* parametrized by a neural network, where POEM-S have been reported to fail).

### 7.3.3 Extensions

In the following, we discuss how different estimators can be used and robustified in the same way as the IPS, for improved performances or additional guarantees - however without sacrificing convexity.

**Variance reduction** The methods presented so far rely on the vanilla IPS estimator. As discussed in Section 6.2.1, it can suffer from large variance leading to a degradation of its performances.. It is therefore natural to investigate whether the DRO approach could be applied to estimators that actively reduce variance. A logical candidate for this task is the self-normalized importance sampling estimator described in Eq. (6.5). This estimator is unfortunately not convex in  $\pi$ , which goes against the efforts undertaken in so far to maintain well-behaved optimization tasks. A simple alternative consist in using a *additive* control variate (instead of a multiplicative one). The resulting estimator writes:

$$\hat{R}_{n,\rho}^{\text{CV}}(\pi) = \frac{1}{n} \sum_{i=1}^n (c(x_i, a_i) - \rho) \omega^\pi(x_i, a_i) + \rho.$$

A robust version of this estimator easily follows, and enjoys the same convex properties of the IPS robust risk. The variance-reduction property of the additive control variate are well-known

and extensively described in the literature - *c.f.* (Owen, 2013) for a review. We recall them in the following Lemma.

**Lemma 7.3.2.** *[Propensity weights as an additive control variate] For all  $\rho$ ,  $\hat{R}_{n,\rho}^{CV}(\pi)$  is an unbiased estimator of  $\text{Risk}(\pi)$ , achieving a better variance than naive IPS whenever:*

$$0 \leq \rho \leq 2 \frac{\text{Cov}(\ell_\pi, \omega_\pi)}{\text{Var}(\omega_\pi)}.$$

*In addition, if the cost is independent of the propensity weights, we obtain  $\rho^* = \mathbb{E}[c]$ .*

In practice, we do not know how to derive the optimal (in terms of variance reduction) coefficient  $\rho^*$  analytically; however as hinted by the last statement in Lemma 7.3.2, one can directly use the cost's empirical mean under  $\pi_0$  as a first approximation. Finally, robust estimators can be directly coupled with the doubly robust approach laid out in Eq. (6.4); this allows to leverage context/actions dependent control variates in a straight-forward fashion, for improved variance reduction.

**Parametric policies.** This paragraph echoes Remark 7.3.1, where we discuss the break in convexity of the ( $\varphi$ -DRO-CRM) objective that arise when parametrizing the policy space. We discuss here how to alleviate this issue when policies are log-concave - *e.g.* exponential policies laid out in Eq. (6.9). In this case, the objective becomes a negative sum of log-concave functions (negative because of Assumption 6.1.2 which yields  $c(x, a) \leq 0$ ) resulting in a non-convex optimization objective. Following (Roux, 2017), one can bypass this non-convexity by constructing a tight convex upper bound of the original objective.

**Lemma 7.3.3.** *[Convex upper-bound for log-concave policies] Let  $\pi_\theta$  be a log-concave (w.r.t  $\theta$ ) policy. For a given  $\theta_0$ , let:*

$$\hat{R}_n^{\text{UP}}(\pi_\theta) = \frac{1}{n} \sum_{i=1}^n \frac{\pi_{\theta_0}(a_i, x_i)}{\pi_0(a_i, x_i)} (1 + \log[\frac{\pi_\theta(a_i, x_i)}{\pi_{\theta_0}(a_i, x_i)}]) c(x_i, a_i).$$

*$\hat{R}_n^{\text{UP}}(\pi_\theta)$  is a convex upper bound of the IPS risk. The closer  $\theta_0$  to  $\theta$ , the tighter the upper bound, with equality at  $\theta_0 = \theta$ .*

The main take-away from Lemma 7.3.3 is that one can build a convex proxy of the original objective, through an iterative procedure that only uses convex losses throughout the whole optimization process. Once again, we can build a robust version of this estimator which can be efficiently optimized. Note that here, the robust estimator will be convex w.r.t the parametrization  $\theta$  as soon as the policy is log-concave.

# Conclusion

---

The goal of this chapter is to provide a brief summary of our contributions to offline policy evaluation and learning. We then identify potential directions for future work, and finally discuss some concurrent and similar work and their ties to our approach.

**Brief summary.** The learning from logged bandit feedback problem presents a salient difficulty compared to the classical supervised setting, for the same estimator applied to different policies can have vastly disparate variance. Motivated by recent progress in the generalized empirical likelihood and distributionally robust optimization literature, we first propose to use new asymptotic confidence intervals for offline policy evaluation. This is mainly motivated by their variance sensitivity properties, which allows for reasonably tight policy-dependent bounds. We demonstrate empirically that despite being asymptotic, these confidence intervals are sound to use even for finite sample sizes. We also investigate their relevance for offline policy optimization. In this context, they lead to a generalization of the CRM principle, a state-of-the-art procedure for this task. Our formulation leads to a variety of different objectives for offline policy optimization. We describe in [Section 7.3.1](#) an exponential-weight version of the original CRM objective that arises from a particular configuration of our approach, and shows that it over-performs the original formulation. In [Section 7.3.2](#) we also show that our approach leads to sounder policy optimization schemes - from an optimization perspective. Indeed, they yield convex objectives which easily undergo stochastic optimization. The former allows for principle minimization, while the latter is much needed in virtually all practical instances, where the size of logged interactions typically prohibit batch strategies.

**Future work.** The experimental results reported in [Section 7.3.2](#) are mostly illustrative; a natural direction for future work involves an exhaustive evaluation of DRO-based method for policy optimization. This could also be the occasion to investigate the practical impact of the extensions we discussed in [Section 7.3.3](#). A major difficulty for achieving such task comes from the well-known instability of comparing offline policy optimization methods; results are highly implementation-dependent, as little changes to the logging policy often leads to sensibly different conclusions. We are convinced that further research on the topic could highly benefit from controlled and standardized learning environments - such as the ones that exist for online reinforcement learning. There have been recent efforts in that direction ([Rohde et al., 2018](#)), yielding some good candidates for such an environment.

**Recent related work.** Concurrently with our contributions to this topic appeared several closely related works, leveraging similar ideas either in the contextual bandit or reinforcement

---

learning setting. For contextual bandits, [Karampatziakis et al. \(2020\)](#) design a new counterfactual estimator, based on the empirical likelihood principle with reverse Kullback-Leibler divergence and with an additional weight constraint. They derive a novel confidence interval for this estimator, analogous to ours (which are tailored for the IPS) and with identical coverage guarantees. Similarly, they report satisfactory empirical coverage for their confidence interval despite being only asymptotical. Alike the rationale presented in this manuscript, they derive a policy optimization objective based on this confidence interval. [Dai et al. \(2020\)](#) apply similar tools as the ones presented in this dissertation, however in a reinforcement learning setting. In many ways, their approach is a generalization of ours to this more challenging problem - up to one major difference; in their formulation, perfect knowledge about the logger  $\pi_0$  is not required. Furthermore, they also provide finite-sample guarantees for generalized empirical likelihood-based confidence intervals, henceforth bringing theoretical confirmation of their relevance for real-world problems.

# Bibliography

---

- Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL <https://proceedings.neurips.cc/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf>.
- Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1–9, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL <http://proceedings.mlr.press/v22/abbasi-yadkori12.html>.
- Marc Abeille and Alessandro Lazaric. Linear Thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165 – 5197, 2017. doi: 10.1214/17-EJS1341SI. URL <https://doi.org/10.1214/17-EJS1341SI>.
- Marc Abeille, Louis Fauray, and Clement Calauzenes. Instance-wise minimax-optimal algorithms for logistic bandits. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3691–3699. PMLR, 13–15 Apr 2021. URL <http://proceedings.mlr.press/v130/abeille21a.html>.
- Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1638–1646, Beijing, China, 22–24 Jun 2014. PMLR. URL <http://proceedings.mlr.press/v32/agarwalb14.html>.
- Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge university press, 2009.
- Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19): 1876–1902, 2009.
- Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time Analysis of the Multi-armed Bandit Problem. *Machine Learning*, 47(2):235–256, 2002.

- Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4(none):384 – 414, 2010. doi: 10.1214/09-EJS521. URL <https://doi.org/10.1214/09-EJS521>.
- Aharon Ben-Tal, Dick den Hertog, Anja De Waegenare, Bertrand Melenberg, and Gijs Rennen. Robust Solutions of Optimization Problems Affected by Uncertain Probabilities. *Management Science*, 59(2):341–357, 2013. ISSN 00251909, 15265501. URL <http://www.jstor.org/stable/23359484>.
- Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic Multi-Armed-Bandit Problem with Non-stationary Rewards. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL <https://proceedings.neurips.cc/paper/2014/file/903ce9225fca3e988c2af215d4e544d3-Paper.pdf>.
- Léon Bottou, Jonas Peters, Joaquin Quiñero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research*, 14(65):3207–3260, 2013. URL <http://jmlr.org/papers/v14/bottou13a.html>.
- Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516 – 1541, 2013. doi: 10.1214/13-AOS1119. URL <https://doi.org/10.1214/13-AOS1119>.
- Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 696–726, Phoenix, USA, 25–28 Jun 2019. PMLR. URL <http://proceedings.mlr.press/v99/chen19b.html>.
- Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Hedging the Drift: Learning to Optimize under Non-Stationarity. *arXiv preprint arXiv:1903.01461*, 2019a.
- Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to Optimize under Non-Stationarity. In *Proceedings of the 22rd International Conference on Artificial Intelligence and Statistics*, pages 1079–1087, 2019b.
- Patrick L Combettes. Perspective Functions: Properties, Constructions, and Examples. *Set-Valued and Variational Analysis*, 26(2):247–264, 2018.
- Bo Dai, Ofir Nachum, Yinlam Chow, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. CoinDICE: Off-Policy Confidence Interval Estimation. In *Advances in Neural Information Processing Systems*, volume 34, 2020. URL <https://papers.nips.cc/paper/2020/file/6aaba9a124857622930ca4e50f5afed2-Paper.pdf>.
- Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic Linear Optimization under Bandit Feedback. In *COLT*, 2008.
- Victor H de la Pena, Michael J Klass, and Tze Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Annals of probability*, pages 1902–1933, 2004.

- Shi Dong, Tengyu Ma, and Benjamin Van Roy. On the Performance of Thompson Sampling on Logistic Bandits. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1158–1160, Phoenix, USA, 25–28 Jun 2019. PMLR. URL <http://proceedings.mlr.press/v99/dong19a.html>.
- John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. *arXiv preprint arXiv:1610.03425*, 2016.
- Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly Robust Policy Evaluation and Optimization. *Statist. Sci.*, 29(4):485–511, 11 2014. doi: 10.1214/14-STS500. URL <https://doi.org/10.1214/14-STS500>.
- Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More Robust Doubly Robust Off-policy Evaluation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1447–1456, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL <http://proceedings.mlr.press/v80/farajtabar18a.html>.
- Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved Optimistic Algorithms for Logistic Bandits. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3052–3060, Virtual, 13–18 Jul 2020a. PMLR. URL <http://proceedings.mlr.press/v119/faury20a.html>.
- Louis Faury, Ugo Tanielian, Elvis Dohmatob, Elena Smirnova, and Flavian Vasile. Distributionally Robust Counterfactual Risk Minimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3850–3857, Apr. 2020b. doi: 10.1609/aaai.v34i04.5797. URL <https://ojs.aaai.org/index.php/AAAI/article/view/5797>.
- Louis Faury, Yoan Russac, Marc Abeille, and Clément Calauzènes. A Technical Note on Non-Stationary Parametric Bandits: Existing Mistakes and Preliminary Solutions. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 619–626. PMLR, 16–19 Mar 2021. URL <http://proceedings.mlr.press/v132/faury21a.html>.
- Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential Experimental Design for Transductive Linear Bandits. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/8ba6c657b03fc7c8dd4dff8e45defcd2-Paper.pdf>.
- Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL <https://proceedings.neurips.cc/paper/2010/file/c2626d850c80ea07e7511bbae4c76f4b-Paper.pdf>.
- David A. Freedman. On Tail Probabilities for Martingales. *The Annals of Probability*, 3(1):100–118, 1975. doi: 10.1214/aop/1176996452. URL <https://doi.org/10.1214/aop/1176996452>.
- Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188. Springer, 2011.

- Elad Hazan, Tomer Koren, and Kfir Y. Levy. Logistic regression: Tight bounds for stochastic and online optimization. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 197–209, Barcelona, Spain, 13–15 Jun 2014. PMLR. URL <http://proceedings.mlr.press/v35/hazan14a.html>.
- Tim Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995. ISSN 00401706. URL <http://www.jstor.org/stable/1269620>.
- D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446. URL <https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1952.10483446>.
- Edward L. Ionides. Truncated Importance Sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008. ISSN 10618600. URL <http://www.jstor.org/stable/27594308>.
- Rémi Jézéquel, Pierre Gaillard, and Alessandro Rudi. Efficient improper learning for online logistic regression. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2085–2108. PMLR, 09–12 Jul 2020. URL <http://proceedings.mlr.press/v125/jezequel20a.html>.
- Nan Jiang and Lihong Li. Doubly Robust Off-policy Value Evaluation for Reinforcement Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 652–661, New York, New York, USA, 20–22 Jun 2016. PMLR. URL <http://proceedings.mlr.press/v48/jiang16.html>.
- Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable Generalized Linear Bandits: Online Computation and Hashing. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/file/28dd2c7955ce926456240b2ff0100bde-Paper.pdf>.
- Kwang-Sung Jun, Lalit Jain, Blake Mason, and Houssam Nassif. Improved Confidence Bounds for the Linear Logistic Model and Applications to Linear Bandits. *arXiv preprint arXiv:2011.11222*, 2021.
- Nikos Karampatziakis, John Langford, and Paul Mineiro. Empirical Likelihood for Contextual Bandits. In *Advances in Neural Information Processing Systems*, volume 34, 2020. URL <https://papers.nips.cc/paper/2020/file/6d34d468ac8876333c4d7173b85efed9-Paper.pdf>.
- Abbas Kazerouni and Lawrence M Wein. Best Arm Identification in Generalized Linear Bandits. *Operations Research Letters*, 2021.
- Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Levente Kocsis and Csaba Szepesvári. Discounted UCB. volume 2, 2006.

- Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. *Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning*, chapter 6, pages 130–166. INFORMS, 2019. doi: 10.1287/educ.2019.0198. URL <https://pubsonline.informs.org/doi/abs/10.1287/educ.2019.0198>.
- Ilja Kuzborskij, Claire Vernade, András György, and Csaba Szepesvári. Confident Off-Policy Evaluation and Selection through Self-Normalized Importance Weighting. *arXiv preprint arXiv:2006.10460*, 2020.
- Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized Exploration in Generalized Linear Bandits. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2066–2076. PMLR, 26–28 Aug 2020. URL <http://proceedings.mlr.press/v108/kveton20a.html>.
- John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, Williams College, Williamstown, MA, USA, pages 282–289, 2001.
- T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. ISSN 0196-8858. doi: [https://doi.org/10.1016/0196-8858\(85\)90002-8](https://doi.org/10.1016/0196-8858(85)90002-8). URL <https://www.sciencedirect.com/science/article/pii/0196885885900028>.
- Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.
- B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by modelselection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000. doi: 10.1214/aos/1015957395. URL <https://doi.org/10.1214/aos/1015957395>.
- Lihong Li, Yu Lu, and Dengyong Zhou. Provably Optimal Algorithms for Generalized Linear Contextual Bandits. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2071–2080. PMLR, 06–11 Aug 2017. URL <http://proceedings.mlr.press/v70/li17c.html>.
- Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline Policy Evaluation across Representations with Applications to Educational Games. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS '14*, page 1077–1084, Richland, SC, 2014. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450327381. URL <http://grail.cs.washington.edu/projects/ordering/orderingpaperExtended.pdf>.
- Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. In *Conference on Learning Theory (COLT)*, 2009. URL <http://dblp.uni-trier.de/db/conf/colt/colt2009.html#MaurerP09>.
- Peter McCullagh and J. A. Nelder. *Generalized linear models*. Chapman and Hall, 1989.

- Hongseok Namkoong and John C Duchi. Stochastic Gradient Methods for Distributionally Robust Optimization with f-divergences. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL <https://proceedings.neurips.cc/paper/2016/file/4588e674d3f0faf985047d4c3f13ed0d-Paper.pdf>.
- Art B Owen. *Empirical likelihood*. CRC press, 2001.
- Art B. Owen. *Monte Carlo theory, methods and examples*. 2013.
- Leandro Pardo. *Statistical inference based on divergence measures*. CRC press, 2018.
- Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.
- Clémence Réda, Emilie Kaufmann, and Andrée Delahaye-Duriez. Top-m identification for linear bandits. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1108–1116. PMLR, 13–15 Apr 2021. URL <http://proceedings.mlr.press/v130/reda21a.html>.
- R Tyrrell Rockafellar. Risk and utility in the duality framework of convex analysis. In *Jonathan M. Borwein Commemorative Conference*, pages 21–42. Springer, 2017.
- David Rohde, Stephen Bonner, Travis Dunlop, Flavian Vasile, and Alexandros Karatzoglou. Recogym: A reinforcement learning environment for the problem of product recommendation in online advertising. *arXiv preprint arXiv:1808.00720*, 2018.
- Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. ISSN 00063444. URL <http://www.jstor.org/stable/2335942>.
- Nicolas Le Roux. Tighter Bounds Lead to Improved Classifiers. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL <https://openreview.net/forum?id=HyAbMKwxe>.
- Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.
- Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted Linear Bandits for Non-Stationary Environments. In *Advances in Neural Information Processing Systems*, pages 12017–12026, 2019.
- Yoan Russac, Olivier Cappé, and Aurélien Garivier. Algorithms for Non-Stationary Generalized Linear Bandits. *arXiv preprint arXiv:2003.10113*, 2020.
- Yoan Russac, Louis Faury, Olivier Cappé, and Aurélien Garivier. Self-Concordant Analysis of Generalized Linear Bandits with Forgetting. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Virtual, 2021.
- Otmane Sakhi, Louis Faury, and Flavian Vasile. Improving Offline Contextual Bandits with Distributional Robustness. *Proceedings of the ACM RecSys Workshop on Reinforcement Learning and Robust Estimators for Recommendation Systems (REVEAL '20)*, 2020.

- Max Simchowitz and Dylan J Foster. Naive Exploration is Optimal for Online LQR. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, 2020.
- Marta Soare, Alessandro Lazaric, and Remi Munos. Best-Arm Identification in Linear Bandits. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL <https://proceedings.neurips.cc/paper/2014/file/f387624df552cea2f369918c5e1e12bc-Paper.pdf>.
- Adith Swaminathan and Thorsten Joachims. Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization. *Journal of Machine Learning Research*, 16(52): 1731–1755, 2015a. URL <http://jmlr.org/papers/v16/swaminathan15a.html>.
- Adith Swaminathan and Thorsten Joachims. The Self-Normalized Estimator for Counterfactual Learning. In *Advances in Neural Information Processing Systems*, volume 28, pages 3231–3239. Curran Associates, Inc., 2015b. URL <https://proceedings.neurips.cc/paper/2015/file/39027dfad5138c9ca0c474d71db915c3-Paper.pdf>.
- Chao Tao, Saúl Blanco, and Yuan Zhou. Best Arm Identification in Linear Bandits with Linear Dimension Dependency. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4877–4886. PMLR, 10–15 Jul 2018. URL <http://proceedings.mlr.press/v80/tao18a.html>.
- Philip Thomas and Emma Brunskill. Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2139–2148, New York, New York, USA, 20–22 Jun 2016. PMLR. URL <http://proceedings.mlr.press/v48/thomasa16.html>.
- Philip S. Thomas, Georgios Theodorou, and Mohammad Ghavamzadeh. High Confidence Off-Policy Evaluation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI’15, page 3000–3006. AAAI Press, 2015. ISBN 0262511290.
- Ahmed Touati and Pascal Vincent. Efficient Learning in Non-Stationary Linear Markov Decision Processes. *arXiv preprint arXiv:2010.12870*, 2021.
- Alexandre B Tsybakov. *Introduction to non-parametric estimation*. Springer Science & Business Media, 2008.
- Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.
- Nikos Vlassis, Aurelien Bibaut, Maria Dimakopoulou, and Tony Jebara. On the Design of Estimators for Bandit Off-Policy Evaluation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6468–6476. PMLR, 09–15 Jun 2019. URL <http://proceedings.mlr.press/v97/vlassis19a.html>.
- Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and Adaptive Off-policy Evaluation in Contextual Bandits. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3589–3597, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL <http://proceedings.mlr.press/v70/wang17a.html>.

- Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards Optimal Off-Policy Evaluation for Reinforcement Learning with Marginalized Importance Sampling. In *Advances in Neural Information Processing Systems*, volume 32, pages 9668–9678. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/4ffb0d2ba92f664c2281970110a2e071-Paper.pdf>.
- Liyuan Xu, Junya Honda, and Masashi Sugiyama. A fully adaptive algorithm for pure exploration in linear bandits. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 843–851. PMLR, 09–11 Apr 2018. URL <http://proceedings.mlr.press/v84/xu18d.html>.
- Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi-hua Zhou. Online Stochastic Linear Optimization under One-bit Feedback. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 392–401, New York, New York, USA, 20–22 Jun 2016. PMLR. URL <http://proceedings.mlr.press/v48/zhangb16.html>.
- Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 2020, 2020.

# Appendix

---

## A Concentration inequalities

In this section, we remind some basic definitions and concentrations results. The proofs are standard and therefore omitted here for the sake of conciseness.

**Definition A.1** (Convergence in distribution). *A sequence of real-valued random variable  $(X_1, \dots, X_n)$  is said to converge in distribution to a random variable  $X$  if  $\lim_{n \rightarrow \infty} F_n(x) = F(x)$  at every  $x$  where  $F$  is continuous, where  $F_n$  and  $F$  are the cumulative distribution functions of  $X_n$  and  $X$ , respectively.*

In what follows convergence in distribution of  $X_n$  to  $X$  will be denoted  $X_n \xrightarrow{d} X$ . We also use the notation  $\bar{X}_n := n^{-1} \sum_{i=1}^n X_i$ .

**Theorem A.1** (Central Limit Theorem). *Let  $X_1, \dots, X_n$  be i.i.d random variables such that  $\mathbb{E}[X_i^2] < +\infty$ , with mean  $\mu$  and variance  $\sigma^2$ . Then:*

$$\frac{\sqrt{n}}{\sigma} (\bar{X}_n - \mu) \xrightarrow{d} Z$$

where  $Z$  is a standard Gaussian random variable.

By denoting  $\Phi$  the cumulative distribution function of a standard Gaussian random variable one easily obtains the following result, providing an asymptotic confidence interval for the mean  $\mu$ :

**Lemma A.1** (Asymptotic confidence interval). *Let  $X_1, \dots, X_n$  be i.i.d random variables such that  $\mathbb{E}[X_i^2] < +\infty$ , with mean  $\mu$  and variance  $\sigma^2$ . For  $\delta \in (0, 1]$ :*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left( \left| \bar{X}_n - \mu \right| \leq \frac{\sigma}{\sqrt{n}} \Phi^{-1}(\delta/2) \right) \geq 1 - \delta$$

A similar result exists when the variance  $\sigma^2$  is unknown and involves the *empirical* variance  $s_n^2 := \frac{1}{n-1} \sum_{i=1}^n X_i^2 - \bar{X}_n^2$ . The central limit theorem is classically extended by replacing  $\sigma$  by  $s_n$  thanks to Slutsky's lemma; cf. (Van der Vaart, 2000, Section 2.9). This leads to a similar asymptotic confidence interval.

**Lemma A.2** (Asymptotic confidence interval with empirical variance). *Let  $X_1, \dots, X_n$  be i.i.d random variables such that  $\mathbb{E}[X_i^2] < +\infty$ , with mean  $\mu$ . For  $\delta \in (0, 1]$ :*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left( \left| \bar{X}_n - \mu \right| \leq \frac{s_n}{\sqrt{n}} \Phi^{-1}(\delta/2) \right) \geq 1 - \delta$$

We now make a few remainders on finite-time concentration inequalities. Below we give one definition (there exists many equivalent ones) of a sub-gaussian random variable.

**Definition A.2** (Sub-gaussian random variable). *A real-valued random variable  $X$  is said to be sub-gaussian if there exists  $\sigma > 0$  such that for all  $\lambda \in \mathbb{R}$ ,*

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2 \sigma^2 / 2)$$

In particular, it can be easily shown that bounded random variables are sub-gaussian.

**Lemma A.3** (Hoeffding's lemma). *Let  $X$  be a real-valued random variable and  $a, b \in \mathbb{R}$  such that  $X \in [a, b]$  almost surely. Then  $X$  is sub-gaussian with proxy-variance  $\sigma = (b - a)^2 / 4$ :*

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2 (b - a)^2 / 8) \quad \forall \lambda \in \mathbb{R}$$

This bound on the moment-generating function can be leveraged along with Markov's inequality to derive finite-time concentration inequalities over sums (or means) of bounded random variables.

**Lemma A.4** (Chernoff-Hoeffding's concentration inequality for bounded r.v.). *Let  $a, b \in \mathbb{R}$  and  $\mu \in \mathbb{R}$ . Let  $(X_1, \dots, X_n)$  be independent bounded random variables such that  $X_i \in [a, b]$  and  $\mathbb{E}[X_i] = \mu$  for all  $i \in [n]$ . Then for all  $\epsilon > 0$ :*

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i \geq \mu + \epsilon\right) \leq \exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$$

Equivalently for any  $\delta \in (0, 1]$ :

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i \leq \mu + |b - a| \sqrt{\frac{\log(1/\delta)}{2n}}\right) \geq 1 - \delta$$

The same results hold for the left tail of  $\frac{1}{n} \sum_{i=1}^n X_i$ . This concentration inequality can be refined if one has knowledge of the variance of the random variables.

**Lemma A.5** (Bernstein's inequality for bounded random variables). *Let  $b \in \mathbb{R}^+$  and  $\mu \in \mathbb{R}$ . Let  $(X_1, \dots, X_n)$  be independent bounded random variables such that  $\mathbb{E}[X_i] = \mu$  and  $|X_i - \mu| \leq b$  almost surely for all  $i \in [n]$ . Then for all  $\epsilon > 0$ :*

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i \geq \mu + \epsilon\right) \leq \exp\left(-\frac{n^2 \epsilon^2 / 2}{\sum_{i=1}^n \sigma_i^2 + bn\epsilon / 3}\right)$$

where  $\sigma_i^2 := \text{Var}(X_i)$  for all  $i \in [n]$ . As a consequence, for  $\delta \in (0, 1]$ :

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i \leq \mu + \frac{2b}{3n} \log(1/\delta) + \sqrt{\frac{2 \log(1/\delta)}{n^2}} \sqrt{\sum_{i=1}^n \sigma_i^2}\right) \geq 1 - \delta$$

Again, similar results hold for the left tail. When the actual variances  $\{\sigma_i\}$  are unknown, they can be replaced by their empirical counterpart.

**Lemma A.6** (Empirical Bernstein inequality, (Maurer and Pontil, 2009, Theorem 3)). *Let  $b \in \mathbb{R}^+$  and  $\mu \in \mathbb{R}$ . Let  $(X_1, \dots, X_n)$  be i.i.d random variables such that  $\mathbb{E}[X_i] = \mu$  and  $|X_i - \mu| \leq b$  almost surely for all  $i \in [n]$ . Then for all  $\delta \in (0, 1]$ :*

$$\mathbb{P} \left( \frac{1}{n} \sum_{i=1}^n X_i \leq \mu + \frac{7b}{3(n-1)} \log(2/\delta) + \sqrt{\frac{2 \log(2/\delta)}{n} s_n^2} \right) \geq 1 - \delta$$

## B Technical lemmas

**Lemma B.1.** *For all  $x \geq 0$ , the following inequality holds:*

$$\frac{1}{x} \left( 1 + \frac{\exp(-x) - 1}{x} \right) \geq \frac{1}{2+x}.$$

*Proof.* It is easy to show that the claimed inequality holds if and only if  $\exp(-x) \geq (2-x)(2+x)^{-1}$ . Let  $h(x) = (2+x)\exp(-x) - (2-x)$ . Easy computations yield that for all  $x$  we have  $h'(x) = -\exp(-x)(1+x) + 1$ . Using the fact that  $\exp(-x) \leq (1+x)^{-1}$  for all  $x \geq 0$  (derived from  $e^x \geq 1+x$ ) we get that:

$$h'(x) \geq -\frac{1+x}{1+x} + 1 = 0.$$

The increasing nature of  $h$  on  $\mathbb{R}^+$ , along with the fact that  $h(0) = 0$  is enough to show that  $\exp(-x) \geq (2-x)(2+x)^{-1}$  for all  $x \geq 0$ . As laid out in the first lines of the proof, this suffices to prove our claim. ■

**Proposition B.1** (Polynomial Inequality). *Let  $b, c \in \mathbb{R}^+$ , and  $x \in \mathbb{R}$ . The following implication holds:*

$$x^2 \leq bx + c \implies x \leq b + \sqrt{c}$$

*Proof.* Let  $f : x \rightarrow x^2 - bx - c$ . Then  $f$  is a strongly-convex function which roots are:

$$\lambda_{1,2} = \frac{1}{2}(b \pm \sqrt{b^2 + 4c})$$

If  $x^2 \leq -b - c$  then by convexity of  $f$  we obtain:

$$\begin{aligned} x &\leq \max(\lambda_1, \lambda_2) \\ &\leq \frac{1}{2}(b + \sqrt{b^2 + 4c}) \\ &\leq b + \sqrt{c} \end{aligned} \quad (\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}, \forall x, y \geq 0)$$

■

The following theorem is extracted from (Abbasi-Yadkori et al., 2011, Lemma 10).

**Lemma B.2** (Determinant-Trace inequality). *Let  $\{x_s\}_{s=1}^\infty$  a sequence in  $\mathbb{R}^d$  such that  $\|x_s\| \leq X$  for all  $s \in \mathbb{N}$ , and let  $\lambda$  be a non-negative scalar. For  $t \geq 1$  define  $\mathbf{V}_t := \sum_{s=1}^{t-1} x_s x_s^\top + \lambda \mathbf{I}_d$ . The following inequality holds:*

$$\det(\mathbf{V}_{t+1}) \leq \left( \lambda + (t-1)X^2/d \right)^d$$

We need a variation of the Elliptical Potential Lemma (Abbasi-Yadkori et al., 2011, Lemma 11) adjusted to handle (increasing) time-varying regulations.

**Lemma B.3** (Elliptical potential with time varying regularization.). *Let  $\{x_s\}_{s=1}^\infty$  a sequence in  $\mathbb{R}^d$  such that  $\|x_s\| \leq X$  for all  $s \in \mathbb{N}$ . Further let  $\{\lambda_s\}_{s=0}^\infty$  be an increasing sequence in  $\mathbb{R}^+$  s.t  $\lambda_1 \geq 1$ . For  $t \geq 1$  define  $\mathbf{V}_t := \sum_{s=1}^{t-1} x_s x_s^\top + \lambda_t \mathbf{I}_d$ . Then:*

$$\sum_{t=1}^T \|x_t\|_{\mathbf{V}_t^{-1}}^2 \leq 2d \max(1, X^2/\lambda_1) \log \left( \lambda_T/\lambda_1 + \frac{TX^2}{\lambda_1 d} \right)$$

*Proof.* By definition of  $\mathbf{V}_t$ :

$$\begin{aligned} |\mathbf{V}_{t+1}| &= \left| \sum_{s=1}^{t-1} x_s x_s^\top + x_t x_t^\top + \lambda_t \mathbf{I}_d \right| \\ &\geq \left| \sum_{s=1}^{t-1} x_s x_s^\top + x_t x_t^\top + \lambda_{t-1} \mathbf{I}_d \right| && (\lambda_t \geq \lambda_{t-1} > 0) \\ &= |\mathbf{V}_t + x_t x_t^\top| \\ &\geq |\mathbf{V}_t| \left| \mathbf{I}_d + \mathbf{V}_t^{-1/2} x_t x_t^\top \mathbf{V}_t^{-1/2} \right| \\ &= |\mathbf{V}_t| \left( 1 + \|x_t\|_{\mathbf{V}_t^{-1}}^2 \right) \end{aligned}$$

and therefore by taking the log on both side of the equation and summing from  $t = 1$  to  $T$ :

$$\begin{aligned} \sum_{t=1}^T \log \left( 1 + \|x_t\|_{\mathbf{V}_t^{-1}}^2 \right) &\leq \sum_{t=1}^T \log |\mathbf{V}_{t+1}| - \log |\mathbf{V}_t| \\ &= \log \left( \frac{\det(\mathbf{V}_{T+1})}{\det(\lambda_1 \mathbf{I}_d)} \right) && \text{(telescopic sum)} \\ &= \log (\det(\mathbf{V}_{T+1})) && (\lambda_1 = 1) \\ &\leq d \log \left( \lambda_T + \frac{TX^2}{d} \right) && \text{(Lemma B.2)} \end{aligned}$$

Remember that for all  $x \in [0, 1]$  we have the inequality  $\log(1 + x) \geq x/2$ . Also note that  $\|x_t\|_{\mathbf{V}_t^{-1}}^2 \leq X^2/\lambda$ . Therefore:

$$\begin{aligned} d \log \left( \lambda_T + \frac{TX^2}{d} \right) &\geq \sum_{t=1}^T \log \left( 1 + \|x_t\|_{\mathbf{V}_t^{-1}}^2 \right) \\ &\geq \sum_{t=1}^T \log \left( 1 + \frac{1}{\max(1, X^2/\lambda_t)} \|x_t\|_{\mathbf{V}_t^{-1}}^2 \right) \\ &\geq \frac{1}{2 \max(1, X^2/\lambda_1)} \sum_{t=1}^T \|x_t\|_{\mathbf{V}_t^{-1}}^2 \end{aligned}$$

which yields the announced result. ■

**Titre :** Intervalles de Confiance Sensibles à la Variance: Applications aux Bandits Paramétrique et Bandits Hors Ligne

**Mots clés :** régions de confiance, processus décisionnel, non-linéarité, contrefactuel.

**Résumé :** Cette thèse présente des contributions récentes au problème d'optimisation sous feedback bandit, au travers de la construction d'intervalles de confiance sensibles à la variance. Nous traitons deux aspects distincts du problème: **(1)** la minimisation du regret pour les bandits à modèle linéaire généralisé (GLMs), une large classe de bandits paramétriques non-linéaires et **(2)** le problème d'optimisation de politique hors ligne sous signal bandit. Concernant **(1)** nous étudions les effets de la non-linéarité dans les GLBs et remettons en question la compréhension actuelle selon laquelle des hauts niveaux de non-linéarité ne peuvent être que préjudiciables à l'équilibre exploration-exploitation. Des algorithmes améliorés suivis d'une nouvelle méthode d'analyse montre que si correctement manipulé, le problème de minimisation du regret dans les GLBs n'est pas nécessairement plus dur que pour leur contrepartie linéaire. Il peut même être si-

gnificativement facilité pour certains membres importants de la famille GLB comme le bandit logistique. Notre approche utilise de nouveaux ensembles de confiance sensibles à la non-linéarité au travers de la variance qu'elle impose à la fonction récompense, accompagnés d'un traitement local de la non-linéarité au travers d'une analyse dite auto-concordante. Concernant **(2)** nous utilisons des résultats de la littérature de l'optimisation robuste afin de construire des intervalles de confiance asymptotiques sensibles à la variance pour l'évaluation contrefactuel de politiques. Cela permet d'assurer du conservatisme (désirable pour des agents averse au risque) lors de la recherche hors-ligne de politiques prometteuses. Cet interval de confiance engendre de nouveaux objectifs contrefactuels qui sont plus adaptés à des applications pratiques, car convexes et de nature composés.

**Title :** Variance-Sensitive Confidence Intervals for Parametric and Offline Bandits

**Keywords :** confidence regions, decision-making, non-linearity, counterfactual.

**Abstract :** In this dissertation we present recent contributions to the problem of optimization under bandit feedback through the design of variance-sensitive confidence intervals. We tackle two distinct topics: **(1)** the regret minimization task in Generalized Linear Bandits (GLBs), a broad class of non-linear parametric bandits and **(2)** the problem of off-line policy optimization under bandit feedback. For **(1)** we study the effects of non-linearity in GLBs and challenge the current understanding that a high level of non-linearity is detrimental to the exploration-exploitation trade-off. We introduce improved algorithms as well as a novel analysis that prove that if correctly handled, the regret minimization task in GLBs is not necessarily harder than for their linear counterparts. It can even be ea-

sier for some important members of the GLB family such as the Logistic Bandit. Our approach leverages a new confidence set which captures the non-linearity of the reward signal through its variance, along with a local treatment of the non-linearity through a so-called self-concordance analysis. For **(2)** we leverage results from the distributionally robust optimization framework to construct asymptotic variance-sensitive confidence intervals for the counterfactual evaluation of policies. This allows to ensure conservatism (sought out by risk-averse agents) while searching off-line for promising policies. Our confidence intervals lead to new counterfactual objectives which, contrary to their predecessors, are more suited for practical deployment thanks to their convex and composite natures.