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 $\|\theta_\star\| \leq S$.*

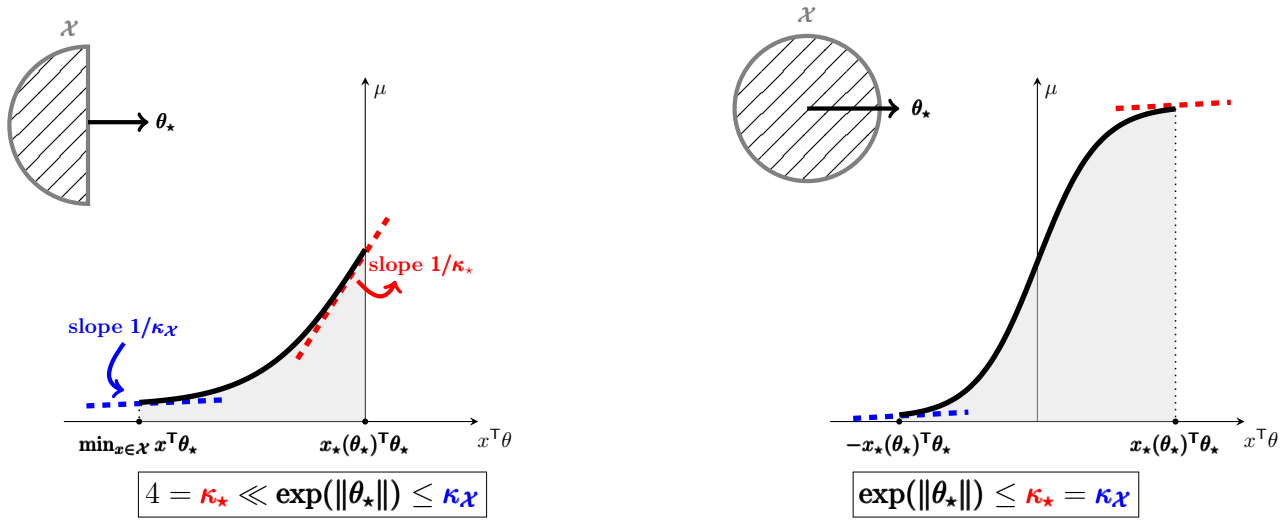
We will denote $\Theta := \mathcal{B}_d(0, S)$. For any $\theta \in \Theta$, we will use the notation $x_\star(\theta) := \arg\max_{x\in\mathcal{X}} x^\mathsf{T}\theta$. At each round t, the agent takes a decision following a policy $\pi : \mathcal{F}_t \to \mathcal{X}$, mapping $\mathcal{F}_t := \sigma(\{x_s, r_{s+1}\}_{s=1}^{t-1})$ (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

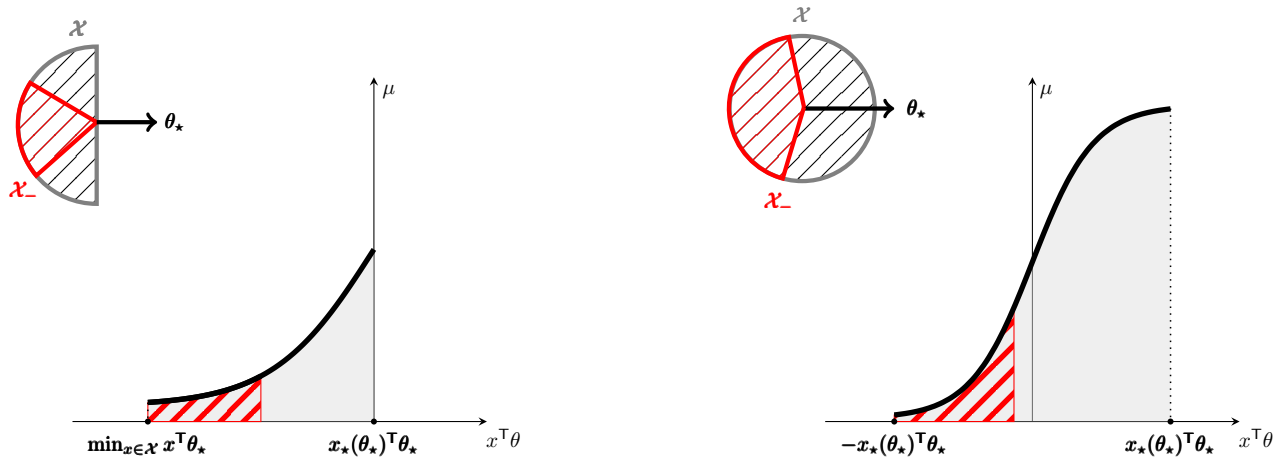
OFULog is the counterpart of the **LB** algorithm **OFUL** of Abbasi-Yadkori et al. (2011). At each round it computes $\hat{\theta}_t$ and the set $\mathcal{C}_t(\delta)$. It then finds an optimistic parameter $\theta_t \in \mathcal{C}_t(\delta)$ and plays x_t the greedy action w.r.t θ_t. Formally:

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

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

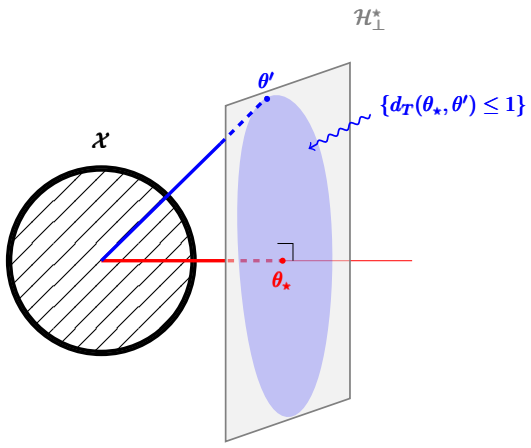


Figure 3: Illustration of the construction behind the local lower-bound.

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 $\mathcal{C}_t(\delta)$:

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

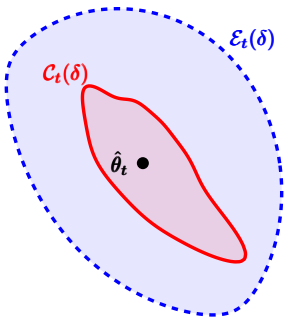


Figure 4: The confidence set $\mathcal{C}_t(\delta)$ and its convex relaxation $\mathcal{E}_t(\delta)$ obtained through a trajectory with: $T = 1000$, $\mathcal{X} = \mathcal{B}_d(0, 1)$ and $\kappa_{\mathcal{X}} = 22$.

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.