Instance-Wise Minimax-Optimal Algorithms for Logistic Bandits

Marc Abeille[⋆]
Criteo AI Lab

Louis Faury[⋆]
Criteo AI Lab
LTCI TélécomParis

Clément Calauzènes
Criteo AI Lab

Abstract

Logistic Bandits have recently attracted substantial attention, by providing an uncluttered yet challenging framework for understanding the impact of non-linearity in parametrized bandits. It was shown by Faury et al. (2020) that the learning-theoretic difficulties of Logistic Bandits can be embodied by a *large* (sometimes prohibitively) problem-dependent constant κ, characterizing the magnitude of the reward's non-linearity. In this paper we introduce a novel algorithm for which we provide a refined analysis. This allows for a better characterization of the effect of non-linearity and yields improved problem-dependent guarantees. In most favorable cases this leads to a regret upper-bound scaling as $\tilde{\mathcal{O}}(d\sqrt{T/\kappa})$, which dramatically improves over the $\tilde{\mathcal{O}}(d\sqrt{T}+\kappa)$ state-of-the-art guarantees. We prove that this rate is *minimax-optimal* by deriving a $\Omega(d\sqrt{T/\kappa})$ problem-dependent lower-bound. Our analysis identifies two regimes (permanent and transitory) of the regret, which ultimately re-conciliates (Faury et al., 2020) with the Bayesian approach of Dong et al. (2019). In contrast to previous works, we find that in the permanent regime non-linearity can dramatically ease the exploration-exploitation trade-off. While it also impacts the length of the transitory phase in a problem-dependent fashion, we show that this impact is mild in most reasonable configurations.

1 INTRODUCTION

Motivation. The Logistic Bandit (**LogB**) model is a sequential decision-making framework that recently received increasing attention in the parametric bandits literature (Li et al., 2010; Dumitrascu et al., 2018; Dong et al., 2019; Faury et al., 2020). This interest can reasonably be attributed to the practical advantages of Logistic Bandits over Linear Bandits (**LB**) (Dani et al., 2008; Abbasi-Yadkori et al., 2011) and to the distinctive learning-theoretical questions that arise in their analysis. On the practical side, **LogB** addresses environments with *binary* rewards (ubiquitous in real-word applications) where it was shown to empirically improve over **LB** approaches (Li et al., 2012). On the theoretical side, **LogB** offers a rigorous framework to study the effects of non-linearity on the exploration-exploitation trade-off for parametrized bandits. It therefore stands as a stepping-stone in generalizing the well-understood **LB** framework to more general and complex reward structures. This particular goal has driven a large part of the research on parametrized bandits, through the study of Generalized Linear Bandits (Filippi et al., 2010; Li et al., 2017) and Kernelized Bandits (Valko et al., 2013; Chowdhury and Gopalan, 2017).

Non-Linearity in LogB. The importance of the non-linearity is fundamentally *problem-dependent* in the **LogB** setting. Interestingly enough, the effects of the non-linearity can be compactly summed-up in a problem-dependent constant, which we will for now denote κ. Intuitively, κ can be understood as a *badness of fit* between the true reward signal and a linear approximation. Given the highly non-linear nature of the logistic function it can become prohibitively large, even for reasonable problem instances. The first known regret upper-bounds for **LogB** were provided by Filippi et al. (2010), scaling as $\widetilde{\mathcal{O}}(\kappa d\sqrt{T})$. This suggests that non-linearity is highly detrimental for the exploration-exploitation trade-off as the more non-linear the reward (*i.e* the bigger κ) the larger the regret.

Recent Work. This conclusion was nuanced by Faury et al. (2020) who introduced an algorithm achieving a regret upper-bound scaling as $\widetilde{\mathcal{O}}(d\sqrt{T}+\kappa)$. Their bound henceforth tells a different story, namely that for large horizons the effect of non-linearity disappears. However, it is not clear if the scaling of the re-

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. ⋆: equal contribution.

gret's first-order term is optimal (w.r.t κ) as to the best of our knowledge there exist no instance-dependent lower-bounds for **LogB**. Furthermore, the presence in the regret bound of a second-order term scaling with κ suggests that the non-linearity can still be particularly harmful for small horizons. A slightly different message on the learning-theoretic difficulties behind the **LogB** was brought by the Bayesian analysis of Dong et al. (2019). They show that in favorable settings the dependency in κ can be removed altogether from the Bayesian regret of Thompson Sampling (whatever the horizon). Yet in worst-case instances (and as κ grows arbitrarily large) their analysis suggests that the problem can remain arbitrarily hard.

Contributions. In this paper, we **(1)** introduce a new algorithm for the Logistic Bandit setting, called **OFULog**. Its analysis distinguishes two regimes of the regret during which the behavior of the algorithm is significantly different: a *long-term* regime and a *transitory* regime. We show that **(2)** in the long-term regime the situation can be much better than what was previously suggested as for a large set of problems the regret scales as $\sqrt{T/\kappa}$. In other words, non-linearity can dramatically ease the exploration-exploitation trade-off. We prove that **(3)** this scaling is optimal by exhibiting a matching *problem-dependent* lower-bound. To the best of our knowledge, this is the first problem-dependent lower-bound for **LogB**. We also **(4)** link the transitory regime to the second-order term in the regret bound of Faury et al. (2020) and to the worst-case analysis of Dong et al. (2019). We show that **(5)** the length of this transitory phase can be much smaller than κ and that **OFULog** can *adapt* to the complexity of the problem to avoid long transitory phases. While the definition of **OFULog** allows for a neat analysis, it can be challenging to implement. To this end, we **(6)** provide a *convex relaxation* of **OFULog**, tractable for finite arm-sets (without sacrificing theoretical guarantees).

2 PRELIMINARIES

Notations Let f and g be two univariate real-valued functions. Throughout the article, we denote $f \lesssim_t g$ or $f = \tilde{\mathcal{O}}(g)$ to indicate that g dominates f up to logarithmic factors. In proof sketches and discussions, we informally use $f \lessapprox g$ to denote $f \leq Cg$ where C is an universal constant. The notation $\dot{f}$ (resp. $\ddot{f}$) will denote the first (resp. second) derivative of f. For any $x \in \mathbb{R}$ we will denote $\|x\|$ its ℓ_2-norm. The notation $\mathcal{B}_d(x, r)$ (resp. $\mathcal{S}_d(x, r)$) will denote the d-dimensional ℓ_2-ball (resp. sphere) centered at x and with radius r. Finally, for two real-valued symmetric matrices A and B, the notation $\mathbf{A} \succeq \mathbf{B}$ indicates that $\mathbf{A}-\mathbf{B}$ is positive semi-definite. When $\mathbf{A}$ is positive semi-definite, we will note $\|x\|_{\mathbf{A}} = \sqrt{x^\mathsf{T}\mathbf{A}x}$. For two scalar a and b, we denote the maximum (resp. minimum) of (a, b) as $a \vee b$ (resp. $a \wedge b$). For an event $E \in \Omega$, we write $E^C = \Omega \backslash E$ and $\mathbb{1}\{E\}$ the indicator function of E.

2.1 Setting

We consider the Logistic Bandit setting, where an agent selects actions (as vectors in $\mathbb{R}^d$) and receives binary, Bernoulli distributed rewards. More precisely at every round t the agent observes an arm-set $\mathcal{X}$ (potentially infinite) and plays an action $x_t \in \mathcal{X}$. She receives a reward r_{t+1} sampled according to a Bernoulli distribution with mean $\mu(x_t^\mathsf{T}\theta_\star)$, where $\mu(z) := (1+e^{-z})^{-1}$ is the *logistic* function, and $\theta_\star \in \mathbb{R}^d$ is *unknown* to the agent. As a result:

$$\mathbb{E}\left[r_{t+1} \,|x_t\right] = \mu\left(x_t^\mathsf{T}\theta_\star\right) \ .$$

The logistic function μ is strictly increasing. It also satisfies a (generalized) *self-concordance* property thanks to the inequality $|\ddot{\mu}| \leq \dot{\mu}$. We will work under the two following standard assumptions.

Assumption 1 (Bounded Arm-Set)**.** *For any $x \in \mathcal{X}$ the following holds:*[1] $\|x\| \leq 1$.

Assumption 2 (Bounded Bandit Parameter)**.** *There exists a* known *constant such that $\|\theta_\star\| \leq S$.*

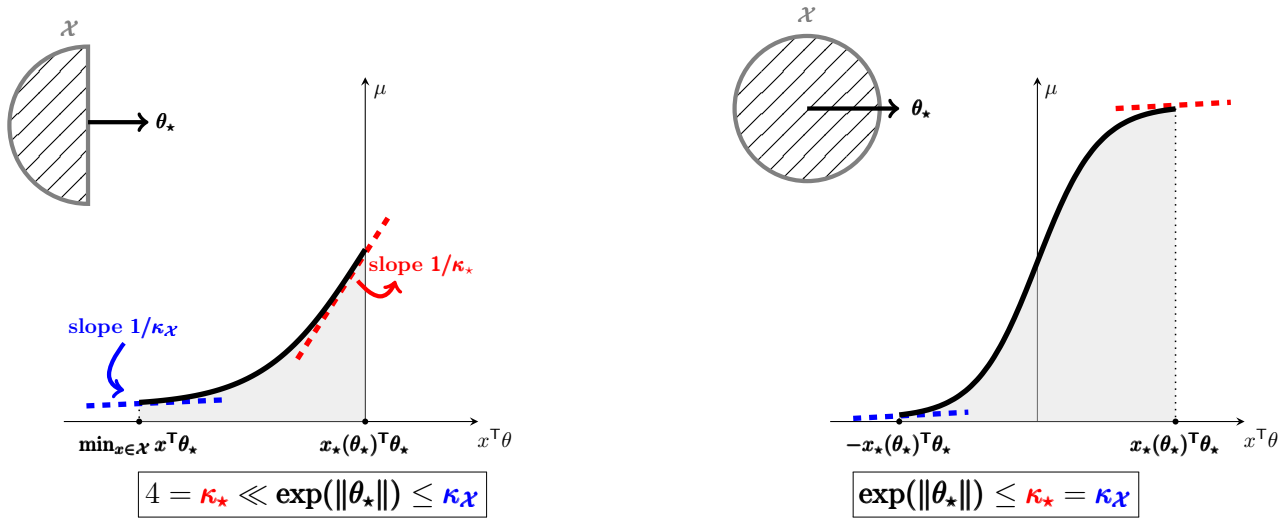
We will denote $\Theta := \mathcal{B}_d(0, S)$. For any $\theta \in \Theta$, we will use the notation $x_\star(\theta) := \arg\max_{x\in\mathcal{X}} x^\mathsf{T}\theta$. At each round t, the agent takes a decision following a policy $\pi : \mathcal{F}_t \to \mathcal{X}$, mapping $\mathcal{F}_t := \sigma(\{x_s, r_{s+1}\}_{s=1}^{t-1})$ (the filtration encoding the information acquired so far) to the arms. The goal of the agent is to minimize her cumulative pseudo-regret up to time T:

$$\text{Regret}_{\theta_\star}^{\pi}(T) := \sum_{t=1}^{T} \mu\left(x_\star(\theta_\star)^\mathsf{T}\theta_\star\right) - \mu\left(x_t^\mathsf{T}\theta_\star\right) \ .$$

We will drop the dependency in π when there is no ambiguity about which policy is considered.

The *conditioning* of μ lies at the center of the analysis of Logistic Bandits. In previous work this conditioning was evaluated through the whole decision-set $\Theta \times \mathcal{X}$ through the problem-dependent quantity $\kappa := \max_{\mathcal{X},\Theta} 1/\dot{\mu}(x^\mathsf{T}\theta)$. In a few words, κ quantifies the level of non-linearity of plausible reward signals and in this sense can be understood as a measure of discrepancy with the linear model. As such, it can be significantly *large* even for reasonable **LogB** problems. We refer the reader to Section 2 of Faury et al. (2020) for a detailed discussion on the importance of this quantity. In this work, we refine the problem-dependant analysis through the use of the following

[1] This assumption is made for ease of exposition, and can easily be relaxed. It can be imposed by re-scaling all actions - which will impact $\|\theta_\star\|$ accordingly.



(a) Assymetric arm-set.

(b) Symmetric arm-set (unit-ball).

Figure 1: Graphical illustration of $\kappa_\star$ and $\kappa_{\mathcal{X}}$ for different decision-sets (top-left). (a) The decision-set spans the left-hand side of the logistic function, $\kappa_{\mathcal{X}}$ and $\kappa_\star$ have (very) different magnitude. (b) The decision-set spans (symmetrically) the whole spectrum of the logistic function, $\kappa_{\mathcal{X}}$ and $\kappa_\star$ have similar magnitudes.

quantities:[2]

$$\kappa_\star(\theta_\star) := 1/\dot{\mu}\left(x_\star(\theta_\star)^\intercal\theta_\star\right) \ ,$$
$$\kappa_{\mathcal{X}}(\theta_\star) := \max_{x\in\mathcal{X}} 1/\dot{\mu}\left(x^\intercal\theta_\star\right) \ .$$

In other words $\kappa_\star$ and $\kappa_{\mathcal{X}}$ measure the *effective* non-linearity around the best action $x_\star(\theta_\star)$ and in the whole parameter-set. Their definitions are illustrated in Figure 1. We have the following ordering: $\kappa_\star \leq \kappa_{\mathcal{X}} \leq \kappa$, with equality between $\kappa_\star$ and $\kappa_{\mathcal{X}}$ for *symmetric* arm-sets (*e.g* $\mathcal{X} = \mathcal{B}_d(0,1)$). Note that the scalings of $\kappa_{\mathcal{X}}$ and κ are fundamentally the same; both grow as $\exp(\|\theta_\star\|)$ and can therefore be *very* large, even in reasonable settings.

2.2 Related Work

Generalized Linear Bandits. Non-linear parametric bandits were first studied by Filippi et al. (2010), who introduced an optimistic algorithm for Generalized Linear Bandits. Their approach was generalized to randomized algorithms (Russo and Van Roy, 2013, 2014; Abeille and Lazaric, 2017) and further refined for the finite-armed setting by Li et al. (2017). Some efforts have also been made to adapt the previous approaches to be fully-online and efficient (Zhang et al., 2016; Jun et al., 2017). All the aforementioned contributions provide regret bounds scaling proportionally to κ, which was recently proven to be sub-optimal for the logistic bandit.

[2]Again, we will drop the dependency in $\theta_\star$ when there is no ambiguity.

Logistic Bandits. Faury et al. (2020) introduced an algorithm which regret bound scales as $\widetilde{\mathcal{O}}(d\sqrt{T} + \kappa d^2)$. This nuances the folk intuition that non-linearity can be only detrimental to the exploration-exploitation trade-off. Indeed, when T is sufficiently large ($T \gtrapprox \kappa^2$) the regret bound is seemingly *independent* of κ and one recovers the regret bound of the **LB** (*e.g* $\widetilde{\mathcal{O}}(d\sqrt{T})$). In other words, the non-linearity no longer plays a part in the exploration-exploitation trade-off. The presence of a second order term (scaling with κd^2) in the regret bound also suggests that under short horizons ($T \lessapprox \kappa^2$) the problem remains *hard* - as the regret bound scales linearly with T. Finally, note that the algorithm of Faury et al. (2020) is impractical: it involves non-convex optimization steps, as well as maintaining a set of constraints (the admissible log-odds) which size grows linearly with time.

A Bayesian Perspective. The nature of the second order term of Faury et al. (2020) and whether it could be improved is still an open question. It is however coherent, to some extent, with the Bayesian analysis of Dong et al. (2019): by letting κ be arbitrarily large (compared to T) they construct arm-sets where no policy can enjoy sub-linear regret. Their construction is particularly worst-case, yet emphasizes that some **LogB** instances are notably hard. On the other hand they also provide a positive result; they exhibit scenarios where the Bayesian regret is upper-bounded by $\sqrt{T}$, *independently* of κ. This stresses that second order dependencies in κ are fundamentally related to the arm-set structure and suggests there is room for improvement.

2.3 Outline and Contributions

In Section 3 we formally introduce **OFULog**, an algorithm for **LogB** based on the **O**ptimism in **F**ace of **U**ncertainty (**OFU**) principle.

We collect our main results in Section 4:
- Theorem 1 provides a regret upper-bound for **OFULog**. It decomposes in two terms $R^{\text{perm}}_{\theta_\star}$ and $R^{\text{trans}}_{\theta_\star}$, each associated with a different regime of the regret: *permanent* and *transitory*. $R^{\text{trans}}_{\theta_\star}$ refines the second-order term of Faury et al. (2020) by introducing the notion of *detrimental* arms, essentially played in a transitory phase. $R^{\text{perm}}_{\theta_\star}$ dominates when T is large and scales as $\widetilde{\mathcal{O}}(d\sqrt{T/\kappa_\star})$.
- Theorem 2 provides a matching problem-dependent lower-bound proving that **OFULog** is minimax-optimal. The main implication is that non-linearity in **LogB** can ease the exploration-exploitation trade-off in the long-term regime, postponing the challenge of non-linearity to the transitory phase.
- Proposition 2 shows that the transitory phase is *short* for reasonable arm-set structures. This confirms that **OFULog**'s second order term ($R^{\text{trans}}_{\theta_\star}$) can be bounded independently of κ. In most unfavorable cases, we retrieve the second order term in Faury et al. (2020).
- Theorem 3 synthesizes the aforementioned improvements. For the commonly studied $\mathcal{X} = \mathcal{B}_d(0,1)$ we prove that **OFULog** enjoys a $\widetilde{\mathcal{O}}(d\sqrt{T/\kappa_{\mathcal{X}}})$ regret.

We provide some intuition behind the proofs of Theorem 1 and Theorem 2 in Section 5.

We address tractability issues in Section 6. In line with previous works **OFULog** requires solving non-convex optimization programs. We circumvent this issue in **OFULog-r** through a *convex* relaxation, at the cost of marginally degrading the regret guarantees.

3 ALGORITHM

3.1 Confidence Set

At the heart of the design of optimistic algorithm is the use of a tight confidence set for $\theta_\star$. We build on Faury et al. (2020) and recall the main ingredients behind its construction. For a *predictable* time-dependent regularizer $\lambda_t > 0$ we define the log-loss as:

$$\mathcal{L}_t(\theta) := -\sum_{s=1}^{t-1} \ell\left(\mu(x_s^\mathsf{T}\theta), r_{s+1}\right) + \lambda_t \|\theta\|^2 \ .$$

where $\ell(x,y) = y\log(x) + (1-y)\log(1-x)$. The log-loss is a strongly convex coercive function and its minimum $\hat{\theta}_t$ is unique and well-defined. We will denote $\mathbf{H_t}(\theta) := \nabla^2\mathcal{L}_t(\theta) \succ 0$ the Hessian of $\mathcal{L}_t$ and:

$$g_t(\theta) := \sum_{s=1}^{t-1} \mu(x_s^\mathsf{T}\theta)x_s + \lambda_t\theta \ .$$

Algorithm 1 OFULog

for $t \geq 1$ **do**
  Set $\lambda_t \leftarrow d\log(t)$.
  (*Learning*) Solve $\hat{\theta}_t = \arg\min_\theta \mathcal{L}_t(\theta)$.
  (*Planning*) Solve $(x_t, \theta_t) \in \arg\max_{\mathcal{X}, \mathcal{C}_t(\delta)} \mu\left(x^\mathsf{T}\theta\right)$.
  Play x_t and observe reward r_{t+1}.
end for

Finally, for $\delta \in (0,1]$ we define:

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

where $\gamma_t(\delta) := \sqrt{\lambda_t}(S + \frac{1}{2}) + \frac{d}{\sqrt{\lambda_t}}\log\left(\frac{4}{\delta}\left(1 + \frac{t}{16d\lambda_t}\right)\right)$. The following proposition ensures that $\mathcal{C}_t(\delta)$ is a confidence set for $\theta_\star$.

Proposition 1 (Lemma 1 in (Faury et al., 2020))**.**

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

The proof is provided in Appendix B and relies on the tail-inequality of (Faury et al., 2020, Theorem 1), adapted to allow time-varying regularizations.[3]

3.2 Algorithm

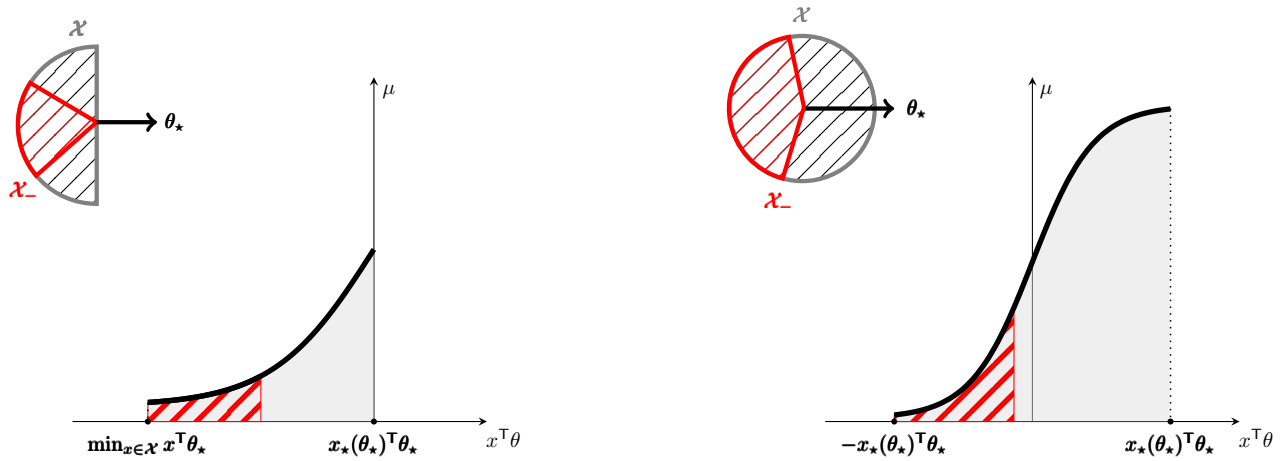
OFULog is the counterpart of the **LB** algorithm **OFUL** of Abbasi-Yadkori et al. (2011). At each round it computes $\hat{\theta}_t$ and the set $\mathcal{C}_t(\delta)$. It then finds an optimistic parameter $\theta_t \in \mathcal{C}_t(\delta)$ and plays x_t the greedy action w.r.t θ_t. Formally:

$$(x_t, \theta_t) \in \underset{x\in\mathcal{X},\, \theta\in\mathcal{C}_t(\delta)}{\arg\max} \ \mu\left(x^\mathsf{T}\theta\right) \ . \tag{1}$$

The pseudo-code for **OFULog** is summarized in Algorithm 1. Notice that we construct $\mathcal{C}_t(\delta)$ with $\lambda_t = d\log(t)$, yielding $\gamma_t(\delta) \lessapprox \sqrt{d\log(t)}$.

Parameter-based versus Bonus-based. **OFULog** and the **LogUCB2** algorithm of Faury et al. (2020) both rely on optimism w.r.t the same confidence set. The main difference resides in how they enforce optimism: optimistic parameter search (**OFULog**) versus exploration bonuses (**LogUCB2**). In contrast with **LB**, the two approaches are not equivalent in a non-linear setting. The parameter-based approach has several key advantages. It (1) allows for a much neater analysis and (2) removes some unnecessary algorithmic complexity. A compelling illustration is that **OFULog** does not require the demanding projection on the set of admissible log-odds of **LogUCB2**. Finally, it (3) yields algorithms that better adapt to the effective complexity of the problem (see Section 4).

[3] Time-varying regularization allows to run **OFULog** without *a-priori* knowledge of the horizon T.



(a) Assymetrical arm-set.

(b) Symmetrical arm-set (unit-ball).

Figure 2: Graphical illustration of $\mathcal{X}_-$.

4 MAIN RESULTS

General Regret Upper-Bound. We first define the set of *detrimental* arms $\mathcal{X}_-$.

Definition (Detrimental arms)**.**

$$\mathcal{X}_- := \begin{vmatrix} \left\{x \in \mathcal{X} \,\middle|\, x^\mathsf{T}\theta_\star \leq -1\right\} & \text{if } x_\star(\theta_\star)^\mathsf{T}\theta_\star > 0\,, \\ \left\{x \in \mathcal{X} \,\middle|\, \dot{\mu}(x^\mathsf{T}\theta_\star) \leq (2\kappa_\star(\theta_\star))^{-1}\right\} & \text{otherwise.} \end{vmatrix}$$

Intuitively, detrimental arms have a large *gap* and carry little *information*. In details, $\mathcal{X}_-$ contains arms x such that $\mu(x^\mathsf{T}\theta_\star) \ll \mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star))$ (large gap) and $\dot{\mu}(x^\mathsf{T}\theta_\star) \approx 0$ (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 $\mathcal{X}_-$ in Figure 2.

Theorem 1 (General Regret Upper-Bound)**.** *The regret of* **OFULog** *satisfies:*

$$\text{Regret}_{\theta_\star}(T) \leq R^{\text{perm}}_{\theta_\star}(T) + R^{\text{trans}}_{\theta_\star}(T)\,,$$

where with high-probability:

$$R^{\text{perm}}_{\theta_\star}(T) \lesssim_T d\sqrt{\frac{T}{\kappa_\star}} \quad \textit{and}$$

$$R^{\text{trans}}_{\theta_\star}(T) \lesssim_T \kappa_{\mathcal{X}} d^2 \wedge \left(d^2 + \sum_{t=1}^{T} \mathbb{1}\left(x_t \in \mathcal{X}_-\right)\right)\,.$$

The proof is deferred to Appendix C.1.

Remark (On the definition of $\mathcal{X}_-$)**.** *We use two alternative definitions for $\mathcal{X}_-$ depending on the sign of $x_\star(\theta_\star)^\mathsf{T}\theta_\star$. 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.*

Problem-Dependent Long-Term Regret. A striking consequence of Theorem 1 arise for large values of the horizon T, when the dominating term is $R^{\text{perm}}_{\theta_\star}(T)$ scaling as $d\sqrt{T/\kappa_\star}$. This is in sharp contrast with previous results as it highlights that non-linearity impacts the first-order regret's term in a positive sense. Indeed the bigger $\kappa_\star$ (cf. Figure 1b) the smaller the (asymptotic) regret. This bound on the long-term regret is actually quite intuitive; in the asymptotic regime the algorithm mostly plays actions around $x_\star(\theta_\star)$. If the reward signal is *flat* in this region, the regret should scale accordingly. It is therefore natural that the regret scales proportionally with the *local* slope $\dot{\mu}(x_\star(\theta_\star)^\mathsf{T}\theta_\star) = 1/\kappa_\star$.

The Long-Term Regret is Minimax. The scaling for the long-term regret is *optimal*: we present in Theorem 2 a matching lower-bound. In contrast to existing the lower-bounds for **LB** our lower-bound is *local*: for *any* nominal instance $\theta_\star$, no policy can ensure a small regret for both $\theta_\star$ and its hardest nearby alternative.[4] Formally, for a small constant $\epsilon > 0$ let us define the *local* minimax regret:

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

Theorem 2 (Local Lower-Bound)**.** *Let $\mathcal{X} = \mathcal{S}_d(0,1)$. For any problem instance $\theta_\star$ and for $T \geq d^2\kappa_\star(\theta_\star)$, there exists ϵ_T small enough such that:*

$$\text{MinimaxRegret}_{\theta_\star,T}(\epsilon_T) = \Omega\left(d\sqrt{\frac{T}{\kappa_\star(\theta_\star)}}\right)\,.$$

[4] This lower-bound has a similar flavor to the lower-bound of Simchowitz and Foster (2020) in a reinforcement learning setting.

The proof is deferred to Appendix D. The *locality* of our lower-bound is necessary to take into account problem-dependent quantities associated with the reference point $\theta_\star$ (*e.g* $\kappa_\star$). Naturally, this local lower bound implies a bound on the global minimax complexity.

Transitory Regret and Detrimental Arms. We now discuss Theorem 1 for smaller values of the horizon T and turn our attention to $R^{\text{trans}}_{\theta_\star}(T)$. In the worst-case, we retrieve the second order term of Faury et al. (2020) - *i.e* $R^{\text{trans}}_{\theta_\star}(T) \leq d^2\kappa$. However Theorem 1 leaves room for improvement, stressing that $R^{\text{trans}}_{\theta_\star}(T)$ is significantly smaller when detrimental arms $\mathcal{X}_-$ are discarded fast enough. Coherently with the Bayesian analysis of Dong et al. (2019) this is achieved by **OFU-Log** for some arm-set structures.

Proposition 2. *The following holds w.h.p:*

$$R^{\text{trans}}_{\theta_\star}(T) \lesssim_T d^2 + dK \qquad \text{if } |\mathcal{X}_-| \leq K\,, \tag{2}$$
$$R^{\text{trans}}_{\theta_\star}(T) \lesssim_T d^3 \qquad \text{if } \mathcal{X} = \mathcal{B}_d(0,1)\,. \tag{3}$$

This result formalizes that **OFULog** quickly discards detrimental arms when (2) there are only a few or (3) the problem's structure is symmetric. The proof is deferred to Appendix C.3.

Remark (Adaptivity)**. OFULog** *effectively* adapts *to the complexity of the problem at hand: its transitory regime varies from d^2 to $\kappa_{\mathcal{X}}d^2$ depending on the arm-set's geometry. To obtain similar behavior, bonus-based approaches (e.g* **LogUCB2***) must hard-code this complexity in the bonus, requiring one design per setting.*

Unit Ball Case. The following result embodies the improvement brought by our analysis; both the regret's first-order and second terms are dramatically smaller than in previous approaches (by an order of $\exp(-\|\theta_\star\|)$).

Theorem 3 (Unit-Ball Regret Upper-Bound)**.** *If $\mathcal{X} = \mathcal{B}_d(0,1)$ the regret of* **OFULog** *satisfies:*

$$\text{Regret}_{\theta_\star}(T) \lesssim_T d\sqrt{\frac{T}{\kappa_{\mathcal{X}}}} + d^2 \quad w.h.p\,.$$

5 HIGH LEVEL IDEAS

5.1 Key Arguments behind Theorem 1

We provide here the main ideas behind the proof of Theorem 1. We assume that the high probability event $\{\theta_\star \in \mathcal{C}_t(\delta)\}$ holds. The optimistic nature of the pair (x_t, θ_t) along with a second-order Taylor expansion of the regret yields:

$$\text{Regret}_{\theta_\star}(T) \leq \underbrace{\sum_{t=1}^{T} \dot\mu(x_t^\mathsf{T}\theta_\star)x_t^\mathsf{T}(\theta_t - \theta_\star)}_{R^{\text{perm}}_{\theta_\star}(T)} + \underbrace{\sum_{t=1}^{T} \ddot\mu(z_t)\{\theta_\star^\mathsf{T}(x_\star(\theta_\star) - x_t)\}^2}_{R^{\text{trans}}_{\theta_\star}(T)}\,.$$

where $z_t \in [x_t^\mathsf{T}\theta_\star, x_\star(\theta_\star)^\mathsf{T}\theta_\star]$.

We start by examining $R^{\text{perm}}_{\theta_\star}(T)$. Leveraging the self-concordance property of the logistic function (cf Appendix F) and the structure of $\mathcal{C}_t(\delta)$ one gets:

$$R^{\text{perm}}_{\theta_\star}(T) \lesssim_T \sqrt{d}\sum_{t=1}^{T} \dot\mu(x_t^\mathsf{T}\theta_\star)\,\|x_t\|_{\mathbf{H_t}^{-1}(\theta_\star)}\,,$$
$$\lesssim_T d\sqrt{\sum_{t=1}^{T} \dot\mu(x_t^\mathsf{T}\theta_\star)}\,.$$

where we last used the Elliptical Potential Lemma (cf. Appendix G) and Cauchy-Schwarz inequality.
A brutal bound of the type $\dot\mu \leq 1/4$ yields $R^{\text{perm}}_{\theta_\star}(T) \lesssim_T d\sqrt{T}$ and retrieves the first order term in (Faury et al., 2020). This bound is however considerably *loose*: an asymptotically optimal strategy often plays $x_\star(\theta_\star)$ (or relatively close actions). Therefore most of the time $\dot\mu(x_t^\mathsf{T}\theta_\star) \approx \dot\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star) = \kappa_\star^{-1}$. Formalizing this intuition (cf. Appendix C.1) yields:

$$R^{\text{perm}}_{\theta_\star}(T) \lessapprox d\sqrt{\frac{T}{\kappa_\star}}\,.$$

We now investigate $R^{\text{trans}}_{\theta_\star}(T)$. First, note that a crude upper-bound directly yields an explicit dependency in $\kappa_{\mathcal{X}}$: from the boundedness of $|\ddot\mu|$ one obtains

$$R^{\text{trans}}_{\theta_\star}(T) \lessapprox d\sum_{t=1}^{T} \|x_t\|^2_{\mathbf{H}_t^{-1}(\theta_\star)} \lessapprox d^2\kappa_{\mathcal{X}}\,.$$

where we used $\mathbf{H_t}(\theta_\star) \succeq \kappa_{\mathcal{X}}^{-1}\sum_{s=1}^{t-1} x_s x_s^\mathsf{T}$ along with the Elliptical Potential Lemma. While it may be unimprovable in some cases, this bound is particularly pessimistic as it discards the good cases where $\ddot\mu(z_t)$ and $\mathbf{H_t}(\theta_\star)$ compensate each other.
We first illustrate this fact with an extreme argument: if $x_t^\mathsf{T}\theta_\star \geq 0$ for all t then $z_t \geq 0$ and $\ddot\mu(z_t) \leq 0$. In this case we obtain $R^{\text{trans}}_{\theta_\star}(T) \leq 0$. This suggests that in more general scenarios the arms $\mathcal{X}$ should be classified depending on their position w.r.t $\theta_\star$. Along with the previous example, this idea hints towards decomposing

$R^{\text{trans}}_{\theta_\star}(T)$ as follows:

$$R^{\text{trans}}_{\theta_\star}(T) \leq \sum_{t=1}^{T} \ddot{\mu}(z_t)\{\theta_\star^\intercal(x_\star(\theta_\star) - x_t)\}^2 \mathbb{1}\left\{x_t^\intercal \theta_\star \leq 0\right\} \ ,$$
$$\lessapprox \sum_{t=1}^{T} \mathbb{1}\left\{x_t^\intercal \theta_\star \leq 0\right\} \ .$$

where we last used the self-concordance of μ. The main point of this last inequality is that $R^{\text{trans}}_{\theta_\star}(T)$ is linked to the number of times the algorithm played detrimental arms. As long as there are few such actions one can therefore expect a good algorithm to have a small associated $R^{\text{trans}}_{\theta_\star}(T)$ - this is the point of Proposition 2. The illustrative discussion we are displaying here is formalized in Theorem 1 by introducing a finer and more general definition for detrimental arms $\mathcal{X}_-$.

5.2 Key Arguments behind Theorem 2

We discuss here the construction of our local lower-bound. Let $\theta_\star$ denote a fixed nominal instance and π a policy which has low-regret when playing against $\theta_\star$. Our strategy is to find an alternative problem θ' which satisfies the two following *conflicting* criteria: **(1)** π has the same behavior against both $\theta_\star$ and θ' and **(2)** θ' is *far* from $\theta_\star$ so that the optimal arms $x_\star(\theta_\star)$ and $x_\star(\theta')$ significantly *differ*.

When playing against $\theta_\star$, we can expect π to produce a trajectory where most of the time $x_t \approx x_\star(\theta_\star)$. Indeed since:

$$\text{Regret}^{\pi}_{\theta_\star}(T) \propto \sum_{t=1}^{T} \|x_t - x_\star(\theta_\star)\|^2 \ ,$$

a small regret against $\theta_\star$ implies an accurate tracking of $x_\star(\theta_\star)$. Notice that when $\mathcal{X} = \mathcal{B}_d(0,1)$ we have $x_\star(\theta_\star)$ is co-linear with $\theta_\star$. As a consequence there are $d-1$ directions (orthogonal to $\theta_\star$) where $\theta_\star$ is poorly estimated. This suggest that parameters laying in $\mathcal{H}^\star_\perp$ (the hyperplane supported by $\theta_\star$, cf. Figure 3) can easily be confused with $\theta_\star$ for the policy π. This notion of *distinguishability* between parameters can be formalized through a discrepancy measures $d_T(\theta_\star, \theta')$ which quantifies how easy it is for π to determine if the rewards it receives are generated by either $\theta_\star$ or θ'. For any $\theta' \in \mathcal{H}^\star_\perp$ it scales as follow:

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

This scaling is rather intuitive; the larger T, the more occasions for π to separate $\theta_\star$ from θ'. Further, the larger $\kappa_\star$, the smaller the conditional variance of the rewards and the longer it takes to correctly estimate an arm's mean reward and determine wether it was

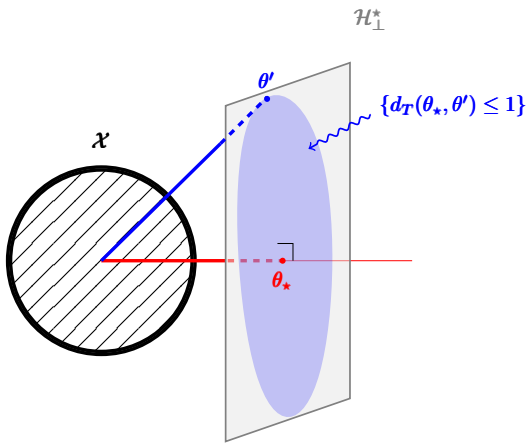


Figure 3: Illustration of the construction behind the local lower-bound.

generated by $\theta_\star$ or θ'. To satisfy **(1)** we must choose θ' so that $d_T(\theta_\star, \theta')$ is small; the trade-off with **(2)** suggests picking θ' such that:

$$\|\theta' - \theta_\star\|^2 \approx \sqrt{\frac{\kappa_\star(\theta_\star)}{T}} \tag{4}$$

Under such conditions, π cannot separate $\theta_\star$ from θ' and must therefore *act* similarly against both parameters (i.e most of the time we will have $x_t \approx x_\star(\theta_\star)$ against θ'). Easy computations show that the regret of π against θ' then writes:

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

which gives the announced behavior after replacing $\|\theta_\star - \theta'\|$ by the scaling suggested by the trade-off between **(1)** and **(2)** presented in Equation (4).

6 TRACTABILITY THROUGH CONVEX RELAXATION

The optimization program presented in Equation (1) and to be solved by **OFULog** is challenging. Indeed, the constraint $\theta \in \mathcal{C}_t(\delta)$ is non-convex and therefore there exist no standard approach for provably approximately solving this program.

A Convex Relaxation. We circumvent this issue by designing a *convex relaxation* for the set $\mathcal{C}_t(\delta)$:

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

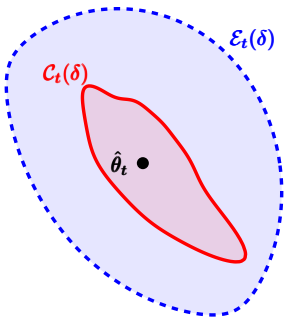


Figure 4: The confidence set $\mathcal{C}_t(\delta)$ and its convex relaxation $\mathcal{E}_t(\delta)$ obtained through a trajectory with: $T = 1000$, $\mathcal{X} = \mathcal{B}_d(0, 1)$ and $\kappa_{\mathcal{X}} = 22$.

where $\beta_t(\delta) := \gamma_t(\delta) + \gamma_t^2(\delta)/\sqrt{\lambda_t}$. The convexity of the log-loss immediatly implies that $\mathcal{E}_t(\delta)$ is convex (illustrated in Figure 4). The following statement ensures that **(1.)** it does relax the confidence set $\mathcal{C}_t(\delta)$ yet **(2.)** preserves core concentration guarantees.

Lemma 1. *The following statements hold:*

1. $\mathcal{C}_t(\delta) \subseteq \mathcal{E}_t(\delta)$.
2. $\forall \theta \in \mathcal{E}_t(\delta)$: $\|\theta - \theta_\star\|_{\mathbf{H}_t(\theta_\star)} = \mathcal{O}(\sqrt{d \log(t)})$ *w.h.p.*

The proof is deferred to Appendix B.3.

Relaxing the Optimistic Planning. Building on $\mathcal{E}_t(\delta)$ we obtain **OFULog-r** where the planning is performed as follows:

$$(x_t, \tilde{\theta}_t) \in \underset{x \in \mathcal{X}, \theta \in \mathcal{E}_t(\delta)}{\arg\max} \; x^\mathsf{T} \theta \; . \tag{5}$$

Note the similarities with the **OFUL** algorithm of Abbasi-Yadkori et al. (2011); the planning consists in the minimization of a *bilinear* objective under *convex* constraints. While solving the program presented in Equation (5) remains challenging in general, a tractable procedure can be developed for finite arm-sets - summarized in Algorithm 2. The following proposition guarantees that it effectively guarantees optimism.

Proposition 3. *Let $(\tilde{x}_t, \tilde{\theta}_t)$ be the pair returned by Algorithm 2. Then:*

$$(\tilde{x}_t, \tilde{\theta}_t) \in \underset{x \in \mathcal{X}, \theta \in \mathcal{E}_t(\delta)}{\arg\max} \; x^\mathsf{T} \theta \; .$$

The main complexity of Algorithm 2 reduces to maximizing a linear objective under convex constraints. The maximizer can therefore be found efficiently by solving the dual problem.

Regret Guarantees. We conclude this section with Corollary 1 proving that relaxing the original optimistic search does not impact the learning performances thus recovering the guarantees of **OFULog**.

Algorithm 2 Planning for **OFULog-r**

input: finite arm-set $\mathcal{X}$, set $\mathcal{E}_t(\delta)$.
for $x \in \mathcal{X}$ **do**
  Solve $\theta_x \leftarrow \arg\max_{\theta \in \mathcal{E}_t(\delta)} x^\mathsf{T} \theta$.
end for
Compute $\tilde{x} \leftarrow \arg\max_{x \in \mathcal{X}} x^\mathsf{T} \theta_x$.
return $(\tilde{x}, \theta_{\tilde{x}})$.

Corollary 1. *Theorem 1, Proposition 2 and Theorem 3 are also satisfied by* **OFULog-r**.

This claim directly follows from Lemma 1.

7 CONCLUSION

In this paper we bring forward an improved characterization of the regret minimization problem in Logistic Bandit through the lense of **OFULog**, a parameter-based optimistic algorithm. Our analysis further describes the impact of non-linearity on the exploration-exploitation trade-off. For a large number of settings, we show that non-linearity *eases* regret minimization in **LogB**. This is embodied by the $\widetilde{\mathcal{O}}(\sqrt{T/\kappa_\star})$ upper-bound of **OFULog**, which we show is optimal by proving a matching, local and problem-dependent lower-bound. Such rates are however conditioned on reaching a permanent regime. The regret associated with the transitory phase acts as a second-order term tied to problem-dependent quantities.

Generalized Linear Bandits. Part of the findings presented here can be easily extended to other generalized linear bandits (namely the $\widetilde{\mathcal{O}}(\sqrt{T/\kappa_\star})$ rate) however with potentially different conclusions. The findings related to the transitory regime are however specific to Logistic Bandits. In general, we believe that attempting to treat all generalized linear bandits in a model-agnostic approach is sub-optimal for a fine characterization of the non-linearity's effect. This should be done in a problem-dependent fashion, relative and specific to the considered model and the singularities behind its non-linear nature.

Efficient Algorithms. An interesting avenue for future work resides in modifying the arguments presented here to develop order-optimal yet fully online algorithms for **LogB**. Jointly achieving efficiency and regret minimax-optimality is still an open question. Improving guarantees for online logistic regression (under a well-specification assumption) and marrying them with our analysis seems like a promising direction to complete this goal.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling Revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

Sayak Ray Chowdhury and Aditya Gopalan. On Kernelized Multi-Armed Bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 844–853, 2017.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic Linear Optimization under Bandit Feedback. In *Conference on Learning Theory*, 2008.

Shi Dong, Tengyu Ma, and Benjamin Van Roy. On the Performance of Thompson Sampling on Logistic Bandits. In *Conference on Learning Theory*, pages 1158–1160, 2019.

Bianca Dumitrascu, Karen Feng, and Barbara Engelhardt. PG-TS: Improved Thompson Sampling for Logistic Contextual Bandits. In *Advances in Neural Information Processing Systems*, pages 4624–4633, 2018.

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, ICML 2020*, 2020.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric Bandits: the Generalized Linear Case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable Generalized Linear Bandits: Online Computation and Hashing. In *Advances in Neural Information Processing Systems*, pages 99–109, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Lihong Li, Wei Chu, John Langford, Taesup Moon, and Xuanhui Wang. An Unbiased Offline Evaluation of Contextual Bandit Algorithms with Generalized Linear Models. In *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, pages 19–36, 2012.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably Optimal Algorithms for Generalized Linear Contextual Bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2071–2080. JMLR. org, 2017.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, pages 12040–12049, 2019.

Daniel Russo and Benjamin Van Roy. Eluder Dimension and the Sample Complexity of Optimistic Exploration. In *Advances in Neural Information Processing Systems*, pages 2256–2264, 2013.

Daniel Russo and Benjamin Van Roy. Learning to Optimize via Posterior Sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

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.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

Michal Valko, Nathan Korda, Rémi Munos, Ilias Flaounas, and Nello Cristianini. Finite-Time Analysis of Kernelised Contextual Bandits. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 654–663, 2013.

Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi-Hua Zhou. Online Stochastic Linear Optimization under One-Bit Feedback. In *International Conference on Machine Learning*, pages 392–401, 2016.

Instance-Wise Minimax-Optimal Algorithms for Logistic Bandits Supplementary Material

ORGANIZATION OF THE APPENDIX

This appendix is organized as follows:

- In Appendix A we introduce useful notations, and introduce some central inequalities.
- In Appendix B we prove that $\mathcal{C}_t(\delta)$ and $\mathcal{E}_t(\delta)$ are confidence sets for $\theta_\star$.
- In Appendix C we prove the different regret upper-bounds announced in the main manuscript.
- In Appendix D we prove the regret lower-bound.
- In Appendix E we give some guarantees for the optimistic solving of **OFULog-r**.
- In Appendix F we prove some key self-concordance results.
- In Appendix G we introduce and prove some auxiliary results, needed for the analysis.
- In Appendix H we display illustrative numerical experiments.

Table of Contents

A NOTATIONS AND FIRST INEQUALITIES

We collect here a list of symbols and definitions that will be used throughout this appendix. Recall the definition of the regularized logistic loss given a sequence of vectors $\{x_i\}_{i=1}^{t-1}$, rewards $\{r_i\}_{i=2}^{t}$ and a (predictable) regularization parameter λ_t:

$$\mathcal{L}_t(\theta) := -\sum_{s=1}^{t-1} \left[r_{s+1} \log \mu(x_s^\mathsf{T}\theta) + (1 - r_{s+1}) \log(1 - \mu(x_s^\mathsf{T}\theta)) \right] + \frac{\lambda_t}{2} \|\theta\|^2 \ .$$

$\mathcal{L}_t(\theta)$ being a strictly convex and coercive function, we can safely define $\hat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \mathcal{L}_t(\theta)$. Define also for all $\theta \in \mathbb{R}^d$:

$$g_t(\theta) := \sum_{s=1}^{t-1} \mu(x_s^\mathsf{T}\theta) x_s + \lambda_t \theta \ , \qquad \mathbf{H_t}(\theta) := \sum_{s=1}^{t-1} \dot{\mu}(x_s^\mathsf{T}\theta) x_s x_s^\mathsf{T} + \lambda_t \mathbf{I_d} \ .$$

For all $x, \theta_1, \theta_2 \in \mathbb{R}^d$ let:

$$\begin{aligned}
\alpha(x, \theta_1, \theta_2) &:= \int_{v=0}^{1} \dot{\mu}\left(x^\mathsf{T}\theta_1 + v x^\mathsf{T}(\theta_2 - \theta_1)\right) dv \ , \\
\tilde{\alpha}(x, \theta_1, \theta_2) &:= \int_{v=0}^{1} (1-v) \dot{\mu}\left(x^\mathsf{T}\theta_1 + v x^\mathsf{T}(\theta_2 - \theta_1)\right) dv \ , \\
\mathbf{G_t}(\theta_1, \theta_2) &:= \sum_{s=1}^{t-1} \alpha(x_s, \theta_1, \theta_2) x_s x_s^\mathsf{T} + \lambda_t \mathbf{I_d} \ , \\
\widetilde{\mathbf{G}}_\mathbf{t}(\theta_1, \theta_2) &:= \sum_{s=1}^{t-1} \tilde{\alpha}(x_s, \theta_1, \theta_2) x_s x_s^\mathsf{T} + \lambda_t \mathbf{I_d} \ .
\end{aligned}$$

Note that since μ is strictly increasing (and therefore $\dot{\mu} \geq 0$) we easily have $\alpha(x, \theta, \theta_2) \geq \tilde{\alpha}(x, \theta, \theta_2)$. It easily follows that $\mathbf{G_t}(\theta_1, \theta_2) \succeq \widetilde{\mathbf{G}}_\mathbf{t}(\theta_1, \theta_2)$. Thanks to the mean-value theorem, we also have for all θ_1, θ_2:

$$g_t(\theta_1) - g_t(\theta_2) = \mathbf{G_t}(\theta_1, \theta_2)(\theta_1 - \theta_2) \ . \tag{6}$$

Also, thanks to Lemmas 7 and 8 we have the following inequalities for any $\theta_1, \theta_2 \in \Theta$:

$$\mathbf{G_t}(\theta_1, \theta_2) \succeq (1 + 2S)^{-1} \mathbf{H_t}(\theta) \text{ for } \theta \in \{\theta_1, \theta_2\} \ , \tag{7}$$

$$\widetilde{\mathbf{G}}_\mathbf{t}(\theta_1, \theta_2) \succeq (2 + 2S)^{-1} \mathbf{H_t}(\theta_1) \ . \tag{8}$$

We will also use the notation:

$$\mathbf{V}_t := \sum_{s=1}^{t-1} x_s x_s^\mathsf{T} + \lambda_t \mathbf{I_d}$$

Thanks to the inequality $\dot{\mu}(x^\mathsf{T}\theta_1) \geq \kappa_{\mathcal{X}}^{-1}(\theta_1)$ for any $x \in \mathcal{X}$ along with $\kappa_{\mathcal{X}}(\theta_1) > 1$ for any θ_1, we have:

$$\mathbf{H}_t(\theta_1) \succeq \kappa_{\mathcal{X}}^{-1}(\theta_1) \mathbf{V_t} \tag{9}$$

B CONFIDENCE SETS

B.1 Concentration Inequality

Our results build on the concentration inequality of (Faury et al., 2020, Theorem 1). We present below a marginally modified version inspired from the proof of Theorem 1 in (Russac et al., 2019), which allows for time-varying (yet predictable) regularization (without ressorting to union bounds). In time, this will allow us to design near-optimal algorithms without the knowledge of the horizon T.

Theorem 4. *Let $\{\mathcal{F}_t\}_{t=1}^{\infty}$ be a filtration. Let $\{x_t\}_{t=1}^{\infty}$ be a stochastic process in $\mathcal{B}_2^d(1)$ such that x_t is $\mathcal{F}_t$-measurable. Let $\{\varepsilon_t\}_{t=2}^{\infty}$ be a real-valued martingale difference sequence such that ε_t is $\mathcal{F}_t$-measurable. Further, assume $|\varepsilon_t| \leq 1$ holds almost surely for all $t \geq 2$ and denote $\sigma_t^2 = \mathbb{E}\left[\varepsilon_t^2 | \mathcal{F}_t\right]$. Let $\{\lambda_t\}_{t=1}^{\infty}$ be a predictable sequence of non-negative scalars. Define:*

$$S_t := \sum_{s=1}^{t-1} \varepsilon_s x_s \,, \qquad \mathbf{H_t} = \sum_{s=1}^{t-1} \sigma_s^2 x_s x_s^\mathsf{T} + \lambda_t \mathbf{I_d} \,.$$

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

$$\mathbb{P}\left(\exists t \in \mathbb{N} \ s.t \ \|S_t\|_{\mathbf{H_t^{-1}}} \geq \frac{2}{\sqrt{\lambda_t}} \log\left(\frac{2^d \lambda_t^{-d/2} \det\left(\mathbf{H_t}\right)^{1/2}}{\delta} \right) + \frac{\sqrt{\lambda_t}}{2} \right) \leq \delta \,.$$

Proof. The proof essentially follows the proof of Theorem 1 in Faury et al. (2020), up to a minor modification to allow for a time-varying regularization. In the following, denote $\bar{\mathbf{H}}_\mathbf{t} := \sum_{s=1}^{t-1} \sigma_s^2 x_s x_s^\mathsf{T}$ and for all $\xi \in \mathcal{B}_d(0,1)$ let:

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

We know thanks to Lemma 5 of Faury et al. (2020) that $M_t(\xi)$ is a super-martingale and hence checks $\mathbb{E}\left[M_t(\xi)\right] \leq 1$ 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 our analysis differs from (Faury et al., 2020). This however doesn't hurt the final result as one can still apply an appropriate stopping time construction. Let τ be a stopping time with respect to $\{\mathcal{F}_t\}_t$. 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 $\mathbb{E}\left[M_\tau(\xi)\right] \leq 1$ for all $\xi \in \mathcal{B}_d(0,1/2)$, . Clearly we have:

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

Following the proof of Theorem 1 in Faury et al. (2020), computing $\bar{M}_\tau$ 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. □

B.2 Confidence Set

Recall the confidence set definition:

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

where:

$$\gamma_t(\delta) = \sqrt{\lambda_t}(S + \frac{1}{2}) + \frac{d}{\sqrt{\lambda_t}} \log\left(\frac{4}{\delta}\left(1 + \frac{t}{16 d \lambda_t}\right) \right) \,. \tag{10}$$

Proposition 1 (Lemma 1 in (Faury et al., 2020)). *Let $\delta \in (0,1]$ and define:*

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

Then $\mathbb{P}\Big(\forall t \geq 1, \theta_\star \in \mathcal{C}_t(\delta)\Big) = \mathbb{P}\left(E_\delta\right) \geq 1-\delta$.

Proof. We trivially have:

$$\left\{ \forall t \geq 1,\ \theta_\star \in \mathcal{C}_t(\delta) \right\} = E_\delta$$

From the optimality conditions of $\hat{\theta}_t$ one easily gets that $g_t(\hat{\theta}_t) = \sum_{s=1}^{t-1} r_{s+1} x_s$. Therefore:

$$\begin{aligned}
\left\| g_t(\hat{\theta}_t) - g_t(\theta_\star) \right\|_{\mathbf{H_t}^{-1}(\theta_\star)} &= \left\| \sum_{s=1}^{t-1} \left(r_{s+1} - \mu(x_s^\mathsf{T}\theta_\star) \right) x_s - \lambda_t \theta_\star \right\|_{\mathbf{H_t}^{-1}(\theta_\star)} \\
&\leq \sqrt{\lambda_t} S + \left\| \sum_{s=1}^{t-1} \varepsilon_{s+1} x_s \right\|_{\mathbf{H_t}^{-1}(\theta_\star)} , \qquad (\|\theta_\star\| \leq S,\ \mathbf{H_t}(\theta_\star) \succeq \lambda_t \mathbf{I_d})
\end{aligned}$$

where we defined for all $s \geq 1$: $\varepsilon_{s+1} := r_{s+1} - \mu(x_s^\mathsf{T}\theta_\star)$. Remember that conditionally on $\mathcal{F}_s$ the rewards are such that $r_{s+1} \sim \text{Bernoulli}(\mu(x_s^\mathsf{T}\theta_\star))$. Therefore:

$$\begin{cases}
\mathbb{E}\left[\epsilon_{s+1} | \mathcal{F}_s\right] = 0 , \\
\mathbb{V}\text{ar}\left[\epsilon_{s+1} | \mathcal{F}_s\right] = \mu(x_s^\mathsf{T}\theta_\star)(1-\mu(x_s^\mathsf{T}\theta_\star)) = \dot{\mu}(x_s^\mathsf{T}\theta_\star) .
\end{cases}$$

If we define $S_t := \sum_{s=1}^{t-1} \varepsilon_{s+1} x_s$ and $\mathbf{H_t} = \mathbf{H_t}(\theta_\star)$ all conditions of Theorem 4 are met and we have:

$$\begin{aligned}
1-\delta &\geq \mathbb{P}\left(\forall t \geq 1,\ \|S_t\|_{\mathbf{H_t}^{-1}(\theta_\star)} \leq \frac{2}{\sqrt{\lambda_t}} \log\left(\frac{2^d \lambda_t^{-d/2} \det\left(\mathbf{H_t}\right)^{1/2}}{\delta} \right) + \frac{\sqrt{\lambda_t}}{2} \right) \\
&\geq \mathbb{P}\left(\forall t \geq 1,\ \|S_t\|_{\mathbf{H_t}^{-1}(\theta_\star)} \leq \gamma_t(\delta) - \sqrt{\lambda_t} S \right) \\
&= \mathbb{P}\left(\forall t \geq 1,\ \sqrt{\lambda_t} S + \left\| \sum_{t=1}^{s-1} \epsilon_{s+1} x_s \right\|_{\mathbf{H_t}^{-1}(\theta_\star)} \leq \gamma_t(\delta) \right) \qquad \text{(def. of } S_t\text{)} \\
&= \mathbb{P}\left(\forall t \geq 1,\ \left\| g_t(\hat{\theta}_t) - g_t(\theta_\star) \right\|_{\mathbf{H_t}^{-1}(\theta_\star)} \leq \gamma_t(\delta) \right) = \mathbb{P}(E_\delta)
\end{aligned}$$

where the second inequality results from simple upper-bounding and the use of Lemma 11. □

B.3 Convex Relaxation

Recall the definition:

$$\mathcal{E}_t(\delta) = \left\{ \theta \in \Theta \,\middle|\, \mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t) \leq \beta_t(\delta)^2 \right\} \quad \text{where } \beta_t(\delta) = \gamma_t(\delta) + \gamma_t^2(\delta)/\sqrt{\lambda_t} . \tag{11}$$

We recall and prove Lemma 1 (we provide here a more detailed version than in the main manuscript).

Lemma 1. *The following statements hold:*

1. $\mathcal{C}_t(\delta) \subseteq \mathcal{E}_t(\delta)$ *for all* $t \geq 1$ *and therefore* $\mathbb{P}\left(\forall t \geq 1,\ \theta_\star \in \mathcal{E}_t(\delta)\right) \geq 1-\delta$.
2. *With probability at least* $1-\delta$, *we have:*

$$\forall \theta \in \mathcal{E}_t(\delta),\ \|\theta - \theta_\star\|_{\mathbf{H_t}(\theta_\star)} \leq (2+2S)\gamma_t(\delta) + 2\sqrt{1+S}\beta_t(\delta) .$$

Therefore if $\lambda_t = d\log(t)$ *with probability at least* $1-\delta$:

$$\forall \theta \in \mathcal{E}_t(\delta),\ \|\theta - \theta_\star\|_{\mathbf{H_t}(\theta_\star)} = \widetilde{\mathcal{O}}(\sqrt{d\log(t)}) .$$

Proof. We start by proving that $\mathcal{C}_t(\delta) \subseteq \mathcal{E}_t(\delta)$. First, we claim Lemma 2, which proof is deferred to Appendix B.4.

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

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

Thanks to exact second-order Taylor expansion of the logistic loss, we have that for all $\theta \in \mathbb{R}^d$:

$$\mathcal{L}_t(\theta) = \mathcal{L}_t(\hat{\theta}_t) + \nabla\mathcal{L}_t(\hat{\theta}_t)^\mathsf{T}(\theta - \hat{\theta}_t) + (\theta - \hat{\theta}_t)^\mathsf{T}\left(\int_{v=0}^1 (1-v)\nabla^2\mathcal{L}_t(\hat{\theta}_t + v(\theta - \hat{\theta}_t))dv\right)(\theta_\star - \hat{\theta}_t)\,.$$

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

$$\begin{aligned}
\mathcal{L}_t(\theta) &= \mathcal{L}_t(\hat{\theta}_t) + (\theta - \hat{\theta}_t)^\mathsf{T}\left(\int_{v=0}^1 (1-v)\nabla^2\mathcal{L}_t(\hat{\theta}_t + v(\theta - \hat{\theta}_t))dv\right)(\theta_\star - \hat{\theta}_t) \\
&= \mathcal{L}_t(\hat{\theta}_t) + (\theta - \hat{\theta}_t)^\mathsf{T}\left(\int_{v=0}^1 (1-v)\mathbf{H_t}(\hat{\theta}_t + v(\theta - \hat{\theta}_t))dv\right)(\theta_\star - \hat{\theta}_t) && (\nabla^2\mathcal{L}_t = \mathbf{H_t}) \\
&= \mathcal{L}_t(\hat{\theta}_t) + \left\|\theta - \hat{\theta}_t\right\|^2_{\widetilde{\mathbf{G}}_\mathbf{t}(\hat{\theta}_t,\theta)} && (\text{def. of } \widetilde{\mathbf{G}}_\mathbf{t}(\hat{\theta}_t,\theta)) \\
&\leq \mathcal{L}_t(\hat{\theta}_t) + \left\|\theta - \hat{\theta}_t\right\|^2_{\mathbf{G_t}(\hat{\theta}_t,\theta)} && (\widetilde{\mathbf{G}}_\mathbf{t} \leq \mathbf{G_t}) \\
&= \mathcal{L}_t(\hat{\theta}_t) + \left\|g_t(\theta) - g_t(\hat{\theta}_t)\right\|^2_{\mathbf{G_t}^{-1}(\hat{\theta}_t,\theta)} && (\text{Equation (6)})\,. \\
&= \mathcal{L}_t(\hat{\theta}_t) + \left\|g_t(\theta) - g_t(\hat{\theta}_t)\right\|^2_{\mathbf{G_t}^{-1}(\hat{\theta}_t,\theta)} && (\mathbf{G_t}(\hat{\theta}_t,\theta) = \mathbf{G_t}(\theta,\hat{\theta}_t))\,.
\end{aligned}$$

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

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

proving that $\theta \in \mathcal{C}_t(\delta) \Rightarrow \theta \in \mathcal{E}_t(\delta)$ and therefore $\mathcal{C}_t(\delta) \subseteq \mathcal{E}_t(\delta)$.

We now prove the second part of Lemma 1. We will assume that E_δ holds, which happens with probability at least $1-\delta$ (cf. Proposition 1). We rely on the following second-order Taylor expansion. For all $\theta \in \mathcal{E}_t(\delta)$:

$$\begin{aligned}
\mathcal{L}_t(\theta) &= \mathcal{L}_t(\theta_\star) + (\theta - \theta_\star)^\mathsf{T}\nabla\mathcal{L}_t(\theta_\star) + (\theta - \theta_\star)^\mathsf{T}\left(\int_{v=0}^1 (1-v)\nabla^2\mathcal{L}_t(\theta_\star + v(\theta - \theta_\star))dv\right)(\theta - \theta_\star) \\
&= \mathcal{L}_t(\theta_\star) + (\theta - \theta_\star)^\mathsf{T}\nabla\mathcal{L}_t(\theta_\star) + \|\theta - \theta_\star\|^2_{\widetilde{\mathbf{G}}_\mathbf{t}(\theta_\star,\theta)}
\end{aligned}$$

Therefore:

$$\begin{aligned}
\mathcal{L}_t(\theta) - \mathcal{L}_t(\theta_\star) - (\theta - \theta_\star)^\mathsf{T}\nabla\mathcal{L}_t(\theta_\star) &= \|\theta - \theta_\star\|^2_{\widetilde{\mathbf{G}}_\mathbf{t}(\theta_\star,\theta)} \\
&\geq (2+2S)^{-1}\|\theta - \theta_\star\|^2_{\mathbf{H_t}(\theta_\star)} && (\text{Equation (8)})
\end{aligned}$$

which can be rewritten as:

$$\begin{aligned}
\|\theta - \theta_\star\|^2_{\mathbf{H_t}(\theta_\star)} &\leq (2+2S)\left|\mathcal{L}_t(\theta) - \mathcal{L}_t(\theta_\star)\right| + (2+2S)\left|(\theta - \theta_\star)^\mathsf{T}\nabla\mathcal{L}_t(\theta_\star)\right| \\
&\leq 2(2+2S)\beta_t(\delta)^2 + (2+2S)\left|(\theta - \theta_\star)^\mathsf{T}\nabla\mathcal{L}_t(\theta_\star)\right| && (\theta, \theta_\star \in \mathcal{E}_t(\delta)) \\
&\leq 2(2+2S)\beta_t(\delta)^2 + (2+2S)\|\theta - \theta_\star\|_{\mathbf{H_t}(\theta_\star)}\|\nabla\mathcal{L}_t(\theta_\star)\|_{\mathbf{H_t}^{-1}(\theta_\star)} && (\text{Cauchy-Schwartz}) \\
&\leq 2(2+2S)\beta_t(\delta)^2 + (2+2S)\gamma_t(\delta)\|\theta - \theta_\star\|_{\mathbf{H_t}(\theta_\star)}
\end{aligned}$$

where we last used:

$$\begin{aligned}\|\nabla\mathcal{L}_t(\theta_\star)\|_{\mathbf{H_t}^{-1}(\theta_\star)} &= \left\|g_t(\theta_*) - \sum_{s=1}^{t-1} r_{s+1}x_s\right\|_{\mathbf{H_t}^{-1}(\theta_\star)} \\ &= \left\|g_t(\theta_\star) - g_t(\hat{\theta}_t)\right\|_{\mathbf{H_t}^{-1}(\theta_\star)} \\ &\leq \gamma_t(\delta)\ . && (E_\delta \text{ holds})\end{aligned}$$

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

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

Solving it (cf. Proposition 7) yields:

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

Finally, note that when $\lambda_t = d\log(t)$ we obtain the following scalings:

$$\begin{aligned}&\gamma_t(\delta) = \mathcal{O}(\sqrt{d\log(t)})\ , && \text{(Equation (10))}\\ &\beta_t(\delta) = \gamma_t(\delta) + \gamma_t^2(\delta)/\sqrt{\lambda_t} = \mathcal{O}(\sqrt{d\log(t)})\ . && \text{(Equation (11))}\end{aligned}$$

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

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

This holds as soon as E_δ does, which happens with probability at least $1-\delta$. □

B.4 Proof of Lemma 2

Lemma 2. *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 Lemma 7 we have:

$$\begin{aligned}\mathbf{G_t}(\theta,\hat{\theta}_t) &= \sum_{s=1}^{t-1}\alpha(x_s,\theta,\hat{\theta}_t)x_sx_s^\mathsf{T} + \lambda_t\mathbf{I_d} \\ &\geq \sum_{s=1}^{t-1}\left(1+|x_s^\mathsf{T}(\theta-\hat{\theta}_t)|\right)^{-1}\dot{\mu}(x_s^\mathsf{T}\theta)x_sx_s^\mathsf{T} + \lambda_t\mathbf{I_d} && (\text{Lemma 7})\\ &\geq \sum_{s=1}^{t-1}\left(1+\|x_s\|_{\mathbf{G_t}^{-1}(\theta,\hat{\theta}_t)}\left\|\theta-\hat{\theta}_t\right\|_{\mathbf{G_t}(\theta,\hat{\theta}_t)}\right)^{-1}\dot{\mu}(x_s^\mathsf{T}\theta)x_sx_s^\mathsf{T} + \lambda_t\mathbf{I_d} && (\text{Cauchy-Schwartz})\\ &\geq \left(1+\lambda_t^{-1/2}\left\|\theta-\hat{\theta}_t\right\|_{\mathbf{G_t}(\theta,\hat{\theta}_t)}\right)^{-1}\sum_{s=1}^{t-1}\dot{\mu}(x_s^\mathsf{T}\theta)x_sx_s^\mathsf{T} + \lambda_t\mathbf{I_d} && (\mathbf{G_t}(\theta,\hat{\theta}_t) \geq \lambda_t\mathbf{I_d})\\ &\geq \left(1+\lambda_t^{-1/2}\left\|\theta-\hat{\theta}_t\right\|_{\mathbf{G_t}(\theta,\hat{\theta}_t)}\right)^{-1}\left(\sum_{s=1}^{t-1}\dot{\mu}(x_s^\mathsf{T}\theta)x_sx_s^\mathsf{T} + \lambda_t\mathbf{I_d}\right)\\ &= \left(1+\lambda_t^{-1/2}\left\|\theta-\hat{\theta}_t\right\|_{\mathbf{G_t}(\theta,\hat{\theta}_t)}\right)^{-1}\mathbf{H_t}(\theta)\\ &= \left(1+\lambda_t^{-1/2}\left\|g_t(\theta)-g_t(\hat{\theta}_t)\right\|_{\mathbf{G_t}^{-1}(\theta,\hat{\theta}_t)}\right)^{-1}\mathbf{H_t}(\theta) && (\text{Equation (6)})\end{aligned}$$

Using this inequality, we therefore obtain that:

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

Solving this polynomial inequality in $\left\|g_t(\theta) - g_t(\hat{\theta}_t)\right\|_{\mathbf{G_t^{-1}}(\theta,\hat{\theta}_t)}$ (cf. Proposition 7) yields :

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

which proves the announced result. □

C REGRET UPPER-BOUNDS

C.1 Proof of Theorem 1

Theorem 1 (General Regret Upper-Bound)**.** *The regret of* **OFULog** *satisfies:*

$$\mathrm{Regret}_{\theta_\star}(T) \leq R^{\mathrm{perm}}_{\theta_\star}(T) + R^{\mathrm{trans}}_{\theta_\star}(T)\,,$$

where with probability at least $1-\delta$*:*

$$R^{\mathrm{perm}}_{\theta_\star}(T) \lesssim_T d\sqrt{\frac{T}{\kappa_\star}} \qquad \textit{and} \qquad R^{\mathrm{trans}}_{\theta_\star}(T) \lesssim_T \kappa_{\mathcal{X}} d^2 \wedge \left(d^2 + \mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)\sum_{t=1}^{T} \mathbb{1}\left(x_t \in \mathcal{X}_-\right)\right)\,.$$

Proof. In the following, we assume the good event $\{\forall t \geq 1, \theta_\star \in \mathcal{C}_t(\delta)\}$ to hold, which happens with probability at least $1-\delta$ according to Proposition 1.

Recall the strategy followed by **OFULog**:

$$(x_t, \theta_t) \in \underset{x\in\mathcal{X}, \theta\in\mathcal{C}_t(\delta)}{\arg\max}\; x^\mathsf{T}\theta\,.$$

and therefore under the good event we have $x_\star(\theta_\star)^\mathsf{T}\theta_\star \leq x_t^\mathsf{T}\theta_t$. We will need the following result, which proof is postponed to Appendix C.2.

Proposition 4. *If* $\theta_* \in \mathcal{C}_t(\delta)$ *then for all* $\theta \in \mathcal{C}_t(\delta)$*:*

$$\|\theta - \theta_\star\|_{\mathbf{H_t}(\theta_\star)} \leq 2(1+2S)\gamma_t(\delta)$$

We start by performing a second Taylor expansion of the regret.

$$\begin{aligned}
\mathrm{Regret}_{\theta_\star}(T) &= \sum_{t=1}^{T} \mu\Big(x_\star(\theta_\star)^\mathsf{T}\theta_\star\Big) - \mu\Big(x_t^\mathsf{T}\theta_\star\Big) \\
&= \sum_{t=1}^{T} \dot\mu\left(x_t^\mathsf{T}\theta_\star\right)\left(x_\star(\theta_\star) - x_t\right)^\mathsf{T}\theta_\star + \sum_{t=1}^{T}\left[\int_{v=0}^{1}(1-v)\ddot\mu\left(x_t^\mathsf{T}\theta_\star + v(x_\star(\theta_\star)-x_t)^\mathsf{T}\theta_\star\right)dv\right]\left\{(x_\star(\theta_\star)-x_t)^\mathsf{T}\theta_\star\right\}^2 \\
&= \underbrace{\sum_{t=1}^{T} \dot\mu\left(x_t^\mathsf{T}\theta_\star\right)\left(x_\star(\theta_\star) - x_t\right)^\mathsf{T}\theta_\star}_{R_1(T)} + \underbrace{\sum_{t=1}^{T}\tilde\vartheta_t\left\{(x_\star(\theta_\star)-x_t)^\mathsf{T}\theta_\star\right\}^2}_{R_2(T)}\,.
\end{aligned}$$

where we defined:

$$\tilde\vartheta_t = \int_{v=0}^{1}(1-v)\ddot\mu\left(x_t^\mathsf{T}\theta_\star + v(x_\star(\theta_\star)-x_t)^\mathsf{T}\theta_\star\right)dv\,. \tag{12}$$

We start by examining $R_1(T)$. We have the following bound:

$$\begin{aligned}
R_1(T) &= \sum_{t=1}^{T} \dot\mu\left(x_t^\mathsf{T}\theta_\star\right)\left(x_\star(\theta_\star) - x_t\right)^\mathsf{T}\theta_\star \\
&\leq \sum_{t=1}^{T} \dot\mu\left(x_t^\mathsf{T}\theta_\star\right) x_t^\mathsf{T}(\theta_t - \theta_\star) && (x_t^\mathsf{T}\theta_t \geq x_\star(\theta_\star)^\mathsf{T}\theta_\star \text{ since } E_\delta \text{ holds}) \\
&= \sum_{t=1}^{T} \dot\mu\left(x_t^\mathsf{T}\theta_\star\right)\|x_t\|_{\mathbf{H_t}^{-1}(\theta_\star)}\|\theta_t - \theta_\star\|_{\mathbf{H_t}(\theta_\star)} && (\text{Cauchy-Schwarz}) \\
&\leq 2(1+2S)\sum_{t=1}^{T}\gamma_t(\delta)\dot\mu\left(x_t^\mathsf{T}\theta_\star\right)\|x_t\|_{\mathbf{H_t}^{-1}(\theta_\star)} && (\text{Proposition 4}, E_\delta \text{ holds}) \\
&\leq 2(1+2S)\bar\gamma_T(\delta)\sum_{t=1}^{T}\dot\mu\left(x_t^\mathsf{T}\theta_\star\right)\|x_t\|_{\mathbf{H_t}^{-1}(\theta_\star)}
\end{aligned}$$

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

In the following, we denote $\tilde{x}_t := \sqrt{\dot{\mu}(x_t^\mathsf{T}\theta_\star)}x_t$ and $\widetilde{\mathbf{V}}_\mathbf{t} := \sum_{s=1}^{t-1} \tilde{x}_s\tilde{x}_s^\mathsf{T} + \lambda_t \mathbf{I_d} = \mathbf{H_t}(\theta_\star)$. We have:

$$
\begin{aligned}
R_1(T) &\le 2(1+2S)\bar{\gamma}_T(\delta) \sum_{t=1}^T \dot{\mu}\left(x_t^\mathsf{T}\theta_\star\right) \|x_t\|_{\mathbf{H_t^{-1}}(\theta_\star)} \\
&\le 2(1+2S)\bar{\gamma}_T(\delta) \sqrt{\sum_{t=1}^T \dot{\mu}\left(x_t^\mathsf{T}\theta_\star\right)} \sqrt{\sum_{t=1}^T \dot{\mu}\left(x_t^\mathsf{T}\theta_\star\right) \|x_t\|^2_{\mathbf{H_t^{-1}}(\theta_\star)}} && \text{(Cauchy-Schwarz)} \\
&\le 2(1+2S)\bar{\gamma}_T(\delta) \sqrt{\sum_{t=1}^T \dot{\mu}\left(x_t^\mathsf{T}\theta_\star\right)} \sqrt{\sum_{t=1}^T \|\tilde{x}_t\|^2_{\widetilde{\mathbf{V}}_\mathbf{t}^{-1}}} \\
&\le 4(1+2S)\bar{\gamma}_T(\delta) \sqrt{d\log\left(\lambda_T + \frac{T}{16d}\right)} \sqrt{\sum_{t=1}^T \dot{\mu}\left(x_t^\mathsf{T}\theta_\star\right)} && \text{(Lemma 12)} \\
&\le C_1 d\log(T) \sqrt{\sum_{t=1}^T \dot{\mu}\left(x_t^\mathsf{T}\theta_\star\right)}
\end{aligned}
$$

where C_1 is a universal (more precisely, independent of $\kappa_\mathcal{X}(\theta_\star)$, d and T), and where we used that $\bar{\gamma}_T(\delta) = \mathcal{O}(\sqrt{d\log(T)})$ since $\lambda_t = d\log(t)$.

Finally, note that by a first-order Taylor expansion of $\dot{\mu}$:

$$
\begin{aligned}
\sum_{t=1}^T \dot{\mu}\left(x_t^\mathsf{T}\theta_\star\right) &= \sum_{t=1}^T \dot{\mu}\left(x_\star(\theta_\star)^\mathsf{T}\theta_\star\right) + \sum_{t=1}^T \left[\int_{v=0}^1 \ddot{\mu}\left(x_\star(\theta_\star)^\mathsf{T}\theta_\star + v(x_t - x_\star(\theta_\star))^\mathsf{T}\theta_\star\right) dv\right] (x_t - x_\star(\theta_\star))^\mathsf{T}\theta_\star \\
&= \frac{T}{\kappa_\star(\theta_\star)} + \sum_{t=1}^T \left[\int_{v=0}^1 \ddot{\mu}\left(x_\star(\theta_\star)^\mathsf{T}\theta_\star + v(x_t - x_\star(\theta_\star))^\mathsf{T}\theta_\star\right) dv\right] (x_t - x_\star(\theta_\star))^\mathsf{T}\theta_\star && \text{(def. } \kappa_\star\text{)} \\
&\le \frac{T}{\kappa_\star(\theta_\star)} + \sum_{t=1}^T \left|\left[\int_{v=0}^1 \ddot{\mu}\left(x_\star(\theta_\star)^\mathsf{T}\theta_\star + v(x_t - x_\star(\theta_\star))^\mathsf{T}\theta_\star\right) dv\right] (x_t - x_\star(\theta_\star))^\mathsf{T}\theta_\star\right| \\
&\le \frac{T}{\kappa_\star(\theta_\star)} + \sum_{t=1}^T \left[\int_{v=0}^1 \left|\ddot{\mu}\left(x_\star(\theta_\star)^\mathsf{T}\theta_\star + v(x_t - x_\star(\theta_\star))^\mathsf{T}\theta_\star\right)\right| dv\right] (x_\star(\theta_\star) - x_t)^\mathsf{T}\theta_\star && (x_\star(\theta_\star)^\mathsf{T}\theta_\star \ge x_t^\mathsf{T}\theta_\star) \\
&\le \frac{T}{\kappa_\star(\theta_\star)} + \sum_{t=1}^T \left[\int_{v=0}^1 \dot{\mu}\left(x_\star(\theta_\star)^\mathsf{T}\theta_\star + v(x_t - x_\star(\theta_\star))^\mathsf{T}\theta_\star\right) dv\right] (x_\star(\theta_\star) - x_t)^\mathsf{T}\theta_\star && (|\ddot{\mu}| \le \mu) \\
&\le \frac{T}{\kappa_\star(\theta_\star)} + \sum_{t=1}^T \alpha(\theta_\star, x_\star(\theta_\star), x_t)(x_\star(\theta_\star) - x_t)^\mathsf{T}\theta_\star && \text{(def. } \alpha\text{)} \\
&= \frac{T}{\kappa_\star(\theta_\star)} + \sum_{t=1}^T \mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star) - \mu(x_t^\mathsf{T}\theta_\star) && \text{(mean-value theorem)} \\
&= \frac{T}{\kappa_\star(\theta_\star)} + \mathrm{Regret}_{\theta_\star}(T)
\end{aligned}
$$

Using that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for all $a, b \ge 0$ we obtain the following intermediate bound on $R_1(T)$:

$$
R_1(T) \le C_1 d\log(T) \left(\sqrt{\frac{T}{\kappa_\star(\theta_\star)}} + \sqrt{\mathrm{Regret}_{\theta_\star}(T)}\right) \tag{13}
$$

We now turn our attention to $R_2(T)$. We start with a crude-bound and retrieve Faury et al. (2020) second-order

term. Indeed from $\tilde{\vartheta}_t \leq 1$ we get that:

$$
\begin{aligned}
R_2(T) &\leq \sum_{t=1}^{T} \left\{(x_\star(\theta_\star) - x_t)^\mathsf{T}\theta_\star\right\}^2 \\
&\leq \sum_{t=1}^{T} \left\{x_t^\mathsf{T}(\theta_t - \theta_\star)\right\}^2 && (x_t^\mathsf{T}\theta_t \geq x_\star(\theta_\star)^\mathsf{T}\theta_\star \text{ since } E_\delta \text{ holds}) \\
&\leq \sum_{t=1}^{T} \|x_t\|^2_{\mathbf{H_t^{-1}}(\theta_\star)} \|\theta_t - \theta_\star\|^2_{\mathbf{H_t}(\theta_\star)} && \text{(Cauchy-Schwarz)} \\
&\leq 4(1+2S)^2\bar{\gamma}_T(\delta)^2 \sum_{t=1}^{T} \|x_t\|^2_{\mathbf{H_t^{-1}}(\theta_\star)} && \text{(Proposition 4, } E_\delta \text{ holds)} \\
&\leq 4(1+2S)^2\bar{\gamma}_T(\delta)^2\kappa_{\mathcal{X}}(\theta_\star) \sum_{t=1}^{T} \|x_t\|^2_{\mathbf{V_t^{-1}}} && \text{(Equation (9))} \\
&\leq 16d(1+2S)^2\bar{\gamma}_T(\delta)^2\kappa_{\mathcal{X}}(\theta_\star) \log\left(\lambda_T + \frac{T}{d}\right) && \text{(Lemma 12)}
\end{aligned}
$$

Introducing C_2 another universal constant (independent of d, T and $\kappa_{\mathcal{X}}(\theta_\star)$);

$$
R_2(T) \leq C_2 d^2 \kappa_{\mathcal{X}}(\theta_\star) \log^2(T) \tag{14}
$$

We now refine this bound to take into account detrimental arms. The following always holds:

$$
R_2(T) = \sum_{t=1}^{T} \tilde{\vartheta}_t \left\{(x_\star(\theta_\star) - x_t)^\mathsf{T}\theta_\star\right\}^2 \mathbb{1}\left(x_t \in \mathcal{X}_-\right) + \sum_{t=1}^{T} \tilde{\vartheta}_t \left\{(x_\star(\theta_\star) - x_t)^\mathsf{T}\theta_\star\right\}^2 \mathbb{1}\left(x_t \in \mathcal{X}_+\right) , \tag{15}
$$

with $\mathcal{X}_+ = \mathcal{X} \setminus \mathcal{X}_-$. 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 $\tilde{\vartheta}_t \leq \alpha(\theta_\star, x_\star(\theta_\star), x_t)$ and therefore:

$$
\begin{aligned}
\sum_{t=1}^{T} \tilde{\vartheta}_t \left\{(x_\star(\theta_\star) - x_t)^\mathsf{T}\theta_\star\right\}^2 \mathbb{1}\left(x_t \in \mathcal{X}_-\right) &\leq \sum_{t=1}^{T} \alpha(\theta_\star, x_\star(\theta_\star), x_t) \left\{(x_\star(\theta_\star) - x_t)^\mathsf{T}\theta_\star\right\}^2 \mathbb{1}\left(x_t \in \mathcal{X}_-\right) \\
&\leq S \sum_{t=1}^{T} \alpha(\theta_\star, x_\star(\theta_\star), x_t) \left\{(x_\star(\theta_\star) - x_t)^\mathsf{T}\theta_\star\right\} \mathbb{1}\left(x_t \in \mathcal{X}_-\right) \\
&= S \sum_{t=1}^{T} \left[\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star) - \mu(x_t^\mathsf{T}\theta_\star)\right] \mathbb{1}\left(x_t \in \mathcal{X}_-\right) \\
&\leq S\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star) \sum_{t=1}^{T} \mathbb{1}\left(x_t \in \mathcal{X}_-\right)
\end{aligned} \tag{16}
$$

where we used $\|\theta_\star\| \leq S$ and $\|x\| \leq 1$ (for any $x \in \mathcal{X}$) in the second-inequality and the mean-value theorem for the equality which follows. We now turn to bounding the most-right term in Equation (15). We start with the case $x_\star(\theta_\star)^\mathsf{T}\theta_\star \geq 0$. We therefore look at the following definition for the detrimental arms:

$$
\mathcal{X}_- = \left\{x \in \mathcal{X} \,|\, x^\mathsf{T}\theta_* \leq -1\right\}
$$

Fix t and assume that $x_t \in \mathcal{X}_+$. Note that when $x_t^\mathsf{T}\theta_\star \geq 0$ we inherit $\tilde{\vartheta}_t \leq 0$ from the fact that $\ddot{\mu}(z) \leq 0$ for all $z \geq 0$. Using this fact ($\ddot{\mu} \leq 0$ on $\mathbb{R}^+$) we can show that when $x_t^\mathsf{T}\theta_\star \leq 0$:

$$
\begin{aligned}
\tilde{\vartheta}_t &\leq \int_{v=0}^{1} (1-v)\ddot{\mu}\left((1-v)x_t^\mathsf{T}\theta_\star\right) dv \\
&\leq \int_{v=0}^{1} \dot{\mu}\left((1-v)x_t^\mathsf{T}\theta_\star\right) dv && (\ddot{\mu} \leq |\ddot{\mu}| \leq \dot{\mu}) \\
&\leq \dot{\mu}(x_t^\mathsf{T}\theta_\star) \int_{v=0}^{1} \exp\left(v|x_t^\mathsf{T}\theta_\star|\right) && \text{(Lemma 9)} \\
&\leq e^1 \dot{\mu}(x_t^\mathsf{T}\theta_\star) && (-1 \leq x_t^\mathsf{T}\theta_\star \leq 0)
\end{aligned}
$$

where in the last inequality we used $x_t^\mathsf{T}\theta_\star \geq -1$ since $x_t \in \mathcal{X}_+$. Packing this results together we showed that:

$$\begin{aligned}\tilde{\vartheta}_t \mathbb{1}(x_t \in \mathcal{X}_+) &\leq e^1 \dot{\mu}(x_t^\mathsf{T}\theta_*)\mathbb{1}(x_t \in \mathcal{X}_+, x_t^\mathsf{T}\theta_\star \leq 0) + 0 \cdot \mathbb{1}(x_t \in \mathcal{X}_+, x_t^\mathsf{T}\theta_\star \geq 0)\\ &\leq e^1 \dot{\mu}(x_t^\mathsf{T}\theta_*)\mathbb{1}(x_t \in \mathcal{X}_+)\\ &\leq e^1 \dot{\mu}(x_t^\mathsf{T}\theta_*)\end{aligned}$$

Therefore we obtain:

$$\begin{aligned}\sum_{t=1}^T \tilde{\vartheta}_t \left\{(x_\star(\theta_\star) - x_t)^\mathsf{T}\theta_\star\right\}^2 \mathbb{1}\left(x_t \in \mathcal{X}_+\right) &\leq e^1 \sum_{t=1}^T \dot{\mu}(x_t^\mathsf{T}\theta_\star)\left\{(x_\star(\theta_\star) - x_t)^\mathsf{T}\theta_\star\right\}^2\\ &\leq e^1 \sum_{t=1}^T \dot{\mu}(x_t^\mathsf{T}\theta_\star)\left\{x_t^\mathsf{T}(\theta_t - \theta_\star)\right\}^2 && \text{(optimism)}\\ &\leq 4e^1(1+2S)^2\bar{\gamma}_T^2(\delta)\sum_{t=1}^T \dot{\mu}(x_t^\mathsf{T}\theta_\star)\left\|x_t\right\|^2_{\mathbf{H_t}^{-1}(\theta_\star)}\end{aligned}$$

Using Lemma 12 with $\tilde{x}_t = \sqrt{\dot{\mu}(x_t^\mathsf{T}\theta_\star)}x_t$ and $\widetilde{\mathbf{V}}_\mathbf{t} := \sum_{s=1}^{t-1}\tilde{x}_s\tilde{x}_s^\mathsf{T} + \lambda_t \mathbf{I_d} = \mathbf{H_t}(\theta_\star)$ finally yields:

$$\sum_{t=1}^T \tilde{\vartheta}_t \left\{(x_\star(\theta_\star) - x_t)^\mathsf{T}\theta_\star\right\}^2 \mathbb{1}\left(x_t \in \mathcal{X}_+\right) \leq 16de^1(1+2S)^2\bar{\gamma}_T^2 \log\left(\lambda_T + \frac{T}{16d}\right) \tag{17}$$

We now consider the case $x_\star(\theta_\star)^\mathsf{T}\theta_\star \leq 0$. The definition of $\mathcal{X}_-$ becomes:

$$\mathcal{X}_- = \left\{x \,\middle|\, \dot{\mu}(x^\mathsf{T}\theta_*) \leq (2\kappa_\star(\theta_\star))^{-1}\right\} = \left\{x \,\middle|\, \dot{\mu}(x^\mathsf{T}\theta_*) \leq \dot{\mu}(x_\star(\theta_\star)^\mathsf{T}\theta_\star)/2\right\}\ .$$

Fix t and assume that $x_t \in \mathcal{X}_+$. Thanks to $|\ddot{\mu}| \leq \dot{\mu}$:

$$\begin{aligned}\tilde{\vartheta}_t &\leq \alpha(\theta_\star, x_\star(\theta_\star), x_t)\\ &\leq \dot{\mu}(x_\star(\theta_\star)^\mathsf{T}\theta_*) && (x_t^\mathsf{T}\theta_\star \leq x_\star(\theta_\star)^\mathsf{T}\theta_\star \leq 0 \text{ and } \dot{\mu} \text{ increasing on } \mathbb{R}^-)\\ &\leq 2\dot{\mu}(x_t^\mathsf{T}\theta_\star) && (x \in \mathcal{X}_+)\end{aligned}$$

Therefore we obtain:

$$\begin{aligned}\sum_{t=1}^T \tilde{\vartheta}_t \left\{(x_\star(\theta_\star) - x_t)^\mathsf{T}\theta_\star\right\}^2 \mathbb{1}\left(x_t \in \mathcal{X}_+\right) &\leq 2\sum_{t=1}^T \dot{\mu}(x_t^\mathsf{T}\theta_\star)\left\{(x_\star(\theta_\star) - x_t)^\mathsf{T}\theta_\star\right\}^2\\ &\leq 2\sum_{t=1}^T \dot{\mu}(x_t^\mathsf{T}\theta_\star)\left\{x_t^\mathsf{T}(\theta_\star - \theta_t)\right\}^2 && \text{(optimism)}\\ &\leq 2\sum_{t=1}^T \dot{\mu}(x_t^\mathsf{T}\theta_\star)\left\|x_t\right\|^2_{\mathbf{H_t}^{-1}(\theta_\star)}\left\|\theta_t - \theta_\star\right\|^2_{\mathbf{H_t}(\theta_\star)} && \text{(Cauchy-Schwarz)}\\ &\leq 8(1+2S)^2\bar{\gamma}_T^2(\delta)\sum_{t=1}^T \dot{\mu}(x_t^\mathsf{T}\theta_\star)\left\|x_t\right\|^2_{\mathbf{H_t}^{-1}(\theta_\star)} && \text{(Proposition 4)}\end{aligned}$$

Using Lemma 12 again yields:

$$\sum_{t=1}^T \tilde{\vartheta}_t \left\{(x_\star(\theta_\star) - x_t)^\mathsf{T}\theta_\star\right\}^2 \mathbb{1}\left(x_t \in \mathcal{X}_+\right) \leq 32d(1+2S)^2\bar{\gamma}_T^2 \log\left(\lambda_T + \frac{T}{16d}\right) \tag{18}$$

Assembling Equation (15)-(16)-(17)-(18) we obtain that:

$$R_2(T) \leq C_3 d^2 \log^2(T) + C_4 \mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)\sum_{t=1}^T \mathbb{1}(x_t \in \mathcal{X}_-)$$

where C_3 and C_4 constants independent of d, T and $\kappa_{\mathcal{X}}$. Merging this result with Equation (14) finally yields:

$$R_2(T) \leq \left[C_2 d^2 \kappa_{\mathcal{X}}(\theta_\star)\log^2(T)\right] \wedge \left[C_3 d^2 \log^2(T) + C_4 \mu(x_\star(\theta_\star)^\intercal \theta_\star)\sum_{t=1}^{T} \mathbb{1}(x_t \in \mathcal{X}_-)\right] \tag{19}$$

We are now ready to finish the proof of Theorem 1. From the decomposition $\text{Regret}_{\theta_\star}(T) = R_1(T) + R_2(T)$ and Equation (13) we have:

$$\text{Regret}_{\theta_\star}(T) \leq C_1 d\log(T)\sqrt{\frac{T}{\kappa_\star(\theta_\star)}} + C_1 d\log(T)\sqrt{\text{Regret}_{\theta_\star}(T)} + R_2(T)$$

This is a second-order polynomial inequation in $\sqrt{\text{Regret}_{\theta_\star}(T)}$. Solving it (cf. Proposition 7) yields:

$$\sqrt{\text{Regret}_{\theta_\star}(T)} \leq C_1 d\log(T) + \sqrt{C_1 d\log(T)\sqrt{\frac{T}{\kappa_\star(\theta_\star)}} + R_2(T)}$$

Using $(a+b) \leq 2(a^2+b^2)$ we obtain:

$$\text{Regret}_{\theta_\star}(T) \leq 2C_1^2 d^2 \log^2(T) + 2C_1 d\log(T)\sqrt{\frac{T}{\kappa_\star(\theta_\star)}} + 2R_2(T)$$

We obtain the announced inequality after plugging Equation (19) in this last inequality. Indeed, ignoring universal constants we obtain:

$$\text{Regret}_{\theta_\star}(T) \leq d\log(T)\sqrt{\frac{T}{\kappa_\star(\theta_\star)}} + d^2\log^2(T) + \left[d^2\kappa_{\mathcal{X}}(\theta_\star)\log^2(T)\right] \wedge \left[d^2\log^2(T) + \mu(x_\star(\theta_\star)^\intercal\theta_\star)\sum_{t=1}^{T}\mathbb{1}(x_t \in \mathcal{X}_-)\right]$$

Slightly re-arranging:

$$\text{Regret}_{\theta_\star}(T) \leq \underbrace{d\log(T)\sqrt{\frac{T}{\kappa_\star(\theta_\star)}}}_{R^{\text{perm}}_{\theta_\star}(T)} + \underbrace{\left[d^2(\kappa_{\mathcal{X}}(\theta_\star)+1)\log^2(T)\right] \wedge \left[2d^2\log^2(T) + \mu(x_\star(\theta_\star)^\intercal\theta_\star)\sum_{t=1}^{T}\mathbb{1}(x_t \in \mathcal{X}_-)\right]}_{R^{\text{trans}}_{\theta_\star}(T)}$$

which finishes the proof. □

C.2 Proof of Proposition 4

Proposition 4. *If $\theta_* \in \mathcal{C}_t(\delta)$ then for all $\theta \in \mathcal{C}_t(\delta)$:*

$$\|\theta - \theta_\star\|_{\mathbf{H_t}(\theta_\star)} \leq 2(1+2S)\gamma_t(\delta)$$

Proof. Let $\theta \in \mathcal{C}_t(\delta)$.

$$\begin{aligned}
\|\theta - \theta_\star\|_{\mathbf{H_t}(\theta_\star)} &\leq \sqrt{1+2S}\,\|\theta - \theta_\star\|_{\mathbf{G_t}(\theta_\star,\theta)} && (\theta_\star, \theta \in \Theta,\ \text{Equation (7)})\\
&= \sqrt{1+2S}\,\|g_t(\theta) - g_t(\theta_\star)\|_{\mathbf{G_t}^{-1}(\theta_\star,\theta)} && (\text{Equation (6)})\\
&\leq \sqrt{1+2S}\left(\left\|g_t(\theta) - g_t(\hat{\theta}_t)\right\|_{\mathbf{G_t}^{-1}(\theta_\star,\theta)} + \left\|g_t(\theta_\star) - g_t(\hat{\theta}_t)\right\|_{\mathbf{G_t}^{-1}(\theta_\star,\theta)}\right)\\
&\leq (1+2S)\left(\left\|g_t(\theta) - g_t(\hat{\theta}_t)\right\|_{\mathbf{H_t}^{-1}(\theta)} + \left\|g_t(\theta_\star) - g_t(\hat{\theta}_t)\right\|_{\mathbf{H_t}^{-1}(\theta_\star)}\right) && (\theta_\star, \theta \in \Theta,\ \text{Equation (7)})\\
&\leq 2(1+2S)\gamma_t(\delta) && (\theta, \theta_\star \in \mathcal{C}_t(\delta))
\end{aligned}$$

which proves the announced result. □

C.3 Proof of Proposition 2

Proposition 2. *The following holds w.h.p:*

$$R^{\text{trans}}_{\theta_\star}(T) \lesssim_T d^2 + dK \qquad \text{if } |\mathcal{X}_-| \leq K\,, \tag{2}$$

$$R^{\text{trans}}_{\theta_\star}(T) \lesssim_T d^3 \qquad \text{if } \mathcal{X} = \mathcal{B}_d(0,1)\,. \tag{3}$$

C.3.1 Proof of Equation (2)

Proof. We assume the event $E_\delta = \{\forall t \geq 1, \theta_\star \in \mathcal{C}_t(\delta)\}$ holds - this happens with high probability (cf. Proposition 1). To bound $R^{\text{trans}}_{\theta_\star}(T)$ we will start from the bound given in the detailed version of Theorem 1 in Appendix C.1, that is with C_1 and C_2 being universal constants:

$$R^{\text{trans}}_{\theta_\star}(T) \leq C_1 d^2 \log^2(T) + C_2 \mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star) \sum_{t=1}^{T} \mathbb{1}\left(x_t \in \mathcal{X}_-\right) \tag{20}$$

Assume that there is a finite number of detrimental arms, *i.e* $|\mathcal{X}_-| = K < \infty$. We will separate three cases 1. $x_\star(\theta_\star)^\mathsf{T}\theta_\star \geq 0$ and 2. $x_\star(\theta_\star)^\mathsf{T}\theta_\star \leq -1$. and 3. $x_\star(\theta_\star)^\mathsf{T}\theta_\star \in [-1, 0]$. Note that 2. and 3. are sub-cases of the more general $x_\star(\theta_\star)^\mathsf{T}\theta_\star \leq 0$. We separate them here to simplify the analysis.

Case 1. $x_\star(\theta_\star)^\mathsf{T}\theta_\star \geq 0$. In this setting we have:

$$\mathcal{X}_- = \left\{x \in \mathcal{X} \middle| x^\mathsf{T}\theta_\star \leq -1\right\}$$

This implies that detrimental arms have a large (constant) gap. Indeed for any $x \in \mathcal{X}_-$:

$$\begin{aligned}\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star) - \mu(x^\mathsf{T}\theta_\star) &\geq \mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star) - \mu(-1) \\ &\geq 1/2 - \mu(-1)\end{aligned}$$

which yields that:

$$\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star) - \mu(x^\mathsf{T}\theta_\star) \geq 1/5 \tag{21}$$

We can use this result to show that **OFULog** plays detrimental arms only logarithmically often. Indeed, for any $x \in \mathcal{X}_-$ let τ_x be the last time-step when x is played, and N_x the number of time x was played over the whole horizon. Formally:

$$\tau_x = \max_t \left\{t \in [T] \,|\, x_t = x\right\} \quad \text{and} \quad N_x = \sum_{t=1}^{T} \mathbb{1}(x_t = x) = \sum_{t=1}^{\tau_x} \mathbb{1}(x_t = x)\,.$$

Fix $x \in \mathcal{X}_-$ and let $\tau = \tau_x$ (*i.e* $x_\tau = x$). Thanks to Equation (21) and the mean-value theorem:

$$\begin{aligned}
1/5 &\leq \mu\left(x_\star(\theta_\star)^\mathsf{T}\theta_\star\right) - \mu\left(x_\tau^\mathsf{T}\theta_\star\right) \\
&\leq \mu\left(x_\tau^\mathsf{T}\theta_\tau\right) - \mu\left(x_\tau^\mathsf{T}\theta_\star\right) && \text{(optimism, } E_\delta \text{ holds)} \\
&\leq \alpha(x_\tau, \theta_\tau, \theta_\star) x_\tau^\mathsf{T}(\theta_\tau - \theta_\star) && \text{(mean-value theorem)} \\
&= \alpha(x_\tau, \theta_\tau, \theta_\star) x_\tau^\mathsf{T} \mathbf{G}_\tau^{-1}(\theta_\tau, \theta_\star)\left(g_\tau(\theta_\tau) - g_\tau(\theta_\star)\right) && \text{(Equation (6))} \\
&\leq \alpha(x_\tau, \theta_\tau, \theta_\star) \left\|x_\tau\right\|_{\mathbf{G}_\tau^{-1}(\theta_\tau, \theta_\star)} \left\|g_\tau(\theta_\tau) - g_\tau(\theta_\star)\right\|_{\mathbf{G}_\tau^{-1}(\theta_\tau, \theta_\star)} && \text{(Cauchy-Schwarz)} \\
&\leq 2\sqrt{1 + 2S}\gamma_\tau(\delta)\alpha(x_\tau, \theta_\tau, \theta_\star) \left\|x_\tau\right\|_{\mathbf{G}_\tau^{-1}(\theta_\tau, \theta_\star)} && (22)
\end{aligned}$$

where we last used $\left\|g_t(\theta_t) - g_t(\theta_\star)\right\|_{\mathbf{G}_\mathbf{t}^{-1}(\theta_t, \theta_\star)} \leq 2\sqrt{1 + 2S}\gamma_t(\delta)$ (cf. proof of Proposition 4). Note also that $\mathbf{G}_\tau(\theta_\tau, \theta_\star) \succeq N_x \alpha(x, \theta_\tau, \theta_\star) x x^\mathsf{T} + \lambda_\tau \mathbf{I_d}$. It is therefore easy to show (for instance, using the Sherman-Morison formula) that $\left\|x_\tau\right\|^2_{\mathbf{G}_\tau^{-1}(\theta_\tau, \theta_\star)} \leq (\alpha(x_\tau, \theta_\tau, \theta_*) N_x)^{-1}$. We therefore finally obtain by injecting this into Equation (22):

$$\begin{aligned}
N_x &\leq 100(1 + 2S)\gamma_\tau(\delta)^2 \alpha(x_\tau, \theta_\tau, \theta_\star) \\
&\leq 25(1 + 2S)\gamma_\tau(\delta)^2 && (\alpha \leq \sup \dot{\mu} \leq 1/4)
\end{aligned}$$

Remember that this results holds for *any* $x \in \mathcal{X}_-$. Henceforth from Equation (20):

$$\begin{aligned}
R^{\text{trans}}_{\theta_\star}(T) &\leq C_1 d^2 \log^2(T) + C_2 \mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star) \sum_{t=1}^{T} \mathbb{1}\left(x_t \in \mathcal{X}_-\right) \\
&\leq C_1 d^2 \log^2(T) + C_2 \sum_{t=1}^{T} \mathbb{1}\left(x_t \in \mathcal{X}_-\right) && (\mu \leq 1) \\
&= C_1 d^2 \log^2(T) + C_2 \sum_{t=1}^{T} \sum_{x \in \mathcal{X}_-} \mathbb{1}(x_t = x) \\
&= C_1 d^2 \log^2(T) + C_2 \sum_{x \in \mathcal{X}_-} N_x \\
&\leq C_1 d^2 \log^2(T) + 25 C_2 (1+2S) \sum_{x \in \mathcal{X}_-} \gamma_{\tau_x}(\delta)^2 \\
&\leq C_1 d^2 \log^2(T) + 25 C_2 (1+2S) K \max_{t \in [T]} \gamma_t(\delta)^2 && (|\mathcal{X}_-| = K)
\end{aligned}$$

Using the fact that $\max_{t\in[T]} \gamma_t(\delta) \lesssim_T \sqrt{d \log(T)}$ we obtain the announced result:

$$R^{\text{trans}}_{\theta_\star}(T) \lesssim_T d^2 + dK$$

Case 2. $x_\star(\theta_\star)^\mathsf{T}\theta_\star < -1$. This necessarily implies $x^\mathsf{T}\theta_\star \leq -1$ and $\mu(x^\mathsf{T}\theta_\star) \leq \mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star) \leq \mu(-1) \leq 1/2$ for any $x \in \mathcal{X}$. We start by characterizing the gap of detrimental arms which are now defined by:

$$\mathcal{X}_- = \left\{ x \in \mathcal{X} \,\middle|\, \dot\mu(x^\mathsf{T}\theta_\star) \leq \dot\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)/2 \right\}$$

From $\dot\mu = \mu(1-\mu)$ we get that for any $x \in \mathcal{X}_-$:

$$\begin{aligned}
\mu(x^\mathsf{T}\theta_\star) &\leq \frac{\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)}{2} \frac{1-\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)}{1-\mu(x^\mathsf{T}\theta_\star)} \\
&\leq \mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)/2 && (\mu(x^\mathsf{T}\theta_\star) \leq \mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star))
\end{aligned}$$

and therefore for any $\mathcal{X}_-$:

$$\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star) - \mu(x^\mathsf{T}\theta_\star) \geq \mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)/2 \tag{23}$$

Note the difference with case 1. since here the gap is no longer lower-bounded by a constant (*i.e* it is problem-dependent). Fix $x \in \mathcal{X}_-$ and let $\tau = \tau_x$ (*i.e* $x_\tau = x$). Using the mean-value theorem we obtain:

$$\begin{aligned}
\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)/2 &\leq \mu\left(x_\star(\theta_\star)^\mathsf{T}\theta_\star\right) - \mu\left(x_\tau^\mathsf{T}\theta_\star\right) \\
&\leq \alpha(\theta_\star, x_\star(\theta_\star), x_\tau)\theta_\star^\mathsf{T}(x_\star(\theta_\star) - x_\tau) \\
&\leq \alpha(\theta_\star, x_\star(\theta_\star), x_\tau) x_\tau^\mathsf{T}(\theta_\tau - \theta_\star) && \text{(optimism)} \\
&\leq \alpha(\theta_\star, x_\star(\theta_\star), x_\tau) x_\tau^\mathsf{T}\mathbf{G}_\tau^{-\mathbf{1}}(\theta_\tau, \theta_\star)(g_\tau(\theta_\tau) - g_\tau(\theta_\star)) && \text{(Equation (6))} \\
&\leq \alpha(\theta_\star, x_\star(\theta_\star), x_\tau) \|x_\tau\|_{\mathbf{G}_\tau^{-\mathbf{1}}(\theta_\tau,\theta_\star)} \|g_\tau(\theta_\tau) - g_\tau(\theta_\star)\|_{\mathbf{G}_\tau^{-\mathbf{1}}(\theta_\tau,\theta_\star)} && \text{(Cauchy-Schwarz)} \\
&\leq 2\sqrt{1+2S}\gamma_\tau(\delta)\alpha(\theta_\star, x_\star(\theta_\star), x_\tau) \|x_\tau\|_{\mathbf{G}_\tau^{-\mathbf{1}}(\theta_\tau,\theta_\star)} \\
&\leq 2\sqrt{1+2S}\gamma_\tau(\delta)\dot\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star) \|x_\tau\|_{\mathbf{G}_\tau^{-\mathbf{1}}(\theta_\tau,\theta_\star)} && (24)
\end{aligned}$$

where we used $\|g_\tau(\theta_t) - g_\tau(\theta_\star)\|_{\mathbf{G}_\tau^{-\mathbf{1}}(\theta_\tau,\theta_\star)} \leq 2\sqrt{1+2S}\gamma_\tau(\delta)$ (cf. proof of Proposition 4) and the fact that $\dot\mu$ is increasing on $[x_\tau^\mathsf{T}\theta_\star, x_\star(\theta_\star)^\mathsf{T}\theta_\star]$ which yields $\alpha(\theta_\star, x_\star(\theta_\star), x_\tau) \leq \dot\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)$. We now need to separate two cases:

2.1. $x^\mathsf{T}\theta_\tau \leq 0$. Thanks to optimism (*i.e* $x^\mathsf{T}\theta_\tau \geq x_\star(\theta_\star)^\mathsf{T}\theta_\star$) and the monotonicity (increasing) of $\dot\mu$ in $\mathbb{R}^-$ we obtain that $\dot\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star) \leq \dot\mu(x^\mathsf{T}\theta_\tau)$. Further:

$$\begin{aligned}
\|x_\tau\|_{\mathbf{G}_\tau^{-\mathbf{1}}(\theta_\tau,\theta_\star)} &\leq \sqrt{1+2S}\,\|x_\tau\|_{\mathbf{H}_\tau^{-\mathbf{1}}(\theta_\tau)} && \text{(Equation (7))} \\
&\leq \sqrt{1+2S}(N_x \dot\mu(x^\mathsf{T}\theta_\tau))^{-1/2} && \text{(Sherman-Morison)} \\
&\leq \sqrt{1+2S}(N_x \dot\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star))^{-1/2} && (25)
\end{aligned}$$

2.2. $x^\mathsf{T}\theta_\tau \geq 0$.

$$\begin{aligned}
\|x_\tau\|_{\mathbf{G}_\tau^{-1}(\theta_\tau,\theta_\star)} &\leq (N_x \alpha(x,\theta_\tau,\theta_\star))^{-1/2} && \text{(Sherman-Morison)}\\
&\leq N_x^{-1/2}\left(\frac{x^\mathsf{T}\theta_\tau - x^\mathsf{T}\theta_\star}{\mu(x^\mathsf{T}\theta_\tau)-\mu(x^\mathsf{T}\theta_*)}\right)^{1/2} && \text{(mean-value theorem)}\\
&\leq N_x^{-1/2}\sqrt{2S}\left(\mu(x^\mathsf{T}\theta_\tau)-\mu(x^\mathsf{T}\theta_*)\right)^{-1/2} && (\|x\|\leq 1, \theta_\tau,\theta_\star\in\Theta)\\
&\leq N_x^{-1/2}\sqrt{2S}\left(1/2-\mu(x^\mathsf{T}\theta_*)\right)^{-1/2} && (x^\mathsf{T}\theta_\tau \geq 0 \Rightarrow \mu(x^\mathsf{T}\theta_\tau)\geq 1/2)\\
&\leq N_x^{-1/2}\sqrt{2S}\left(1/2-\mu(-1)\right)^{-1/2} && (x^\mathsf{T}\theta_\star \leq 0 \Rightarrow \mu(x^\mathsf{T}\theta_\star)\leq \mu(-1))\\
&\leq 5N_x^{-1/2}\sqrt{2S}\\
&\leq 5(N_x\dot\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star))^{-1/2}\sqrt{2S} && (0\leq\dot\mu\leq 1) \qquad (26)
\end{aligned}$$

Therefore combining Equations (25) and (26) we obtain that whichever we are in case 2.1 or 2.2, for any $x\in\mathcal{X}_-$:

$$\|x_\tau\|_{\mathbf{G}_\tau^{-1}(\theta_\tau,\theta_\star)} \leq C_3\left(N_x\dot\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)\right)^{-1/2}\gamma_\tau(\delta)$$

where C_3 is a constant hiding universal terms and S dependencies. Plugging this result in Equation (24) and introducing a similar constant C_4 we obtain that:

$$\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)/2 \leq C_4 N_x^{-1/2}\left(\dot\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)\right)^{1/2}\gamma_\tau(\delta)$$

Therefore for any $x\in\mathcal{X}_-$:

$$\begin{aligned}
N_x &\leq 4C_4^2\frac{\dot\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)}{\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)^2}\gamma_\tau(\delta)^2\\
&\leq \frac{4C_4^2}{\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)}\gamma_\tau(\delta)^2 && (\dot\mu\leq\mu) \qquad (27)
\end{aligned}$$

Henceforth from Equation (20):

$$\begin{aligned}
R^{\text{trans}}_{\theta_\star}(T) &\leq C_1 d^2\log^2(T) + C_2\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)\sum_{t=1}^T \mathbb{1}\left(x_t\in\mathcal{X}_-\right)\\
&= C_1 d^2\log^2(T) + C_2\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)\sum_{t=1}^T\sum_{x\in\mathcal{X}_-}\mathbb{1}(x_t=x)\\
&= C_1 d^2\log^2(T) + C_2\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)\sum_{x\in\mathcal{X}_-}N_x\\
&\leq C_1 d^2\log^2(T) + 4C_2C_4\sum_{x\in\mathcal{X}_-}\gamma_{\tau_x}(\delta)^2 && \text{(Equation (27))}\\
&\leq C_1 d^2\log^2(T) + 25C_2K\max_{t\in[T]}\gamma_t(\delta)^2 && (|\mathcal{X}_-|=K)
\end{aligned}$$

Using the fact that $\max_{t\in[T]}\gamma_t(\delta)\lesssim_T \sqrt{d\log(T)}$ we obtain the announced result:

$$R^{\text{trans}}_{\theta_\star}(T)\lesssim_T d^2+dK$$

Case 3. $x_\star(\theta_\star)^\mathsf{T}\theta_\star\in[-1,0]$. Recall the definition of $\mathcal{X}_-$ in this case:

$$\mathcal{X}_- = \left\{x\in\mathcal{X}\,\middle|\,\dot\mu(x^\mathsf{T}\theta_\star)\leq\dot\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)/2\right\}$$

We can directly re-use the characterization of the sub-optimality gap for detrimental arms of Equation (23). This yields that for any $x\in\mathcal{X}_-$:

$$\begin{aligned}
\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)-\mu(x^\mathsf{T}\theta_\star) &\geq \dot\mu(x_\star(\theta_\star)^\mathsf{T}\theta_\star)/2\\
&\geq \dot\mu(-1)/2 \geq 9/200
\end{aligned}$$

We are therefore in the same configuration as in case 1 (the sub-optimality gap of detrimental arms is lower-bounded by a non-problem dependent constant). Following the same reasoning yields to the announced claim. This finishes the proof.

□

C.3.2 Proof of Equation (3)

Proof. As in the proof of Equation (2), we work under the assumption that the event $E_\delta = \{\forall t \geq 1, \theta_\star \in \mathcal{C}_t(\delta)\}$ holds, which happens with high probability (cf. Proposition 1). We focus here on the case where $\mathcal{X} = \mathcal{B}_d(0,1)$, which implies that any parameter θ is co-linear with its associated optimal arm. More precisely: $x_\star(\theta) = \theta/\|\theta\|$ for any $\theta \in \Theta$. Further, this guarantees that $x_\star(\theta)^\mathsf{T}\theta \geq 0$ for all $\theta \in \Theta$. In particular, $x_\star(\theta_\star)^\mathsf{T}\theta_\star \geq 0$ and we have the following definition for the detrimental arms:

$$\mathcal{X}_- = \left\{x \in \mathcal{X} \middle| x^\mathsf{T}\theta_\star \leq -1\right\}.$$

The objective of the proof is to bound the number of time detrimental arms are played by **OFULog** within T rounds. We collect this in the following set:

$$\mathcal{T} := \{t \leq T \text{ s.t } x_t \in \mathcal{X}_-\}, \ . \tag{28}$$

To do so, we start by decomposing 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

$$\begin{aligned}\tau_i &= \max\left\{t \in \mathcal{T}, t \notin \bigcup_{j<i}\mathcal{T}_j\right\},\\ \mathcal{T}_i &= \left\{t \leq \tau_i, t \notin \bigcup_{j<i}\mathcal{T}_j \geq 0,\, x_t^\mathsf{T}\theta_{\tau_i},\, x_t \in \mathcal{X}_-\right\}.\end{aligned} \tag{29}$$

Such construction immediately implies that $\{\mathcal{T}_i\}_{i\geq 1}$ is a partition of $\mathcal{T}$.

Proposition 5. *Let $\mathcal{T}$ and $\{\mathcal{T}_i\}_{i\geq 1}$ be defined as in Equation (28) and Equation (29), and let N be the number of subsets $\{\mathcal{T}_i\}$. Then:*

$$\bigcup_{i=1}^{N}\mathcal{T}_i = \mathcal{T}; \qquad \mathcal{T}_i \cap \mathcal{T}_j = \emptyset, \forall i \neq j; \qquad N \leq (d+1).$$

Proof of Proposition 5. 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 τ_i for $i = 1, \dots, N$, we have that

$$\forall j > i, \quad x_{\tau_i}^\mathsf{T}\theta_{\tau_j} < 0,$$

and since θ_{τ_i} is co-linear with x_{τ_i}, we obtain

$$\forall j, i \in [N], \quad x_{\tau_i}^\mathsf{T}x_{\tau_j} < 0.$$

We conclude by using Lemma. 19 in Dong et al. (2019), which states that it can only exists 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 5, 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.

First, notice that τ_i is the last time step in $\mathcal{T}_i$ and that for all $t \in \mathcal{T}_i$, $x_t^\mathsf{T}\theta_* \leq -1$ (from the definition of $\mathcal{T}$) while $x_t^\mathsf{T}\theta_{\tau_i} \geq 0$ (from the construction of the partition). Hence, for all $t \in \mathcal{T}_i$:

$$\begin{aligned}
\mu(0) - \mu(-1) &\leq \mu(x_t^\mathsf{T}\theta_{\tau_i}) - \mu(x_t^\mathsf{T}\theta_\star) \\
&= \alpha(x_t, \theta_{\tau_i}, \theta_\star) x_t^\mathsf{T}(\theta_{\tau_i} - \theta_\star) && \text{(mean-value theorem)} \\
&\leq \alpha(x_t, \theta_{\tau_i}, \theta_\star) \|x_t\|_{\mathbf{G}_{\tau_i}^{-1}(\theta_{\tau_i}, \theta_\star)} \|\theta_{\tau_i} - \theta_\star\|_{\mathbf{G}_{\tau_i}(\theta_{\tau_i}, \theta_\star)} && \text{(Cauchy-Schwarz)} \\
&\leq 2\sqrt{1+2S}\gamma_{\tau_i}(\delta)\alpha(x_t, \theta_{\tau_i}, \theta_\star) \|x_t\|_{\mathbf{G}_{\tau_i}^{-1}(\theta_{\tau_i}, \theta_\star)} && (E_\delta \text{ holds, Proposition 4}) \\
&\leq \frac{\sqrt{1+2S}}{2}\gamma_{\tau_i}(\delta) \|x_t\|_{\mathbf{G}_{\tau_i}^{-1}(\theta_{\tau_i}, \theta_\star)} \,. && (\alpha \leq \sup \dot{\mu} \leq 1/4) \qquad (30)
\end{aligned}$$

Further, for all $t \in \mathcal{T}_i$, $x_t^\mathsf{T}\theta_\star \leq -1$ and $x_t^\mathsf{T}\theta_{\tau_i} \geq 0$ leads to

$$\begin{aligned}
\alpha(x_t, \theta_{\tau_i}, \theta_\star) &= \frac{\mu(x_t^\mathsf{T}\theta_{\tau_i}) - \mu(x_t^\mathsf{T}\theta_\star)}{x_t^\mathsf{T}(\theta_{\tau_i} - \theta_\star)} \\
&\geq \frac{\mu(0) - \mu(-1)}{2S} \,. && (\|x\| \leq 1,\, \theta_{\tau_i}, \theta_\star \in \Theta)
\end{aligned}$$

As a result, let $\bar{\mathbf{V}}_{\tau_\mathbf{i}} := \sum_{s \in \mathcal{T}_i} x_s x_s^\mathsf{T} + \lambda_{\tau_i}\mathbf{I_d}$, one obtains,

$$\mathbf{G}_{\tau_\mathbf{i}}(\theta_{\tau_i}, \theta_\star) \succcurlyeq \sum_{s \in \mathcal{T}_i} \alpha(x_s, \theta_{\tau_i}, \theta_\star) x_s x_s^\mathsf{T} + \lambda_{\tau_i}\mathbf{I_d} \succeq \frac{\mu(0) - \mu(-1)}{2S}\bar{\mathbf{V}}_{\tau_\mathbf{i}},$$

which combined with Equation (30) leads to:

$$(\mu(0) - \mu(-1))^{3/2} \leq \sqrt{S/2}\sqrt{1+2S}\gamma_{\tau_i}(\delta)\|x_t\|_{\bar{\mathbf{V}}_{\tau_\mathbf{i}}^{-1}}. \qquad (31)$$

Taking the square and summing over $t \in \mathcal{T}_i$ yields:

$$\begin{aligned}
(\mu(0) - \mu(-1))^3 |\mathcal{T}_i| &\leq (S/2)(1+2S)\gamma_{\tau_i}(\delta)^2 \sum_{t \in \mathcal{T}_i} \|x_t\|^2_{\bar{\mathbf{V}}_{\tau_\mathbf{i}}^{-1}} \\
&\leq (S/2)(1+2S)\gamma_{\tau_i}(\delta)^2 \mathrm{Tr}\left(\bar{\mathbf{V}}_{\tau_\mathbf{i}}^{-1} \sum_{t \in \mathcal{T}_i} x_t x_t^\mathsf{T}\right) \\
&\leq (S/2)(1+2S)\gamma_{\tau_i}(\delta)^2 d
\end{aligned}$$

and therefore $|\mathcal{T}_i| \leq C_5 d\gamma_{\tau_i}^2(\delta)$. Since $\max_{t \in [T]} \gamma_t(\delta) \lesssim_T \sqrt{d\log(T)}$ we obtain

$$|\mathcal{T}| = \sum_{i=1}^N |\mathcal{T}_i| \leq C_5(d+1)d \max_{i=1,\dots,N} \gamma_{\tau_i}^2(\delta) \leq C_6 d^3 \log(T).$$

which we plug in Equation (20) to obtain the desired result,

$$R_{\theta_\star}^{\text{trans}}(T) \leq C_1 d^2 \log^2(T) + C_6 d^3 \log(T).$$

Here C_5 and C_6 are universal constants hiding dependencies in poly(S).

□

C.4 Proof of Theorem 3

Theorem 3 (Unit-Ball Regret Upper-Bound)**.** *If $\mathcal{X} = \mathcal{B}_d(0,1)$ the regret of* **OFULog** *satisfies:*

$$\text{Regret}_{\theta_\star}(T) \lesssim_T d\sqrt{\frac{T}{\kappa_\mathcal{X}}} + d^2 \quad w.h.p\,.$$

Proof. The result is easily obtained by merging Theorem 1 with Equation (3) in Proposition 2. □

D REGRET LOWER-BOUND

We give below a statement of Theorem 2 which is more detailed than its version in the main text. In particular we emphasize the fact that ϵ_T is *small* enough that $\theta_\star$ and all the alternative packing $\{\|\theta - \theta_\star\| \leq \epsilon_T\}$ have roughly the same problem-dependent constants (cf **2.** in Theorem 2).

Theorem 2 (Local Lower-Bound)**.** *Let $\mathcal{X} = \mathcal{S}_d(0,1)$. For any problem instance $\theta_\star$ and for $T \geq d^2\kappa_\star(\theta_\star)$, there exist ϵ_T small enough such that:*

1. $\text{MinimaxRegret}_{\theta_\star,T}(\epsilon_T) = \Omega\left(d\sqrt{\frac{T}{\kappa_\star(\theta_\star)}}\right)$

2. $\frac{5}{6}\kappa_\star(\theta_\star) \leq \kappa_\star(\theta) \leq \frac{6}{5}\kappa_\star(\theta_\star) \qquad \forall \theta \in \left\{\|\theta - \theta_\star\| \leq \epsilon_T\right\}$

D.1 Proof of Theorem 2

The strategy for proving this result is the following: for any policy π,[5] we will assume that for a well-chosen set Ξ we have:

$$\forall \theta \in \Xi, \quad \text{Regret}_\theta^\pi(T) = \mathcal{O}\left(d\sqrt{\frac{T}{\kappa_\star(\theta)}}\right) .$$

We shall arrive to a contradiction of the form:

$$\exists \theta \in \Xi \quad \text{s.t} \quad \text{Regret}_\theta^\pi(T) = \Omega\left(d\sqrt{\frac{T}{\kappa_\star(\theta)}}\right) .$$

Proof. In the following, we fix the policy π. 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 θ. We refer the interested reader to (Lattimore and Szepesvári, 2020, Section 4.7) for a thorough definition of this probability space. To simplify notations, we will denote $\mathbb{P}_\theta = \mathbb{P}_{\pi\theta}$ the probability measure of the random sequence $\{x_1, r_2, .., x_T, r_{T+1}\}$, obtained by having π interact with the environment parameter θ. Recall that we work in a logistic bandit setting, meaning that at any round t:

$$\mathbb{P}_\theta(r_t|x_t) = \text{Bernoulli}(\mu(x_t^\mathsf{T}\theta))$$

where $\mu(z) = (1+\exp(-z))^{-1}$ is the logistic function. Note that when $\mathcal{X} = \mathcal{S}_d(0,1)$ we have $\kappa_\star(\theta) = \kappa_\mathcal{X}(\theta)$ for any θ. We therefore use the notation $\kappa(\theta)$ for short. We will need the following result, of which we defer the proof to Appendix D.3.

Proposition 6. *For all $\theta \in \mathbb{R}^d$ the following holds:*

$$\text{Regret}_\theta(T) \geq \frac{\|\theta\|}{\kappa(\theta)} \sum_{i=1}^d \mathbb{E}_\theta\left[\sum_{t=1}^T [x(\theta) - x_t]_i^2\right] \tag{32}$$

Further if $\|\theta\| \geq 1$:

$$\text{Regret}_\theta(T) \geq \frac{1}{6}\mathbb{E}_\theta\left[\sum_{t=1}^T \dot{\mu}(x_t^\mathsf{T}\theta)\,\|x(\theta) - x_t\|^2\right] \tag{33}$$

In this proof we will assume that $\|\theta_\star\| \geq 1$ (which implies that $\kappa(\theta_\star) \geq 5$).[6] Let $\{e_i\}_{i=1}^d$ the canonical basis of $\mathbb{R}^d$ and without loss of generality assume that $\theta_\star = \|\theta_*\|\, e_1$. With such notations, we now introduce the set of *unidentifiable* parameters:

$$\Xi := \left\{\theta_* + \epsilon \sum_{2=1}^d v_i e_i\,, \quad v \in \{-1,1\}^d\right\}$$

[5] The policy is arbitrary, we only ask that at round t its actions are $\mathcal{F}_t$-adapted.

[6] This assumption 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 κ.

where ϵ is a (small) positive scalar to be tuned later. For now, we will only make the following assumption on ϵ:

$$\varepsilon \leq \|\theta_\star\| / \sqrt{d-1} \tag{34}$$

Intuitively, Ξ is a set of slightly perturbed versions of $\theta_\star$. The goal is to set ϵ small enough so the parameters are indiscernible for a policy interacting with each of them, however large enough so the policy can't perform well on all problems. Note that all the elements θ of Ξ have the same norm, and henceforth the same $\kappa(\theta) =: \kappa_\epsilon$. As anticipated earlier, we are going to make the hypothesis that for all $\theta \in \Xi$, the regret is dominated by $d\sqrt{T/\kappa_\epsilon}$. Note that if this assumption does not hold, then by definition there exists $\theta \in \Xi$ such that $\mathrm{Regret}_\theta(T) = \Omega(d\sqrt{T/\kappa_\epsilon})$ and the proof is over.

Hypothesis. *There exists a universal constant C such that:*

$$\forall \theta \in \Xi, \quad R_\theta(T) \leq Cd\sqrt{\frac{T}{\kappa_\epsilon}} \tag{H1}$$

Without loss of generality, we will take $C = 1$[7].

Starting from Equation (33) we are going to provide a first lower-bound of the regret for any $\theta \in \Xi$. To do so, introduce for any direction $i \in [d, 2]$ the event:

$$A_i(\theta) := \left\{ [x_\star(\theta) - x_\star(\theta_\star)]_i \cdot \left[\frac{1}{T} \sum_{t=1}^{T} x_t - x_\star(\theta_\star) \right]_i \geq 0 \right\}$$

We have the following lower-bound, which proof is deferred to Appendix D.4.

Lemma 3. *For any $\theta \in \Xi$ we have:*

$$\mathrm{Regret}_\theta(T) \geq \frac{T\epsilon^2}{2\kappa_\epsilon \|\theta_\star\|} \sum_{i=2}^{d} \mathbb{P}_\theta(A_i(\theta))$$

The goal is now to find one $\theta \in \Xi$ 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 $\mathrm{Flip}_i(\cdot)$ which for any $\theta \in \Xi$ changes the sign of the i^{th} coordinate of θ. Formally, let:

$$[\mathrm{Flip}_i(\theta)]_i = -[\theta]_i \quad \text{and} \quad [\mathrm{Flip}_i(\theta)]_j = [\theta]_j \ \text{ for all } \ j \neq i \tag{35}$$

In the following Lemma, we show that the average value of $\sum_{i=2}^{d} \mathbb{P}_\theta(A_i(\theta))$ over Ξ is linked to the average relative entropy (denoted D_{KL}) between *flipped* versions of θ.

Lemma 4 (Averaging Hammer)**.** *The following holds:*

$$\frac{1}{|\Xi|} \sum_{\theta \in \Xi} \sum_{i=2}^{d} \mathbb{P}_\theta(A_i(\theta)) \geq \frac{d}{4} - \frac{\sqrt{d}}{2} \sqrt{\frac{1}{|\Xi|} \sum_{\theta \in \Xi} \sum_{i=2}^{d} D_{KL}\left(\mathbb{P}_\theta, \mathbb{P}_{\mathrm{Flip}_i(\theta)}\right)}$$

The proof is deferred to Appendix D.5. We now have to characterize this average relative entropy. This is done in the following Lemma, which proof is presented in Appendix D.6.

Lemma 5 (Average Relative Entropy)**.** *Under Hypothesis* (H1) *we have:*

$$\frac{1}{|\Xi|} \sum_{\theta \in \Xi} \sum_{i=2}^{d} D_{KL}\left(\mathbb{P}_\theta, \mathbb{P}_{\mathrm{Flip}_i(\theta)}\right) \leq \frac{2}{\kappa_\epsilon} dT\epsilon^4 \exp(4\epsilon) + 4d\epsilon^2 \exp(4\epsilon)(6 + \frac{d}{2}\epsilon^2)\sqrt{\frac{T}{\kappa_\epsilon}}$$

[7] This assumption is made to avoid clutter and is not necessary. Keeping C only impacts our lower-bound by a universal constant, independent of the problem.

Combining Lemmas 4 and 5 we therefore obtain that:

$$\frac{1}{|\Xi|}\sum_{\theta\in\Xi}\sum_{i=2}^{d}\mathbb{P}_\theta(A_i(\theta)) \geq \frac{d}{4}\left[1-2\left(2\epsilon^4\frac{T}{\kappa_\epsilon}+24\epsilon^2\sqrt{\frac{T}{\kappa_\epsilon}}+2d\epsilon^4\sqrt{\frac{T}{\kappa_\epsilon}}\right)^{1/2}\exp(2\epsilon)\right]$$

Because this results holds for an average over Ξ, it must still be true for at least one $\tilde{\theta}\in\Xi$. In other words, there exists $\tilde{\theta}\in\Xi$ 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 we therefore have that it exists $\tilde{\theta}\in\Xi$ such that:

$$\mathrm{Regret}_{\tilde{\theta}}(T) \geq dT\frac{\epsilon^2}{8\left\|\theta_\star\right\|\kappa_\epsilon}\left[1-2\left(2\epsilon^4\frac{T}{\kappa_\epsilon}+24\epsilon^2\sqrt{\frac{T}{\kappa_\epsilon}}+2d\epsilon^4\sqrt{\frac{T}{\kappa_\epsilon}}\right)^{1/2}\exp(2\epsilon)\right]$$

We only have left to tune ϵ to prove our result. Taking $\epsilon^2=\frac{1}{32}\sqrt{\frac{\kappa_\epsilon}{T}}$ yields, after some computations that:

$$\mathrm{Regret}_{\tilde{\theta}}(T) \geq \frac{1}{256\left\|\theta_\star\right\|}d\sqrt{\frac{T}{\kappa_\epsilon}}\left(1-2\left(\frac{24576}{32^4}+\frac{2}{32^4}d\sqrt{\frac{\kappa_\epsilon}{T}}\right)^{1/2}\exp\left(\frac{2}{\sqrt{32}}\sqrt{\frac{\kappa_\epsilon}{T}}\right)\right)$$

When $T\geq d^2\kappa$ (and therefore $T\geq\kappa$) we obtain:

$$\begin{aligned}\mathrm{Regret}_{\tilde{\theta}}(T) &\geq \frac{1}{256\left\|\theta_\star\right\|}d\sqrt{\frac{T}{\kappa_\epsilon}}\left(1-\left(\frac{98312}{32^4}\right)^{1/2}\exp\left(\frac{1}{\sqrt{8}}\right)\right)\\ &\geq \frac{1}{512\left\|\theta_\star\right\|}d\sqrt{\frac{T}{\kappa_\epsilon}}\end{aligned}$$

To sum-up, we have shown that when Hypothesis (**H1**) holds, there exists $\tilde{\theta}\in\Xi$ such that $\mathrm{Regret}_{\tilde{\theta}}(T)=\Omega(d\sqrt{\frac{T}{\kappa_\epsilon}})$. Note that if Hypothesis (**H1**) did not hold, then by definition such a parameter would also exist. This proves part **1.** of the claim; indeed by setting $\epsilon_T^2=\frac{1}{32}\sqrt{\kappa_\epsilon/T}$ we have shown that for *any* policy π if $T\geq d^2\kappa_\epsilon$:

$$\max_{\|\theta-\theta_\star\|^2\leq d\epsilon_T^2}\mathrm{Regret}_\theta^\pi(T)=\Omega\left(d\sqrt{\frac{T}{\kappa_\epsilon}}\right)$$

and therefore since $\kappa_\epsilon\geq\kappa_\star(\theta_\star)$ there exists $\tilde{\epsilon}_T$ small enough ($\tilde{\epsilon}_T=\sqrt{d}\epsilon_T$) such that:

$$\mathrm{MinimaxRegret}_{\theta_\star,T}(\tilde{\epsilon}_T)=\Omega\left(d\sqrt{\frac{T}{\kappa_\star(\theta_\star)}}\right)$$

This formulation is somehow a degradation of the result we obtained, because we showed that under $\tilde{\theta}$ (the hard nearby instance) the regret is $\Omega(d\sqrt{T/\kappa_\epsilon})$ and therefore *directly involves* the problem-dependent constant $\kappa(\tilde{\theta})=\kappa_\epsilon$. This degradation is however mild: our bound is *local* and ϵ_T is *small*. As a result, $\theta_\star$ and any nearby alternative $\theta\in\Xi$ fundamentally have the same problem-dependent constants. We now turn this intuition rigorous and prove part **2.** of the Theorem. By Lemma 9 for any θ:

$$\dot{\mu}\left(x_\star(\theta)^\mathsf{T}\theta\right)\exp\left(-\left|x_\star(\theta_\star)^\mathsf{T}\theta_\star-x_\star(\theta)^\mathsf{T}\theta\right|\right)\leq\dot{\mu}\left(x_\star(\theta_\star)^\mathsf{T}\theta_\star\right)\leq\dot{\mu}\left(x_\star(\theta)^\mathsf{T}\theta\right)\exp\left(\left|x_\star(\theta_\star)^\mathsf{T}\theta_\star-x_\star(\theta)^\mathsf{T}\theta\right|\right)$$

which yields that if $\left\|\theta-\theta_\star\right\|^2\leq d\epsilon_T^2$:

$$\dot{\mu}\left(x_\star(\theta)^\mathsf{T}\theta\right)\exp\left(-\sqrt{d}\epsilon_T\right)\leq\dot{\mu}\left(x_\star(\theta_\star)^\mathsf{T}\theta_\star\right)\leq\dot{\mu}\left(x_\star(\theta)^\mathsf{T}\theta\right)\exp\left(\sqrt{d}\epsilon_T\right)$$

We obtain the desired result by noting that $d\epsilon_T^2=(1/32)d\sqrt{\kappa(\theta)/T}\leq 1/32$ when $T\geq d^2\kappa(\theta)$: □

D.2 A Global Lower-Bound

As announced in the main text, this local-minimax bound easily implies a global one. We state it here for the sake of completeness.

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

$$\mathrm{Regret}_\theta^\pi(T) = \Omega\left(d\sqrt{\frac{T}{\kappa}}\right)$$

Proof. This result is a direct consequence of Theorem 2. The proof only requires to select a nominal instance $\theta_\star$ which ℓ_2-norm is large enough so that for any $\theta \in \Xi$ we have $\kappa_\star(\theta) = \kappa$. □

D.3 Proof of Proposition 6

Proposition 6. *For all $\theta \in \mathbb{R}^d$ the following holds:*

$$\mathrm{Regret}_\theta(T) \geq \frac{\|\theta\|}{\kappa(\theta)} \sum_{i=1}^{d} \mathbb{E}_\theta\left[\sum_{t=1}^{T} [x(\theta) - x_t]_i^2\right] \tag{32}$$

Further if $\|\theta\| \geq 1$:

$$\mathrm{Regret}_\theta(T) \geq \frac{1}{6}\mathbb{E}_\theta\left[\sum_{t=1}^{T} \dot{\mu}(x_t^\mathsf{T}\theta)\, \|x(\theta) - x_t\|^2\right] \tag{33}$$

Proof. We start by proving the second result. By definition of the regret:

$$\begin{aligned}
\mathrm{Regret}_\theta(T) &= \mathbb{E}_\theta\left[\sum_{t=1}^{T} \mu(x_\star(\theta)^\mathsf{T}\theta) - \mu(x_t^\mathsf{T}\theta)\right] \\
&= \mathbb{E}_\theta\left[\sum_{t=1}^{T} \alpha\left(\theta, x_\star(\theta), x_t\right)\left(x_\star(\theta)^\mathsf{T}\theta - x_t^\mathsf{T}\theta\right)\right] && \text{(mean-value theorem)} \\
&\geq \mathbb{E}_\theta\left[\sum_{t=1}^{T} \frac{\dot{\mu}(x_t^\mathsf{T}\theta)}{1 + |\theta^\mathsf{T}(x_\star(\theta) - x_t)|}\left(x_\star(\theta)^\mathsf{T}\theta - x_t^\mathsf{T}\theta\right)\right] && \text{(Lemma 7)} \\
&\geq \frac{1}{1 + 2\|\theta\|}\mathbb{E}_\theta\left[\sum_{t=1}^{T} \dot{\mu}(x_t^\mathsf{T}\theta)\left(x_\star(\theta)^\mathsf{T}\theta - x_t^\mathsf{T}\theta\right)\right] && (\|x\| \leq 1\, \forall x \in \mathcal{X}) \\
&\geq \frac{\|\theta\|}{1 + 2\|\theta\|}\mathbb{E}_\theta\left[\sum_{t=1}^{T} \dot{\mu}(x_t^\mathsf{T}\theta)\left(1 - x_t^\mathsf{T}\frac{\theta}{\|\theta\|}\right)\right] \\
&\geq \frac{\|\theta\|}{2 + 4\|\theta\|}\mathbb{E}_\theta\left[\sum_{t=1}^{T} \dot{\mu}(x_t^\mathsf{T}\theta)\, \|x_\star(\theta) - x_t\|^2\right]
\end{aligned}$$

where in the last line we used that for all $x, y \in \mathcal{S}_d(0,1)$ we have $1 - x^\mathsf{T}y = \frac{1}{2}\|x - y\|^2$. Using the fact that $\|\theta\| \geq 1$ yields the second result.

A similar bound can be written by using $\alpha(\theta, x_\star(\theta), x_t) \geq \dot{\mu}(x_\star(\theta)^\mathsf{T}\theta)$. Namely, we obtain:

$$\begin{aligned}
\text{Regret}_\theta(T) &\geq \mathbb{E}_\theta\left[\sum_{t=1}^T \dot{\mu}(x_\star(\theta_\star)^\mathsf{T}\theta)\left(x_\star(\theta)^\mathsf{T}\theta - x_t^\mathsf{T}\theta\right)\right] \\
&\geq \frac{\|\theta\|}{\kappa_\star(\theta)}\mathbb{E}_\theta\left[\sum_{t=1}^T \|x_\star(\theta) - x_t\|^2\right] \\
&\geq \frac{\|\theta\|}{\kappa_\star(\theta)}\mathbb{E}_\theta\left[\sum_{t=1}^T \|x_\star(\theta) - x_t\|^2\right] && (\|\theta_\star\| \geq 1) \\
&\geq \frac{\|\theta\|}{\kappa_\star(\theta)}\mathbb{E}_\theta\left[\sum_{t=1}^T \sum_{i=1}^d [x_\star(\theta) - x_t]_i^2\right]
\end{aligned}$$

Using the linearity of the expectation delivers the first claim. □

D.4 Proof of Lemma 3

Lemma 3. *For any $\theta \in \Xi$ we have:*

$$\text{Regret}_\theta(T) \geq \frac{T\epsilon^2}{2\kappa_\epsilon \|\theta_\star\|}\sum_{i=2}^d \mathbb{P}_\theta(A_i(\theta))$$

Proof. From Proposition 6 we have that:

$$\begin{aligned}
\text{Regret}_\theta(T) &\geq \frac{\|\theta\|}{\kappa(\theta)}\sum_{i=1}^d \mathbb{E}_\theta\left[\sum_{t=1}^T [x_\star(\theta) - x_t]_i^2\right] \\
&\geq \frac{\|\theta\|}{\kappa(\theta)}\sum_{i=1}^d \mathbb{E}_\theta\left[\sum_{t=1}^T [x_\star(\theta) - x_t]_i^2 \mathbb{1}\{A_i(\theta)\}\right] \\
&= \frac{\|\theta\|}{\kappa(\theta)}\sum_{i=1}^d \mathbb{E}_\theta\left[\sum_{t=1}^T [x_\star(\theta) - x_\star(\theta_\star) + x_\star(\theta_\star) - x_t]_i^2 \mathbb{1}\{A_i(\theta)\}\right] \\
&= \frac{\|\theta\|}{\kappa(\theta)}\sum_{i=1}^d [x_\star(\theta) - x_\star(\theta_\star)]_i^2 \mathbb{E}_\theta\left[\mathbb{1}\{A_i(\theta)\}\right] \\
&\quad + \frac{\|\theta\|}{\kappa(\theta)}\sum_{i=1}^d \mathbb{E}_\theta\left[\sum_{t=1}^T [x_\star(\theta_\star) - x_t]_i^2 \mathbb{1}\{A_i(\theta)\}\right] \\
&\quad + \frac{2T\|\theta\|}{\kappa(\theta)}\sum_{i=1}^d \mathbb{E}_\theta\left[\mathbb{1}\{A_i(\theta)\}\left[x_\star(\theta_\star) - \frac{1}{T}\sum_{t=1}^T x_t\right]_i [x_\star(\theta) - x_\star(\theta_\star)]_i\right] \\
&\geq \frac{\|\theta\|}{\kappa(\theta)} T \sum_{i=1}^d [x_\star(\theta) - x_\star(\theta_\star)]_i^2 \mathbb{E}_\theta\left[\mathbb{1}\{A_i(\theta)\}\right]
\end{aligned}$$

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:

$$\begin{aligned}
\text{Regret}_\theta(T) &\geq T\frac{\|\theta\|}{\kappa_\epsilon}\frac{\epsilon^2}{\|\theta_\star\|^2 + (d-1)\epsilon^2}\sum_{i=2}^d \mathbb{E}_\theta\left[\mathbb{1}\{A_i(\theta)\}\right] \\
&\geq T\frac{\|\theta\|}{\kappa_\epsilon}\frac{\epsilon^2}{2\|\theta_\star\|^2}\sum_{i=2}^d \mathbb{E}_\theta\left[\mathbb{1}\{A_i(\theta)\}\right] && \text{(Equation (34))} \\
&= \frac{T\epsilon^2}{2\kappa_\epsilon\|\theta_\star\|}\sum_{i=2}^d \mathbb{P}_\theta(A_i(\theta))
\end{aligned}$$

□

D.5 Proof of Lemma 4

Lemma 4 (Averaging Hammer)**.** *The following holds:*

$$\frac{1}{|\Xi|}\sum_{\theta\in\Xi}\sum_{i=2}^{d}\mathbb{P}_\theta(A_i(\theta)) \geq \frac{d}{4} - \frac{\sqrt{d}}{2}\sqrt{\frac{1}{|\Xi|}\sum_{\theta\in\Xi}\sum_{i=2}^{d} D_{KL}\left(\mathbb{P}_\theta, \mathbb{P}_{\mathrm{Flip}_i(\theta)}\right)}$$

Proof. Let us fix $\theta \in \Theta$ and $i \in [2, d]$. Note that:

$$\begin{aligned}
\mathbb{P}_{\mathrm{Flip}_i(\theta)}(A_i(\mathrm{Flip}_i(\theta)) &\geq \mathbb{P}_\theta(A_i(\mathrm{Flip}_i(\theta))) - D_{\mathrm{TV}}\left(\mathbb{P}_\theta, \mathbb{P}_{\mathrm{Flip}_i(\theta)}\right) \\
&\geq \mathbb{P}_\theta(A_i(\mathrm{Flip}_i(\theta))) - \sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\mathbb{P}_\theta, \mathbb{P}_{\mathrm{Flip}_i(\theta)}\right)} \qquad \text{(Pinsker inequality)} \\
&\geq \mathbb{P}_\theta(A_i^C(\theta)) - \sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\mathbb{P}_\theta, \mathbb{P}_{\mathrm{Flip}_i(\theta)}\right)} \qquad (36)
\end{aligned}$$

where D_{KL} denotes the relative entropy, and where we used the fact that:

$$\begin{aligned}
A_i(\mathrm{Flip}_i(\theta)) &= \left\{ [x_\star(\mathrm{Flip}_i(\theta)) - x_\star(\theta_\star)]_i \cdot \left[\frac{1}{T}\sum_{t=1}^{T} x_t - x_\star(\theta_\star)\right]_i \geq 0 \right\} && \text{(definition)} \\
&= \left\{ [x_\star(\mathrm{Flip}_i(\theta))]_i \cdot \left[\frac{1}{T}\sum_{t=1}^{T} x_t\right]_i \geq 0 \right\} && (x_\star(\theta_\star)_i = 0) \\
&= \left\{ -[x_\star(\theta)]_i \cdot \left[\frac{1}{T}\sum_{t=1}^{T} x_t\right]_i \geq 0 \right\} && ([\mathrm{Flip}_i(\theta)]_i = -[\theta]_i) \\
&= A_i(\theta)^C
\end{aligned}$$

In the following, we denote $\Xi_i^+ := \{\theta \in \Xi \text{ such that } \mathrm{sign}([\theta]_i) > 0\}$ and $\Xi_i^- := \{\theta \in \Xi \text{ such that } \mathrm{sign}([\theta]_i) < 0\}$. Then by averaging over Ξ:

$$\begin{aligned}
\frac{1}{|\Xi|}\sum_{\theta\in\Xi}\sum_{i=2}^{d}\mathbb{P}_\theta(A_i(\theta)) &= \frac{1}{|\Xi|}\sum_{i=2}^{d}\sum_{\theta\in\Xi}\mathbb{P}_\theta(A_i(\theta)) \\
&= \frac{1}{|\Xi|}\sum_{i=2}^{d}\sum_{\theta\in\Xi_i^+}\left(\mathbb{P}_\theta(A_i(\theta)) + \mathbb{P}_{\mathrm{Flip}_i(\theta)}(A_i(\mathrm{Flip}_i(\theta)))\right) \\
&\geq \frac{1}{|\Xi|}\sum_{i=2}^{d}\sum_{\theta\in\Xi_i^+}\mathbb{P}_\theta(A_i(\theta)) + \mathbb{P}_\theta(A_i^C(\theta)) - \sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\mathbb{P}_\theta, \mathbb{P}_{\mathrm{Flip}_i(\theta)}\right)} \qquad \text{(Equation (36))} \\
&\geq \frac{1}{|\Xi|}\sum_{i=2}^{d}\sum_{\theta\in\Xi_i^+} 1 - \sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\mathbb{P}_\theta, \mathbb{P}_{\mathrm{Flip}_i(\theta)}\right)}
\end{aligned}$$

Repeating the same operation but referencing to Ξ_i^- we easily get that:

$$\begin{aligned}
\frac{2}{|\Xi|}\sum_{\theta\in\Xi}\sum_{i=2}^{d}\mathbb{P}_\theta(A_i(\theta)) &\geq \frac{1}{|\Xi|}\sum_{i=2}^{d}\sum_{\theta\in\Xi_i^+\cup\Xi_i^-} 1-\sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\mathbb{P}_\theta,\mathbb{P}_{\mathrm{Flip}_i(\theta)}\right)} \\
&= \frac{1}{|\Xi|}\sum_{i=2}^{d}\sum_{\theta\in\Xi} 1-\sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\mathbb{P}_\theta,\mathbb{P}_{\mathrm{Flip}_i(\theta)}\right)} \\
&= (d-1)-\sum_{i=2}^{d}\frac{1}{|\Xi|}\sum_{\theta\in\Xi}\sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\mathbb{P}_\theta,\mathbb{P}_{\mathrm{Flip}_i(\theta)}\right)} \\
&\geq \frac{d}{2}-\sum_{i=2}^{d}\frac{1}{|\Xi|}\sum_{\theta\in\Xi}\sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\mathbb{P}_\theta,\mathbb{P}_{\mathrm{Flip}_i(\theta)}\right)} && (d\geq 1) \\
&\geq \frac{d}{2}-\frac{1}{\sqrt{2}}\sum_{i=2}^{d}\sqrt{\frac{1}{|\Xi|}\sum_{\theta\in\Xi}D_{\mathrm{KL}}\left(\mathbb{P}_\theta,\mathbb{P}_{\mathrm{Flip}_i(\theta)}\right)} && \text{(Jensen inequality)} \\
&\geq \frac{d}{2}-\sqrt{\frac{d-1}{2}}\sqrt{\sum_{i=2}^{d}\frac{1}{|\Xi|}\sum_{\theta\in\Xi}D_{\mathrm{KL}}\left(\mathbb{P}_\theta,\mathbb{P}_{\mathrm{Flip}_i(\theta)}\right)} && \text{(Cauchy-Schwartz)} \\
&\geq \frac{d}{2}-\sqrt{d}\sqrt{\sum_{i=2}^{d}\frac{1}{|\Xi|}\sum_{\theta\in\Xi}D_{\mathrm{KL}}\left(\mathbb{P}_\theta,\mathbb{P}_{\mathrm{Flip}_i(\theta)}\right)}
\end{aligned}$$

which proves the announced result.

D.6 Proof of Lemma 5

Lemma 5 (Average Relative Entropy). *Under Hypothesis* (**H1**) *we have:*

$$\frac{1}{|\Xi|}\sum_{\theta\in\Xi}\sum_{i=2}^{d}D_{KL}\left(\mathbb{P}_\theta,\mathbb{P}_{\mathrm{Flip}_i(\theta)}\right)\leq\frac{2}{\kappa_\epsilon}dT\epsilon^4\exp(4\epsilon)+4d\epsilon^2\exp(4\epsilon)(6+\frac{d}{2}\epsilon^2)\sqrt{\frac{T}{\kappa_\epsilon}}$$

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. The proof is deferred to Section D.7.

Lemma 6 (Relative Entropy Decomposition). *For any θ,θ' we have that:*

$$D_{\mathrm{KL}}\left(\mathbb{P}_\theta,\mathbb{P}_{\theta'}\right)\leq\mathbb{E}_\theta\left[\sum_{t=1}^{T}\frac{\left(\mu(x_t^\mathsf{T}\theta)-\mu(x_t^\mathsf{T}\theta')\right)^2}{\dot\mu(x_t^\mathsf{T}\theta')}\right]$$

Applying this result between $\mathbb{P}_\theta$ and $\mathbb{P}_{\mathrm{Flip}_i(\theta)}$ yields:

$$\begin{aligned}
D_{\mathrm{KL}}(\mathbb{P}_\theta,\mathbb{P}_{\mathrm{Flip}_i(\theta)}) &\leq \mathbb{E}_\theta\left[\sum_{t=1}^{T}\frac{\left(\mu(x_t^\mathsf{T}\theta)-\mu(x_t^\mathsf{T}\mathrm{Flip}_i(\theta)\right)^2}{\dot\mu(x_t^\mathsf{T}\mathrm{Flip}_i(\theta))}\right] \\
&\leq \mathbb{E}_\theta\left[\sum_{t=1}^{T}\frac{\alpha^2(x_t,\theta,\mathrm{Flip}_i(\theta))}{\dot\mu(x_t^\mathsf{T}\mathrm{Flip}_i(\theta))}\left\{x_t^\mathsf{T}(\theta-\mathrm{Flip}_i(\theta))\right\}^2\right] && \text{(mean-value theorem)}
\end{aligned}$$

We are now going to link $\alpha^2(x_t,\theta,\mathrm{Flip}_i(\theta))$ to $\dot\mu(x_t^\mathsf{T}\mathrm{Flip}_i(\theta))$ and $\dot\mu(x_t^\mathsf{T}\theta)$ thanks to the self-concordance. Indeed, it is easy to show (see the proof of Lemma 7) 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:

$$\begin{aligned}
\alpha(x_t,\theta,\mathrm{Flip}_i(\theta))&\leq\dot\mu(x_t^\mathsf{T}\theta)\exp\left(\left|x_t^\mathsf{T}(\theta-\mathrm{Flip}_i(\theta)\right|\right) \quad \text{and} \\
\alpha^2(x_t,\theta,\mathrm{Flip}_i(\theta))&\leq\dot\mu(x_t^\mathsf{T}\mathrm{Flip}_i(\theta))\exp\left(\left|x_t^\mathsf{T}(\theta-\mathrm{Flip}_i(\theta)\right|\right)
\end{aligned}$$

Plugging this in the relative entropy decomposition we obtain:

$$\begin{aligned}
D_{\mathrm{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\mathrm{Flip}_i(\theta)}) &\leq \mathbb{E}_\theta\left[\sum_{t=1}^T \frac{\alpha^2(x_t, \theta, \mathrm{Flip}_i(\theta))}{\dot\mu(x_t^\mathsf{T}\mathrm{Flip}_i(\theta))}\left\{x_t^\mathsf{T}(\theta - \mathrm{Flip}_i(\theta))\right\}^2\right] \\
&\leq \mathbb{E}_\theta\left[\sum_{t=1}^T \dot\mu(x_t^\mathsf{T}\theta)\left\{x_t^\mathsf{T}(\theta - \mathrm{Flip}_i(\theta))\right\}^2\right]\exp\left(2\left|x_t^\mathsf{T}(\theta - \mathrm{Flip}_i(\theta)\right|\right) \\
&\leq \exp(4\epsilon)\mathbb{E}_\theta\left[\sum_{t=1}^T \dot\mu(x_t^\mathsf{T}\theta)\left\{x_t^\mathsf{T}(\theta - \mathrm{Flip}_i(\theta))\right\}^2\right] \\
&\leq 2\epsilon^2\exp(4\epsilon)\mathbb{E}_\theta\left[\sum_{t=1}^T \dot\mu(x_t^\mathsf{T}\theta)[x_t]_i^2\right] \\
&= 2\epsilon^2\exp(4\epsilon)\mathbb{E}_\theta\left[\sum_{t=1}^T \dot\mu(x_t^\mathsf{T}\theta)\left[x_t - x_\star(\theta) + x_\star(\theta)\right]_i^2\right] \\
&\leq 4\epsilon^2\exp(4\epsilon)\mathbb{E}_\theta\left[\sum_{t=1}^T \dot\mu(x_t^\mathsf{T}\theta)\left[x_t - x_\star(\theta)\right]_i^2 + \sum_{t=1}^T \dot\mu(x_t^\mathsf{T}\theta)\left[x_\star(\theta)\right]_i^2\right]
\end{aligned}$$

where we last used the fact that $(a+b)^2 \leq 2(a^2+b^2)$. Therefore by summing over d:

$$\begin{aligned}
\sum_{d=2}^d D_{\mathrm{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\mathrm{Flip}_i(\theta)}) &\leq 4\epsilon^2\exp(4\epsilon)\mathbb{E}_\theta\left[\sum_{t=1}^T\sum_{i=2}^d \dot\mu(x_t^\mathsf{T}\theta)\left[x_t - x_\star(\theta)\right]_i^2 + \sum_{t=1}^T\sum_{i=2}^d \dot\mu(x_t^\mathsf{T}\theta)\left[x_\star(\theta)\right]_i^2\right] \\
&\leq 4\epsilon^2\exp(4\epsilon)\mathbb{E}_\theta\left[\sum_{t=1}^T\sum_{i=1}^d \dot\mu(x_t^\mathsf{T}\theta)\left[x_t - x_\star(\theta)\right]_i^2 + \sum_{t=1}^T\sum_{i=1}^d \dot\mu(x_t^\mathsf{T}\theta)\left[x_\star(\theta)\right]_i^2\right] \\
&\leq 4\epsilon^2\exp(4\epsilon)\mathbb{E}_\theta\left[\sum_{t=1}^T \dot\mu(x_t^\mathsf{T}\theta)\left\|x_t - x_\star(\theta)\right\|^2 + d\frac{\epsilon^2}{\left\|\theta_\star\right\|^2 + (d-1)\epsilon^2}\sum_{t=1}^T \dot\mu(x_t^\mathsf{T}\theta)\right] \\
&\leq 4\epsilon^2\exp(4\epsilon)\mathbb{E}_\theta\left[\sum_{t=1}^T \dot\mu(x_t^\mathsf{T}\theta)\left\|x_t - x_\star(\theta)\right\|^2 + \frac{d}{2}\epsilon^2\sum_{t=1}^T \dot\mu(x_t^\mathsf{T}\theta)\right]
\end{aligned}$$

where we used Equation (34) and the fact that $\|\theta_\star\| \geq 1$. Using Proposition 6 (more precisely Equation (33)) we obtain:

$$\sum_{d=2}^d D_{\mathrm{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\mathrm{Flip}_i(\theta)}) \leq 4\epsilon^2\exp(4\epsilon)\left(6\mathrm{Regret}_\theta(T) + \frac{d}{2}\epsilon^2\mathbb{E}_\theta\left[\sum_{t=1}^T \dot\mu(x_t^\mathsf{T}\theta)\right]\right) \tag{37}$$

We finish the proof by resorting to a Taylor expansion of $\dot\mu(x_t^\mathsf{T}\theta)$. Formally:

$$\sum_{t=1}^T \dot\mu(x_t^\mathsf{T}\theta) \leq \sum_{t=1}^T\left[\dot\mu(x_\star(\theta)^\mathsf{T}\theta) + \left|\int_{v=0}^1 \ddot\mu(x_\star(\theta)^\mathsf{T}\theta + v\theta^\mathsf{T}(x_t - x_\star(\theta)))dv\right|\left|\theta^\mathsf{T}(x_\star(\theta) - x_t)\right|\right]$$

Using the fact that $|\ddot\mu| \leq \mu$ and $x_\star(\theta)^\mathsf{T}\theta \geq x_t^\mathsf{T}\theta$ we obtain that:

$$\begin{aligned}
\mathbb{E}_\theta\left[\sum_{t=1}^T \dot\mu(x_t^\mathsf{T}\theta)\right] &\leq \mathbb{E}_\theta\left[\sum_{t=1}^T\left[\dot\mu(x_\star(\theta)^\mathsf{T}\theta) + \alpha(\theta, x_\star(\theta), x_t)\theta^\mathsf{T}(x_\star(\theta) - x_t)\right]\right] \\
&= \frac{T}{\kappa_\epsilon} + \mathbb{E}_\theta\left[\sum_{t=1}^T \alpha(\theta, x_\star(\theta), x_t)\theta^\mathsf{T}(x_\star(\theta) - x_t)\right] \\
&= \frac{T}{\kappa} + \mathrm{Regret}_\theta(T)
\end{aligned}$$

where we used the mean value theorem in the last line (see for instance the beginning of the proof of Proposition 6. Plugging this result in Equation (37) we obtain:

$$\sum_{d=2}^{d} D_{\mathrm{KL}}(\mathbb{P}_\theta, \mathbb{P}_{\mathrm{Flip}_i(\theta)}) \leq 4\epsilon^2 \exp(4\epsilon) \left(6\mathrm{Regret}_\theta(T) + \frac{d}{2}\epsilon^2 \left(\frac{T}{\kappa_\epsilon} + \mathrm{Regret}_\theta(T) \right) \right)$$

Averaging over Ξ and since by Hypothesis (**H1**) we know that $\mathrm{Regret}_\theta(T) \leq d\sqrt{T/\kappa_\epsilon}$ we obtain the announced result. □

D.7 Proof of Lemma 6

Lemma 6 (Relative Entropy Decomposition)**.** *For any θ, θ' we have that:*

$$D_{\mathrm{KL}}\left(\mathbb{P}_\theta, \mathbb{P}_{\theta'}\right) \leq \mathbb{E}_\theta \left[\sum_{t=1}^{T} \frac{\left(\mu(x_t^\mathsf{T}\theta) - \mu(x_t^\mathsf{T}\theta')\right)^2}{\dot{\mu}(x_t^\mathsf{T}\theta')} \right]$$

Proof. Denote $P_x^\theta = \mathbb{P}_\theta(r|x)$. Thanks to (Lattimore and Szepesvári, 2020, Section 24.1) we have:

$$\begin{aligned} D_{\mathrm{KL}}\left(\mathbb{P}_\theta, \mathbb{P}_{\theta'}\right) &= \mathbb{E}_\theta \left[\sum_{t=1}^{T} D_{\mathrm{KL}}\left(P_{x_t}^\theta, P_{x_t}^{\theta'} \right) \right] \\ &= \mathbb{E}_\theta \left[\sum_{t=1}^{T} D_{\mathrm{KL}}\left(\mathrm{Bernoulli}(x_t^\mathsf{T}\theta), \mathrm{Bernoulli}(x_t^\mathsf{T}\theta') \right) \right] \\ &\leq \mathbb{E}_\theta \left[\sum_{t=1}^{T} D_{\chi^2}\left(\mathrm{Bernoulli}(x_t^\mathsf{T}\theta), \mathrm{Bernoulli}(x_t^\mathsf{T}\theta') \right) \right] \end{aligned}$$

where we used $D_{\mathrm{KL}} \leq D_{\chi^2}$ (Tsybakov, 2008, Chapter 2). Using the expression of the χ^2-divergence for Bernoulli random variables finishes the proof. □

E TRACTABILITY OF OFULog-r

E.1 Proof of Proposition 3

Proposition 3. *Let $(\tilde{x}_t, \tilde{\theta}_t)$ be the pair returned by Algorithm 2. Then:*

$$(\tilde{x}_t, \tilde{\theta}_t) \in \operatorname*{arg\,max}_{x \in \mathcal{X}, \theta \in \mathcal{E}_t(\delta)} x^\mathsf{T}\theta \,.$$

Proof. Recall that we assume the arm-set $\mathcal{X}$ to be finite. For any $x \in \mathcal{X}$ denote:

$$\theta_x \in \operatorname*{arg\,max}_{\theta \in \mathcal{E}_t(\delta)} x^\mathsf{T}\theta \tag{38}$$

which is well-defined, as the maximizer of a concave function under a convex constraint. We can now write:

$$\tilde{x}_t \in \operatorname*{arg\,max}_{x \in \mathcal{X}} x^\mathsf{T}\theta_x \tag{39}$$

Since we have $\tilde{\theta}_t = \theta_{\tilde{x}_t}$ we can prove that the planning of **OFULog-r** is indeed optimistic:

$$\begin{aligned}
\tilde{x}_t^\mathsf{T}\tilde{\theta}_t &= \tilde{x}^\mathsf{T}\theta_{\tilde{x}_t} \\
&\geq x^\mathsf{T}\theta_x && \text{(Equation (39))} \\
&\geq x^\mathsf{T}\theta && \text{(Equation (38))}
\end{aligned}$$

which holds for any $x \in \mathcal{X}$ and $\theta \in \mathcal{E}_t(\delta)$. This finishes the proof. □

E.2 Proof of Corollary 1

Corollary 1. *Theorem 1, Proposition 2 and Theorem 3 are also satisfied by* **OFULog-r**.

Proof. The proof is fairly simple, as this result directly follows from Lemma 1. To see this, note that we only need two ingredients to repeat the proof of Theorem 1:

(1) We rely on optimism to enforce $x_t^\mathsf{T}\tilde{\theta}_t \geq x_\star(\theta_\star)^\mathsf{T}\theta_\star$. This fact this holds (with high probability) as thanks to Lemma 1 we have $\mathcal{C}_t(\delta) \subseteq \mathcal{E}_t(\delta)$ and therefore $\theta_\star \in \mathcal{E}_t(\delta)$ for all $t \geq 1$ with probability at least $1 - \delta$.

(2) We bound the deviation $\|\theta - \theta_\star\|_{\mathbf{H_t}(\theta_\star)}$ by $\widetilde{\mathcal{O}}(\sqrt{d\log(t)})$ terms for any $\theta \in \mathcal{C}_t(\delta)$ (cf. Proposition 4). The same property holds for any $\theta \in \mathcal{E}_t(\delta)$ thanks to Lemma 1.

As a result of **(1)** and **(2)** proving that **OFULog-r** satisfies Theorem 1 follows rigorously the same line of proof. The same arguments hold for proving that **OFULog-r** satisfies Proposition 2 and Theorem 3.

□

F SELF-CONCORDANCE RESULTS

In this section we state some useful generalized self-concordance results. The first technical result is from (Faury et al., 2020, Lemma 9). We provide a proof for the sake of completeness.

Lemma 7. *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 $z_1, z_2 \in \mathcal{Z}$:*

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

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

$$\begin{aligned}
& -1 \leq \frac{\ddot{f}(z)}{\dot{f}(z)} \leq 1 \\
\Rightarrow \quad & -|z_1 - z_0| \leq \int_{z_1 \wedge z_0}^{z_1 \vee z_0} \frac{\ddot{f}(z)}{\dot{f}(z)} dz \leq |z_1 - z_0| && (z_0 \in \mathcal{Z}) \\
\Leftrightarrow \quad & -|z_1 - z_0| \leq \log\left(\dot{f}(z_1 \vee z_0)/\dot{f}(z_1 \wedge z_0)\right) \leq |z_1 - z_0| \\
\Leftrightarrow \quad & \dot{f}(z_1 \wedge z_0)\exp\left(-|z_1 - z_0|\right) \leq \dot{f}(z_1 \vee z_0) \leq \dot{f}(z_1 \wedge z_0)\exp\left(|z_1 - z_0|\right). && (40)
\end{aligned}$$

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 Equation (40) we easily get:

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

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 Equation (40) instead. We have therefore proven the announced result, but only for $z = z_1$. The proof is concluded by realizing than 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). □

We now state a second result, which proof closely follows the one of Lemma 7.

Lemma 8. *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 $z_1, z_2 \in \mathcal{Z}$:*

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

Proof. From Equation (40) it can easily be extracted that for all $v \geq 0$:

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

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

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

where we defined:

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

Finally, we use Lemma 10 which guarantees that $g(z) \geq (2 + z)^{-1}$ for all $z \geq 0$ to prove the claimed result. □

We will need one last technical result obtained from the self-concordance property. Its proof can be extracted from Equation (40) in the proof of Lemma 7.

Lemma 9. *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 $z_1, z_2 \in \mathcal{Z}$:*

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

G AUXILIARY RESULTS

Lemma 10. *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 yields 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 7 (Polynomial Inequality)**.** *Let $b, c \in \mathbb{R}^+$, and $x \in \mathbb{R}$. The following implication holds:*

$$x^2 \leq bx + c \Longrightarrow x \leq b + \sqrt{c}$$

Proof. Let $f : x \to x^2 - bx - c$. 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:

$$\begin{aligned} x &\leq \max(\lambda_1, \lambda_2) \\ &\leq \frac{1}{2}(b + \sqrt{b^2 + 4c}) \\ &\leq b + \sqrt{c} && (\sqrt{x+y} \leq \sqrt{x} + \sqrt{y},\ \forall x, y \geq 0) \end{aligned}$$

□

The following theorem is extracted from (Abbasi-Yadkori et al., 2011, Lemma 10).

Lemma 11 (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 λ be a non-negative scalar. For $t \geq 1$ define $\mathbf{V}_t := \sum_{s=1}^{t-1} x_s x_s^\mathsf{T} + \lambda \mathbf{I}_d$. The following inequality holds:*

$$\det(\mathbf{V}_{t+1}) \leq \left(\lambda + (t-1)X^2/d\right)^d$$

We need a slight-variation of the Elliptical Potential Lemma (Abbasi-Yadkori et al., 2011, Lemma 11) adjusted to handle (increasing) time-varying regulations.

Lemma 12 (Elliptical potential)**.** *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 $\lambda_1 = 1$. For $t \geq 1$ define $\mathbf{V}_t := \sum_{s=1}^{t-1} x_s x_s^\mathsf{T} + \lambda_t \mathbf{I}_d$. Then:*

$$\sum_{t=1}^{T} \|x_t\|^2_{\mathbf{V}_t^{-1}} \leq 2d(1+X^2)\log\left(\lambda_T + \frac{TX^2}{d}\right)$$

Proof. By definition of $\mathbf{V}_t$:

$$\begin{aligned}
|\mathbf{V}_{t+1}| &= \left|\sum_{s=1}^{t-1} x_s x_s^\mathsf{T} + x_t x_t^\mathsf{T} + \lambda_t \mathbf{I_d}\right| \\
&\geq \left|\sum_{s=1}^{t-1} x_s x_s^\mathsf{T} + x_t x_t^\mathsf{T} + \lambda_{t-1} \mathbf{I_d}\right| && (\lambda_t \geq \lambda_{t-1}) \\
&= \left|\mathbf{V}_t + x_t x_t^\mathsf{T}\right| \\
&\geq |\mathbf{V}_t| \left|\mathbf{I}_d + \mathbf{V}_t^{-1/2} x_t x_t^T \mathbf{V}_t^{-1/2}\right| \\
&= |\mathbf{V}_t| \left(1 + \|x_t\|^2_{\mathbf{V}_t^{-1}}\right)
\end{aligned}$$

and therefore by taking the log on both side of the equation and summing from $t = 1$ to T:

$$\begin{aligned}
\sum_{t=1}^{T} \log\left(1 + \|x_t\|^2_{\mathbf{V}_t^{-1}}\right) &\leq \sum_{t=1}^{T} \log|\mathbf{V}_{t+1}| - \log|\mathbf{V}_t| \\
&= \log\left(\frac{\det(\mathbf{V}_{T+1})}{\det(\lambda_1 \mathbf{I}_d)}\right) && \text{(telescopic sum)} \\
&= \log\left(\det(\mathbf{V}_{T+1})\right) && (\lambda_1 = 1) \\
&\leq d \log\left(\lambda_T + \frac{TX^2}{d}\right) && \text{(Lemma 11)}
\end{aligned}$$

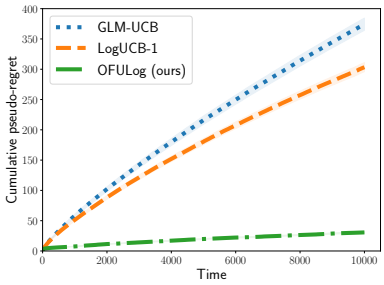
Remember that for all $x \in [0, 1]$ we have the inequality $\log(1 + x) \geq x/2$. Also note that $\|x_t\|^2_{\mathbf{V}_t^{-1}} \leq X^2/\lambda$. Therefore:

$$\begin{aligned}
d \log\left(\lambda_T + \frac{TX^2}{d}\right) &\geq \sum_{t=1}^{T} \log\left(1 + \|x_t\|^2_{\mathbf{V}_t^{-1}}\right) \\
&\geq \sum_{t=1}^{T} \log\left(1 + \frac{1}{\max(1, X^2/\lambda_t)} \|x_t\|^2_{\mathbf{V}_t^{-1}}\right) \\
&\geq \frac{1}{2\max(1, X^2/\lambda_1)} \sum_{t=1}^{T} \|x_t\|^2_{\mathbf{V}_t^{-1}} \\
&\geq \frac{1}{2(1 + X^2)} \sum_{t=1}^{T} \|x_t\|^2_{\mathbf{V}_t^{-1}}
\end{aligned}$$

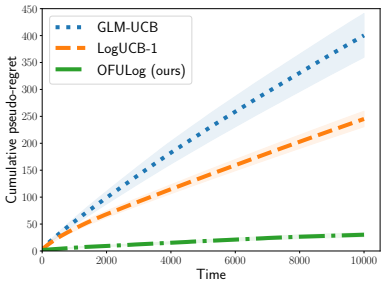
which yields the announced result. □

H NUMERICAL EXPERIMENTS

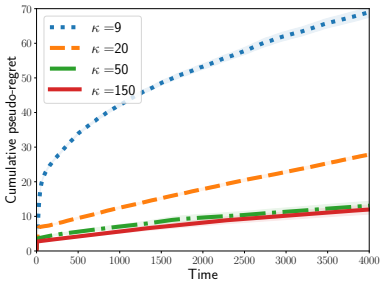
We present here a few illustrative experiments. We compare the three following algorithms: **GLM-UCB** (Filippi et al., 2010), **LogUCB1** (Faury et al., 2020) and **OFULog** (this work). We didn't implement **LogUCB2** (Faury et al., 2020): it is intractable as it relies on non-convex minimization that cannot be bypassed. These results, presented in Figure 5, corroborate our theoretical analysis: **(1)** our algorithm displays a clear advantage over previous approaches (Figures 5a and 5b) **(2)** a higher level of non-linearity (i.e higher values of κ) is actually beneficial (Figure 5c) for **OFULog**. Remember that this cannot be the case for other approaches as by design, the performances of **GLM-UCB** and **LogUCB1** can only degrade when κ increases. The arm-set is composed of 40 arms, drawn uniformly at random on the 2-dimensional ball at the beginning of each run. For each experiment, we average the regret curves over 50 independent runs and report standard-deviation in shaded colors.



(a) Regret curves for $\kappa = 50$ in $d = 2$ with 40 arms.



(b) Regret curves for $\kappa = 400$ in $d = 2$ with 40 arms.



(c) Regret curves of **OFULog** in $d = 2$ with 40 arms for different κ.

Figure 5: Illustrative numerical experiments. Shaded areas represent 1-standard deviation of the cumulative regret, aggregated over 50 independent experiments.